MICROBIOME AND MICROBIAL INFORMATICS

EDITED BY: Zheng Zhang, Juntao Liu, Pengfei Ding, Ivan Erill and Zhenhua Ming

PUBLISHED IN: Frontiers in Microbiology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-83250-682-0
DOI 10.3389/978-2-83250-682-0

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

MICROBIOME AND MICROBIAL INFORMATICS

Topic Editors:
Zheng Zhang, Shandong University, China
Juntao Liu, Shandong University, Weihai, China
Pengfei Ding, University of Maryland, Baltimore County, United States
Ivan Erill, University of Maryland, Baltimore County, United States
Zhenhua Ming, Guangxi University, China

Citation: Zhang, Z., Liu, J., Ding, P., Erill, I., Ming, Z., eds. (2022). Microbiome and Microbial Informatics. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83250-682-0

Table of Contents

frontiers | Frontiers in Microbiology

TYPE Editorial
PUBLISHED 19 October 2022
DOI 10.3389/fmicb.2022.1054811

OPEN ACCESS

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

*CORRESPONDENCE
Zheng Zhang
zhangzheng@sdu.edu.cn

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

RECEIVED 27 September 2022
ACCEPTED 10 October 2022
PUBLISHED 19 October 2022

CITATION
Ding P, Ming Z, Liu J, Erill I and Zhang Z (2022) Editorial: Microbiome and microbial informatics. *Front. Microbiol.* 13:1054811. doi: 10.3389/fmicb.2022.1054811

Editorial: Microbiome and microbial informatics

Pengfei Ding[1], Zhenhua Ming[2], Juntao Liu[3], Ivan Erill[4] and Zheng Zhang[5]*

[1]Department of Chemistry and Biochemistry, University of Maryland Baltimore County, Baltimore, MD, United States, [2]State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Life Science and Technology, Guangxi University, Nanning, China, [3]School of Mathematics and Statistics, Shandong University, Weihai, China, [4]Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, MD, United States, [5]State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao, China

KEYWORDS

microbiome, bioinformatics, evolutionary and genomic microbiology, metagenomics, data mining, computational techniques

Editorial on the Research Topic
Microbiome and microbial informatics

The advancement of genome sequencing technologies and metagenomic analysis has allowed researchers to study microorganisms, as well as their functions and microbial-based interactions in natural and industrial environments. Nevertheless, the large amounts of information resulting from these studies must be stored, structured, indexed, analyzed, and correlated with existing experimental data. The requirement has led to the exploitation of bioinformatics solutions at the cross-over point of information science and microbiology.

In this context, we are pleased to note that the Microbiome and Microbial Informatics Research Topic has drawn the contributions of lots of well-respected researchers in the field all over the world, including those from China, India, Germany, the Netherlands, Saudi Arabia and Australia. We received 28 submissions, 17 of which were accepted for publication after peer review. These publications focused on new insights, novel developments, current challenges, and future perspectives in the field of microbiome and microbial informatics. We sincerely thank all researchers who have agreed to contribute to our Research Topic.

Profiting from the current rapid progress of artificial intelligence techniques, the aggregation of statistical analysis methodologies and predictions for large-scale data has evolved for a variety of fields associated with data science. Jiang et al. summarized the application and advancement on machine learning and deep learning in microbiology. They illustrated and contrasted the benefits and drawbacks of distinct algorithmic tools in four dimensions: microbiome and taxonomy, microbial ecology, pathogen and epidemiology, and drug discovery, demonstrating the development prospects of computational microbiology from the perspective of machine learning. As an example, by combining data augmentation techniques, López et al. utilized machine learning methods to investigate the predictability of smoking habits from class imbalanced saliva

microbiome data to account for class imbalance. In doing so, they successfully addressed the class imbalance problem in microbiome data, resulting in a reliable prediction of smoking habits.

The rapid development of high-throughput, culture-independent analytical techniques has brought a wealth of experimental data that have significantly facilitated the human microbiome study. To study the explicit microbial variance in the human face, Wei et al. reassessed data from 822 shotgun-metagenomic sequencing of Han Chinese individuals in conjunction with 97 North American samples from the NIH Human Microbiome Project (HMP). This study explores the fine-scale facial location-related variations of skin microbiomes to provide an in-depth understanding of ecological processes that underlie facial microbial changes. Wang et al. analyzed 404 datasets from human oral saliva samples and made comparisons with other human part samples to reveal the diversity and biogeography of human oral saliva microbial communities. Using high-throughput sequencing of 16S rRNA V3-V4 hypervariable regions, Huang, Deng et al. assessed the fecal microbiota profiles of healthy individuals from three representative Han populations in Guangdong Province, China. On the basis of genus-level OTU abundance, the random forest prediction model indicated that there may be potential to distinguish individuals according to their fecal microbial community. Li et al. analyzed the evolution of the gut microbiome in Tibetan populations in the Minjiang River Basin. This study demonstrates that altitude of habitation is a vital factor influencing the enterotype of the Tibetan population microbiome. Sindi et al. concluded that short-term maternal dietary interventions during lactation could significantly alter the functional potential of the gut microbiome of breastfed infants. Another study by Peng et al. showed that esterases from *Bifidobacteria* undertake albiflorin conversion in the gut and play an important role in the metabolism of natural compounds including ester bonds. *Bifidobacteria*-mediated metabolism of ester bonds has the potential to facilitate the exploitation of novel enzymes and probiotic adjuvant compounds for therapeutic use.

The development of omics technologies has greatly increased our understanding of the interaction between microbes and agricultural animals and plants. Na et al. evaluated the effects of adding six common commercial lactic acid bacterial additives in the microbial communities and condition of fermentation of alfalfa silage. The study demonstrated that lactic acid bacterial additives enhanced the quality of fermentation and changed the microbial communities of alfalfa silage. Zhang et al. investigated the reaction in fungal subcommunities in a corn-wheat rotation plow land managed by long-term conservation tillage. Their findings indicated that the use of no-tillage and straw mulching practices had a negative impact on the complicacy of plentiful and medium fungal networks, but did not prominently affect rare fungal networks. Their study informs our learning on the reaction in fungal subcommunities to preservation tillage management technologies, and provides a new view on how fungal subcommunity assemble. Wen et al. reported the discovery of a new NAD(P)-dependent alcohol dehydrogenase from *Gluconobacter frateurii* NBRC 3264, which displayed great potentiality for application in processes involving high-yield bioconversion of D-allulose and could therefore be used for the manufacturing production of D-allulose.

Molecular diagnostics are extensively applied in clinical microbiology studies, such as routine detection and epidemiological analysis of infectious microbes. Liao et al. presented a concise multilocus sequence typing protocol for *Staphylococcus aureus* and demonstrated the effectiveness of portable sequencing technology for accurate, rapid, and routine molecular typing.

Molecular taxonomy and environmental adaptation have been deeply studied due to the increased genomic information of some microbial species. For example, Du et al. isolated a novel pathogenic bacterium, *Haemobacillus shengwangii*, from a blood sample of a critically ill patient. They classified *H. shengwangii* as a member of the Thermicanaceae family, for which they report the first high-quality genome, by utilizing single-molecule real-time sequencing and next-generation sequencing technologies. Mahata et al. combined morphological descriptions, phylogenetics and single-nucleotide polymorphism analysis to characterize the distribution and relative abundance of *Aspergillus* species from *Foeniculum vulgare*. The integration of morphological features with molecular systematics is regarded as an essential element of taxonomic studies. Huang, Peng et al. isolated and identified 22 fungal strains from the Beibu Gulf coral using serial dilution and internal transcribed spacer sequence analysis. The isolates were further divided into three branches by phylogenetic analysis. Their study provided eight fungal isolates with potential activity against *Vibrio* species, and two alkaloid-type antibiotics with anti-Vibrio effects were characterized from the bioactive strain *Fusarium equiseti* BBG10. Liu et al. characterized the diversity and function of laccase family genes in the fungus *Schizophyllum commune* 20R-7-F01 genome, which was isolated from deep sea sediment. Their findings contribute to further our understanding of laccase genes in white-rot fungi and pave the way for further exploring the relationship between the laccase gene family and anaerobic degradation of lignin by *Schizophyllum commune*. Yuan et al. conducted comparative genomic and functional analyses of *Paenibacillus peoriae* ZBSF16, a species with potential for biocontrol against grapevine diseases. Their analysis provided insight into the plant growth-promoting and biocontrol mechanisms of this bacterium, identifying conserved genes involved in both plant-growth promotion and antibiotic production.

Modern microbiology studies lead to increased adoption of high-throughput techniques and big data methods to

provide faster, unbiased and more reproducible results than traditional studies with insufficient data or time-consuming pure experimental techniques. We created this Research Topic with the hope that the contributions submitted to it would prove useful for a wide audience, but in particular to microbiologists, computational biologists and bioinformaticians. We believe that the high-quality contributions published within this Research Topic, together with the diversity of microorganisms and environments studied and the broad array of experimental and computational techniques used, have amply achieved our goal.

Author contributions

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

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

frontiers in Microbiology

ORIGINAL RESEARCH
published: 18 February 2022
doi: 10.3389/fmicb.2022.832320

Substantive Morphological Descriptions, Phylogenetic Analysis and Single Nucleotide Polymorphisms of *Aspergillus* Species From *Foeniculum vulgare*

Pranab Kumar Mahata[1], Regina Sharmila Dass[1]*, Archana Pan[2] and Babylakshmi Muthusamy[3,4]

[1] Fungal Genetics and Mycotoxicology Laboratory, Department of Microbiology, School of Life Sciences, Pondicherry University, Pondicherry, India, [2] Centre for Bioinformatics, School of Life Sciences, Pondicherry University, Pondicherry, India, [3] Institute of Bioinformatics, International Tech Park, Bengaluru, India, [4] Manipal Academy of Higher Education, Manipal, India

OPEN ACCESS

Edited by:
Zheng Zhang,
Shandong University, China

Reviewed by:
Kamel A. Abd-Elsalam,
Agricultural Research Center, Egypt
Fuguo Xing,
Institute of Food Science and Technology, Chinese Academy of Agricultural Sciences (CAAS), China

****Correspondence:***
Regina Sharmila Dass
reginadass@gmail.com

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

Received: *09 December 2021*
Accepted: *24 January 2022*
Published: *18 February 2022*

Citation:
Mahata PK, Dass RS, Pan A and Muthusamy B (2022) Substantive Morphological Descriptions, Phylogenetic Analysis and Single Nucleotide Polymorphisms of Aspergillus Species From Foeniculum vulgare. Front. Microbiol. 13:832320. doi: 10.3389/fmicb.2022.832320

Ascomycetous fungi are found associated with a wide variety of substrates which range from fresh water to marine ecosystems, tropical to temperate forest soils and deserts, throughout the world over. These demystifying fungi exist as endophytes, pathogens and saprobes. They have been studied due to their ability to contaminate foods and feedstuffs, causing an elaboration of mycotoxins. The objectives of the study included extensive analyses of the morphological features of fungi, especially *Aspergilli*, which have been presented while studying them on specific mycological media. It is also an elaborate compilation of substantive macro- and micro-morphological characterization of different *Aspergilli* isolated from the spice *Foeniculum vulgare* used in India and other countries in the world. Further, a first of its kind attempt has been made to study their relative abundance and frequency of occurrence, molecular phylogeny and genetic relatedness to characterize the *Aspergilli* into specific sections, groups and clades. Single nucleotide polymorphism (SNP) analysis was carried out to evaluate the functional consequences of nucleotide variations, synonymous and non-synonymous mutations in the protein structure. The study resulted in a total of 3,506 *Aspergillus* isolates, which were obtained from seventy (70) fennel samples, representing 14 *Aspergillus* species. The two most frequently found species were *A. niger* and *A. flavus* with a relative abundance of 32.24 and 11.63%, respectively. The taxonomy and current placements have been reappraised with suggestions and prospects for future research from six sections namely *Terrei, Flavi*, *Fumigati*, *Nidulantes*, *Nigri,* and *Versicolores*. In addition, a total number of 27 isolates were studied and deposited at the National Centre for Biotechnology Information (NCBI) and five *Aspergillus* species have been identified and are being reported for the first time from the fennel seeds, based on partial sequence analysis of the official fungal barcode namely, Internal Transcribed Spacer (ITS) and a functional gene, beta tubulin gene locus, coupled with phenotypic characterization. SNPs for specific DNA regions have been used to identify variants in

Aspergilli obtained from Indian fennel seeds for the first time. The need for a polyphasic approach of morphological identification and genetic characterization of *Aspergilli* from *Foeniculum vulgare* is addressed and presented here in adequate detail. Our current work makes extensive use of partial beta-tubulin gene sequences analyses to evaluate the association between SNPs in five *Aspergillus* species sections.

Keywords: fungal genomics, *Aspergillus*, polyphasic, phylogenetic analysis, single nucleotide polymorphisms (SNPs), single nucleotide variations (SNVs), non-synonymous and synonymous mutation

INTRODUCTION

Foeniculum vulgare Mill, commonly known as fennel in English and *saunf* in Hindi, belongs to the Umbelliferae (Apiaceae) family, is a flowering plant species, and is accredited by the International Code of Botanical Nomenclature (ICBN) (Badgujar et al., 2014). It is a perennial type of herb and is available in many regions like the United States, Northern Europe, southern Canada, Asia and Australia. This spice is flavorful, aromatic with culinary and medicinal uses. Fennel seeds are considered as a rich source of protein (2%), dietary fiber (10%), vitamin A (4.5%), thiamin (1%), vitamin C (14%) and dietary minerals like calcium (4%), iron (4%), magnesium (4%), and manganese (7%) with essential fatty acids. They are known to possess properties like antioxidant, antitumor, cryoprotective, hyper protective, hyperglycemic, and estrogenic activities (Javidnia et al., 2003; Samadi-Noshahr et al., 2021). It has a memory enhancing effect and can reduce stress (Koppula and Kumar, 2013). More importantly, *F. vulgare* has been used in the control of infectious diseases of bacterial, viral, fungal, mycobacterial and protozoan origin (Rather et al., 2016).

Statistics reveal that fennel production in India has doubled (58,265 tons) ever since the year 2000–2001 (27,332 tons), indicating the consumption trend in the sub-continent. Fennel is not just produced and processed in India, but it also exported. India exported around 23,562,460, and 20,295,380 million metric tons in the years 2018 and 2020–2021, respectively (Anonymous, 2020). The Ministry of Food Processing has estimated a loss of nearly 93 crores INR, due to harvest and post-harvest losses of India's agricultural products (Moloney, 2019), fennel being one of them.

F. vulgare, like most cereals and grains can be infected by range of mycoflora, some of which may significantly damage the economic value of the crop. *Aspergillus* spp. infection and several other genera have been of specific concern in the recent years, because these fungal phytopathogens produce toxic metabolites (mycotoxins) which represent significant contaminants of food (Ahmad et al., 2014), feed (Streit et al., 2012), and agricultural commodities, spices being one of them (Makhlouf et al., 2019). The genus *Aspergillus* consists of a few over one hundred mold species, ever since its first characterization nearly 300 years ago. It is mitosporic and conidial group of fungi, where in some species exhibit teleomorphic stages and hence are classified under the division Ascomycota (Bennett, 2010). Apart from their wide use in the industry for benefits (Casas López et al., 2004), *Aspergillus* species are prolific producers of secondary metabolites known as mycotoxins (Goto et al., 1996). Some species of *Aspergillus* cause diseases in humans (Alshehri and Palanisamy, 2020; Kashyap, 2020), animals and birds (Arné et al., 2021). More than sixty *Aspergillus* sp. are pathogens which pose a major health concern (Pal et al., 2014). *Aspergilli* are the common contaminants of food and feedstuffs. Spices (for example, fennel seeds) occupy a prominent role in the day-to-day culinary preparations in India (Siruguri and Bhat, 2015) and are not free from being contaminated by a variety of mycotoxins. Post-harvest contamination and spoilage during storage deteriorates the nutritive value, make them unfit for consumption because of the production and elaboration of toxic secondary metabolites. The genus *Aspergillus* comprising of *A. flavus*, *Aspergillus* section *Nigri*, *A. oryzae*, *A. parasiticus*, *A. terreus,* and *A. versicolor* are frequent contaminants in agricultural commodities like paddy, milled rice, peanuts, maize, millets, and wheat. These *Aspergilli* produce several mycotoxins at different relative humidities and storage periods, namely aflatoxin B_1, B_2, G_1, G_2, ochratoxin A, and fumonisin B_1 (Amadi and Adeniyi, 2009). Aflatoxins are toxigenic, carcinogenic, mutagenic, teratogenic, immunosuppressive, and are produced by *Aspergilli* as by-products which can also contaminate crops like peanut and cotton (Bhatnagar-Mathur et al., 2015; Soni et al., 2020) in addition to spices. Of the multitude of mycotoxins produced by *Aspergillus* species, aflatoxins B_1 (Nurtjahja et al., 2019), fumonisin B_2 (Han et al., 2017), and ochratoxin A (Magnoli et al., 2007), seem to be quite stable during storage of seeds and spices. Despite extensive toxicological studies with aflatoxins and other mycotoxins, the significance to human health, in case of several mycotoxins remains unclear. Synergistic interactions due to the co-occurrence of toxins is yet to be established.

Aspergillus species have medical and commercial importance. The genus *Aspergillus* has economic importance in different aspects like fermentation industry and enzyme production (Malathi and Chakraborty, 1991). Members of the genus produce a wide range of secondary metabolites, including lovastatin, a cholesterol-lowering drug, antitumor metabolites, etc. (Casas López et al., 2004). Few members of the genus are also the source of natural products which can be used to treat human diseases (Hiort et al., 2004).

Challenges in Identification of *Aspergillus* Species by Morphological Characterization

Mycologists have traditionally used phenotypic characteristics as a sole means for fungal identification. Classification of

Aspergilli, which comprise highly speciose lineages, requires techniques such as fungal barcoding for accurate species identification. *Aspergilli* have been recognized since ancient times by the formation of a common morphological structure referred to as the "*Aspergillum*" which consists of an asexual reproductive and distinctive entity, with a characteristic stalk-like "conidiophore" bearing a bulbous vesicle, which in turn bears the phialides and metulae, on which chains of microconidia are arranged (Bennett, 2010) in basipetal succession. Morphological identification of *Aspergillus* spp. by using macro-morphological characters like colony diameter, color, obverse, and reverse colony characters, etc., and microscopic features like conidial heads, vesicles shape, length of vesicles, stipes, seriation, metula, conidia color, size shape and ornamentations, asci, ascospore size, shape, color, etc., would only help in the preliminary assignment of generic names, leading to an inadequacy in identification to assign species. The limitations posed by morphological identification are many. These tests are extremely time consuming, cumbersome, requiring skilled personnel and expertise in carrying out the whole process. Hence, fungal phylogeny which uses genetic and molecular tools, are being used alongside the conventional techniques to address the gap of misidentification (Raja et al., 2017).

To overcome this constraint, a set of adoptable standard procedures like the use of nuclear ribosomal genes (ITS) most commonly used in fungal identification, considered as the official barcoding marker and methods used for construction of phylogenetic tree, which facilitates species identification has been carried out (Raja et al., 2017). The ITS and beta-tubulin genes have not been systematically used in *Aspergillus* taxonomy from *F. vulgare* till date, and have thus been appropriately evaluated in this study. In addition, our purpose was to examine protein-coding functional sequences as a major genetic marker for assessing intraspecies diversity. Therefore, the beta-tubulin gene is being considered in the current study since researchers (Cleveland and Sullivan, 1985) have reported that multiple tubulin genes are required in all but the simplest eukaryotic organisms (Edgcomb et al., 2001), for the formation, maintenance and preservation of the essential microtubules. The beta-tubulin gene appears to occur as a single-copy (α_1, α_2, β_1, γ tubulins) in *Aspergillus niger,* single-copy (α_1, γ tubulins), two-copy (β_1) or three-copy sequences (α_2 tubulins) in *Aspergillus nidulans* (Zhao et al., 2014). Among the various genes with basic functionality of coding for proteins in all eukaryotes (Mages et al., 1995; Schütze et al., 1999), the beta-tubulin gene (Keeling et al., 2000; Ayliffe et al., 2001; Mukherjee et al., 2003; Juuti et al., 2005), is also essential for sexual development in *A. nidulans* (Kirk and Morris, 1991) and has received enormous attention due to its highly conserved nature. This is especially important while phylogenetic constructions are undertaken with a broad range of organisms or for a specific genus.

Single nucleotide polymorphisms (SNPs) are single-nucleotide substitutions of one base for another and are definitely the most prevalent set of genetic variety. There are two sorts of techniques to examine SNPs: genomic and functional. SNPs have a variety of consequences, for instance, drug resistance, effects on mRNA splicing, nucleocytoplasmic export, etc. There is no information available on SNPs in *Aspergilli* isolated from spices. The current analysis has made efforts to identify synonymous and non-synonymous mutations, to examine them evolutionarily, and to reclassify *Aspergilli* into their respective Sections for appropriate taxonomic identification.

With this background, the predominant objectives of this study constituted, the elaborate and detailed macro- and micro-morphological studies of *Aspergillu*s species, molecular characterization of *Aspergillus* isolates obtained from *Foeniculum vulgare* samples using universal fungal bar-coding of the Internal Transcribed Spacer (ITS) and a functional genetic marker namely, beta–tubulin (β) gene sequences. This analysis was also performed with the primary objective of molecular fungal barcoding to substantiate our morphological analysis. Further, phylogenetic and single nucleotide polymorphism analyses have also been performed, for the first time from *Aspergilli* isolated from Indian fennel seed samples.

MATERIALS AND METHODS

Collection of Samples

A total number of seventy (70) samples of the Indian spice *Foeniculum vulgare* were collected from different retail markets from several regions, namely New Delhi, Lucknow, Pinjore, West Midnapore, and Puducherry from India. These samples were from urban, semi-urban and rural zones from different geographical regions of India. The samples were purchased in portions of 100 g and stored in sterile polyethylene zip-lock pouches. All samples were adequately labeled with suitable codes. The pouches were then transported to the laboratory and analyzed sequentially at the earliest. Representative spice samples were stored at 4°C for further analysis.

Macro-Morphological Cultural Studies of *Aspergilli*

Standard methods of agar plating technique as described by the International Seed Testing Association (ISTA) were chosen for the mycofloral analyses of the Indian fennel samples. Briefly, 25 g of the spice samples was weighed, washed thrice with distilled water and then rinsed with 1% sodium hypochlorite (NaOCl) solution for one min to inactivate surface contaminants. Subsequently, the seeds were rinsed in distilled water to remove any adhering sodium hypochlorite. The seeds were briefly allowed to stand for about 5–10 min after they had been placed in sterile petriplates lined with sterile tissue paper and plated using standard procedures. The petri plates containing the seeds were incubated in upright position at 25–27°C for 7 days. The results were expressed as percentage of infection and the frequency of occurrence was calculated (Reddy et al., 2009; dos Santos-Ciscon et al., 2019).

Aspergillus isolates were obtained after plating *Foeniculum vulgare* spice samples on mycological/fungi-specific media namely fungal agar (FA) and potato dextrose agar (PDA) media (HiMedia, Mumbai, India) by following standard agar plating procedures (Xie et al., 2007; Hamzah et al., 2018). All media

were prepared with the mandatory addition of a thermostable, antibacterial antibiotic namely chloramphenicol (Sigma-Aldrich, Steinhein, Germany) at the rate of 20 mg/1,000 mL of the fungi-specific media proposed. Briefly, representative samples were washed thrice with distilled water, dried for about 15 min and then plated @ 25 seeds per petri plate (100 × 15 mm, S-line, Borosil®, Mumbai, India), using aseptic techniques. The plates were then incubated at 25°C ± 2°C for 5–7 days with alternating periods of 12 h light and 12 h darkness. On the 7th day, *Aspergillus* colonies (green, greenish yellow, ochre/yellow, black, brownish-black, dark-brown, biscuit-brown, blue) were observed and isolated from the spice samples onto PDA medium. Spore suspensions of the *Aspergilli* were prepared and about 20 μL was transferred onto fresh PDA media. Colonies arising from single germinating spores were checked using Stereo Binocular Microscope (Magnus MSZ-Bi, Model: 13M1009), chosen for further studies and identified based on descriptions made in Fungal keys and Manuals (Thom and Raper, 1945; Varga and Samson, 2008) and Monographs (Refai et al., 2014).

Macro-Morphological Microscopic Studies of *Aspergilli*

Stereo binocular microscopic observations were carried out to detect presence of different species of *Aspergilli* and distinguish them from other fungal genera like *Alternaria*, *Curvularia*, *Fusaria*, *Penicillia*, *Mucor, Mycelia sterilia*, *Trichothecium,* etc., encountered during the analysis. Fungal flora were isolated from the fennel samples on potato dextrose agar (PDA) and czapek dox agar (CDA) media and identified using fungal manuals and keys (Thom and Raper, 1945; Refai et al., 2014). While mycoflora were being studied, emphasis was laid to the observation and isolation of *Aspergilli*. Macromorphological studies were carried out on aforesaid agar media. The isolates were inoculated either by single point or three-point inoculation method in glass petriplates (100 × 15 mm, S-line, Borosil®, Mumbai, India). After 7 days of incubation colony diameters, color, texture, sporulation pattern, obverse and reverse colony colors/characters and presence or absence ascomata (in older cultures incubated for 15–25 days) were determined.

Micro-Morphological Microscopic Studies of *Aspergilli*

The microscopic features of the isolated *Aspergilli* were sequentially studied using Light Microscopy (Olympus CH20i), Scanning Electron (SE; Hitachi, Model E-1010) and Differential Interference Contrast (DIC) microscopic analysis. Prominence was given to study characteristics like hyphal nature, septa, conidiophores, vesicle shape, sterigmata arrangement, and conidia formation. Standard protocols were followed while studying the microscopic features as outlined in identification Keys and Manuals (Thom and Raper, 1945; Varga and Samson, 2008). The light and scanning electron micrographs were taken at the Fungal Genetics and Mycotoxicology Laboratory, Department of Microbiology, Pondicherry University, while the DIC microscopic images (Nikon Upright Motorised Microscope, ECLIPSE Ni series, Nikon Corporation, Tokyo) were taken at the Fungal Biotechnology Laboratory, Department of Biotechnology, School of Life Sciences, Pondicherry University.

Preparation of Fungal Cultures for Genomic DNA Isolation

As and when microscopic analyses were being performed pure cultures of *Aspergilli* were inoculated onto PDA medium and incubated at 28°C for 7 days. Fungal mycelium from each petriplate was scraped off the medium using disposable sterile blades (Carbon steel Scalpel Blade No. 11, Olrada, India) and transferred to sterile pestle and mortar. The mycelial biomass (approximately close to 100 mg was collected) was used for DNA isolation.

Isolation of Fungal Genomic DNA

DNA was isolated from fungal isolates grown on synthetic PDA plates and DNA isolation was carried out using a Standard Kit (Gene JET plant Genomic DNA Purification Kit- K0791, Thermo Fisher Scientific, Vilnius, Lithuania). 100 mg of the fungal mycelial biomass was transferred to a sterile pestle and mortar. To this liquid nitrogen (200 mL) was carefully added along the walls of the mortar and crushed adequately in clockwise fashion for 5 min. The crushed mycelial biomass was immediately transferred to a new, sterile microcentrifuge tube (1.5 mL; Tarsons, Kolkata, India). Soon after, Lysis Buffer A (350 μL) and Lysis Buffer B (50 μL) were added in succession. A tiny quantity of RNase A (20 μL) was transferred to the microcentrifuge tubes containing the lysis buffers and crushed mycelial biomass. The specimen was then subjected to heat treatment for 10 min at 65°C using a sterile water bath with occasional stirring. After the heat treatment, the precipitation solution (130 μL) was added as per manufacturer's instructions. The microcentrifuge tubes were adequately mixed and kept on frost for 5 min. The tubes were then spun at ≥ 20,000 g (≥ 14,000 rpm) for 5 min. The tubes were removed, the supernatant (precisely 450–550 μL) was collected and transferred to a clean microcentrifuge tube, and an equal volume of plant gDNA binding solution and 96% ethanol (400 μL) was added and vigorously mixed. Thereafter, 600–700 μL of the prepared solution was transferred, isolate-wise, to the new spin columns. The supernatant thus collected was centrifuged at 6,000 g (∼8,000 rpm) for about a min. Using the same column, the supernatant solution was removed and the remaining mixture was centrifuged for 1 min. The wash buffer I (500 μL) was added into the spin-columns and centrifuged at 8,000 g (∼10,000 rpm) for 1 min, followed by the addition of wash buffer II (500 μL) in the spin columns again. This time the spin columns were centrifuged at high speed of ≥ 20,000 g (≥ 14,000 rpm) for 3 min. The contents of the collection tube within the spin columns were discarded, and the column contents were transferred to fresh, new microcentrifuge tubes. For genomic DNA elution, the elution buffer (30 μL) was placed mid-point of the spin column layer, incubated at room temperature for 5 min. The microcentrifuge tubes were spun at a speed of 8,000 g (∼10,000 rpm) for 1 min. A second elution step was carried out using the elution buffer (20 μL). The purified DNA was stored at −20°C for further use.

Polymerase Chain Reaction of the Internal Transcribed Spacer and β-Tubulin Genes

A total number of 27 isolates were selected for a phylogenetic analysis of the ITS and β-tubulin gene using (Tam et al., 2014) as one of the references. Partial amplification of the Internal Transcribed Spacer (ITS) region and functional gene β-tubulin was performed using the primers ITS1 (5′-TCCGTAGGTGAACCTGCGG-3′) and ITS4 (5′-TCCTCCGCTTATTGATATGC-3′) for ITS (Op De Beeck et al., 2014; Tam et al., 2014), bT2a (5′-GGTAACCAAATCGGTGCTGCTTTC-3′) and bT2b (5′-ACCCTCAGTGTAGTGACCCTTGGC-3′) for β-tubulin (Tam et al., 2014) respectively. Polymerase Chain Reaction (PCR) reactions were performed in a 25 μL volume, containing 12.5 μL (Taq DNA Polymerase 2x Master Mix RED, Ampliqon PCR Enzymes and Reagents, Stenhuggervej, Denmark), Forward primer (10 pmoL) 2.0 μL, Reverse primer (10 pmoL) 2.0 μL, Template DNA X μL (genomic DNA: 10–500 ng), Molecular biology grade water X μL (HiMedia®, Mumbai, India) for ITS gene, 12.5 μL (Taq DNA Polymerase 2x Master Mix RED, Ampliqon PCR Enzymes and Reagents), Forward primer (10 μM) 1.0 μL, Reverse primer (10 μM) 1.0 μL, Template DNA X μL (genomic DNA: 10–500 ng), Molecular biology grade water X μL (HiMedia®, Mumbai, India) for β-tubulin gene amplification. The cycling protocol consisted of an initial denaturation step of 94°C for 1 min 30 s, 32 cycles of denaturation at 95°C for 35 s, annealing for 55 s at 55°C for ITS and an initial denaturation step of 95°C for 3 min, 32 cycles of denaturation at 94°C for 30 s, annealing for 45 s at 65.5°C for β-tubulin and extension at 72°C for 1 min followed by a final extension of 10 min for ITS and 13 min for β-tubulin at 72°C. PCR products were sent to (Biokart India Pvt. Ltd., Bangalore, India) for purification and sequencing.

Phylogenetic Analyses

A consensus sequence for each locus of ITS and partial β-tubulin gene was generated using the forward and reverse sequence with CAP3 programme[1] (Huang and Madan, 1999). Newly generated sequences have been deposited in GenBank[2] (**Table 1**). A similarity search was performed for the newly generated sequences using BLASTn of the National Centre for Biotechnology Information.[3] The homologous sequences for each isolate belonging to the corresponding species were retrieved from the GenBank database. The retrieved sequences were aligned using MAFTT v 7[4] (Katoh and Standley, 2013) with default parameters and alignments were edited with BioEdit v7.2.5.0 (Hall, 1999) and were manually adjusted as and when required. The low-quality bases at the start and end positions of the sequences were removed.

The retrieved homologous sequences for each genetic marker were aligned separately with ClustalX2 v2.1 (Thompson et al., 1997) using default parameters. Phylogenetic analyses of individual loci were performed by maximum parsimony (MP) method using PAUP v 4.0 beta win (Swofford and Sullivan, 2003) (heuristic search option with 1,000 random taxon additions and tree bisection and reconnection (TBR) as the branch swapping algorithm). The efficacy and the robustness

TABLE 1 | Section-wise details of *Aspergillus* isolates used in the current study, GenBank accession numbers of ITS and β-tubulin gene sequences.

Sl. No.	FGM lab isolate code	Section	Anamorph/ Teleomorph	Accession number (ITS)	Accession number (β-tubulin)
1	51	*Terrei*	*Aspergillus aureoterreus*	–	MN791093
2	18		*Aspergillus aureoterreus**	–	MN791096
3	1		*Aspergillus aureoterreus*	–	MN791109
4	16		*Aspergillus terreus**	MN392907	MN791095
5	52		*Aspergillus terreus*	MN264636	–
6	61	*Flavi*	*Aspergillus flavus**	–	MN791106
7	22		*Aspergillus tamarii**	MN326529	MN791098
8	75		*Aspergillus tamarii*	MN317364	MN791108
9	S44		*Aspergillus tamarii*	–	MN791115
10	63		*Aspergillus tamarii*	MN263247	–
11	24	*Fumigati*	*Aspergillus fumigatus*	–	MN791099
12	37		*Aspergillus fumigatus*	MN317367	MN791100
13	31		*Aspergillus fumigatus**	MN264637	MN791103
14	17	*Nidulantes*	*Aspergillus nidulans*	–	MN791097
15	45		*Aspergillus nidulans**	MN309877	MN791101
16	9		*Aspergillus nidulans*	–	MN791102
17	71		*Aspergillus nidulans*	MN317365	MN791107
18	2		*Aspergillus nidulans/Emericella nidulans*	–	MN791116
19	49		*Aspergillus quadrilineatus/ Emericella quadrilineata*	–	MN791104
20	58		*Aspergillus quadrilineatus*/ Emericella quadrilineata*	–	MN791105
21	4		*Aspergillus latus**	–	MN791110
22	11		*Aspergillus latus*	–	MN791111
23	12		*Aspergillus latus*	–	MN791112
24	S8	*Nigri*	*Aspergillus awamori*	–	MN791113
25	S24		*Aspergillus awamori**	–	MN791114
26	83	*Versicolores*	*Aspergillus sydowii**	MN298848	–
27	29	Unassigned	*Aspergillus* species	MN294688	–

**Isolates subjected to macro- and micro-morphological analyses.*
ITS, Internal Transcribed Spacer gene; FGM Laboratory, Fungal Genetics and Mycotoxicology Laboratory; All fungal isolates were isolated, cultured and studied by Pranab Kumar Mahata and Regina Sharmila Dass.

[1] http://doua.prabi.fr/software/cap3

[2] https://www.ncbi.nlm.nih.gov/genbank/

[3] https://blast.ncbi.nlm.nih.gov/Blast.cgi

[4] https://mafft.cbrc.jp/alignment/server/

of the parsimonious trees were evaluated with 1,000 bootstrap replications (Hillis and Bull, 1993). Further, the maximum likelihood (ML) method was used for the analysis of individual loci via the CIPRES Science Gateway.[5] Analysis was carried out using RA_xML v. 7.4.2 (Stamatakis et al., 2012). A general time-reversible model (GTR) was applied, including estimation of invariable sites and a discrete gamma distribution with four rate classes (GTRG + I). The reproducibility of the branches was evaluated by bootstrap analysis with 1,000 replicates.

The trees that were being generated were envisaged with FigTree v1.4.0 (Rambaut and Drummond, 2009). Bootstrap (bs) percentages of analysis were labeled at the nodes. Values less than 70% bs were removed. Branches with values 95% and above bs were thickened.

Single Nucleotide Polymorphisms

Each of the sequences of the fungal isolates listed in **Table 1** were aligned with the reference genomes of respective species that are available on the NCBI using BLAST algorithm. Variants observed in these sequences were fetched manually and the corresponding variants on the protein sequences were annotated manually from the protein annotations of β-tubulin genes of the respective species.

Statistical Analysis

To learn more about the *Aspergillus* fungal isolates, we employed descriptive statistics on the data. The statistical studies were carried out using Origin 8.6 (Northampton, United States).

RESULTS

Relative Abundance in *Foeniculum vulgare*

As shown in **Figure 1** among all the isolates collected from *F. vulgare* (n = 70), we were able to identify fourteen (14) different species of the *Aspergillus*. *Aspergillus niger* (32.24%) ranked first, followed by *A. flavus* (11.63%), *A. terreus* (2.6%), *A. nidulans* (1.65%), *A. tamarii* (0.97%), *A. species* (0.96%), *Emericella quadrilineata* (0.45%), *A. fumigatus* (0.43%), *A. latus* (0.4%), *A. aureoterreus* (0.37%), *A. awamori* (0.37%), *A. brasiliensis* (0.29%), *A. ochraceous* (0.06%), and *A. sydowii* (0.03%).

Aspergillus Diversity

The cultivation of fungi from seventy (70) fennel samples yielded a total of 3,506 *Aspergillus* isolates. Apart from *Aspergilli*, genera like *Alternaria* species, *Curvularia* species, *Drechslera* species, *Fusarium* species, *Mucor* species, *Mycelia sterilia*, *Neurospora crassa*, *Penicillium* species, *Trichothecium* species and Yeasts were also isolated based purely on morphological and microscopical analyses. The detailed cultural morphology and microscopic analyses of the *Aspergilli* has been carried out and has been presented with descriptions (**Figures 2–5**). *A. niger* (Relative abundance: 32.24%) followed by *A. flavus* (Relative abundance: 11.63%) were found to be the two most frequently occurring species. The fourteen (14) *Aspergillus* species were isolated from Indian fennel seeds, collected from different geographical zones of India, and studied in detail using morphological and genetic methods for molecular phylogeny. The Internal Transcribed Spacer gene of the ribosomal DNA (rDNA-ITS) sequencing identified 20 isolates under five *Aspergillus* Sections namely *Terrei*, *Versicolores*, *Flavi*, *Fumigati* and *Nidulantes* (Sub-genus II) and corresponding to the family *Trichocomaceae*. The isolates included *A. nidulans*, *A. quadrilineatus*, *A. miyajii*, *A. rugulosus*, *A. terreus*, *A. tamarii*, *A. fumigatus*, *A. sydowii* and *Aspergillus* species (Section Unassigned). Isolates sequenced for β-tubulin gene sequences revealed the identity of 23 *Aspergilli* with species namely *A. nidulans*, *A. quadrilineatus*, *A. latus*, *A. terreus*, *A. aureoterreus*, *A. flavus*, *A. tamarii*, *A. fumigatus* and *A. awamori* from five sections namely *Nidulantes*, *Terrei*, *Flavi*, *Fumigati*, and *Nigri* (Sub-genus I) with no known teleomorphs. The *Aspergilli* namely *A. quadrilineatus*, *A. latus*, *A. aureoterreus*, *A. awamori* and *Aspergillus* species are being reported for the very first time from fennel seeds. The details of the GenBank depositions are also shown (**Table 1**).

Phylogenetic Analyses

The PCR products of the ITS (Amplicon size: 600-bp) and β-tubulin (Amplicon size: 400-bp) genetic markers were analyzed (**Figures 6A,B**). Based on sequence analysis of ITS 1, ITS 4 and β-tubulin genes, eleven (11) monophyletic groups of *Aspergilli* have been well resolved into six Sections: *Terrei*, *Flavi*, *Fumigati*, *Nidulantes, Nigri* and *Versicolores*. The ITS marker, which is the most suitable, profoundly recommended and widely accepted fungal barcode (Schoch et al., 2012) served handy in the recognition, identification, and classification of species in their respective clades. Both ITS regions and β-tubulin gene phylogenetic trees are depicted in the phylograms. Data creation and post-sequencing of *Aspergillus* genes were examined, with results showing that the β-tubulin-based phylogenetic trees had a greater level of precision than those generated by the ITS regions. Analysis of β-tubulin-based phylogenetic trees for our isolates namely *Aspergillus aureoterreus* (**Figure 7**), *Emericella quadrilineata* (**Figure 8**), *Aspergillus latus* (**Figure 9**), and *Aspergillus awamori* (**Figure 10**) showed higher resolution. In addition, ITS-based phylogenetic trees for *Aspergillus terreus*, *Aspergillus nidulans*, *Aspergillus sydowii*, and *Aspergillus* sp., as well as β-tubulin-based phylogenetic trees for isolates of *Aspergillus flavus*, *Aspergillus tamarii*, and *Aspergillus fumigatus*, were created.

Isolates studied in the present analysis, namely, *A. aureoterreus* FOEVPRB18 (**Figure 7**) was closely associated with the type strain *A. aureoterreus* CMV010F6 (MK451161), which has been deposited very recently at NCBI by researchers, while conducting studies to update the taxonomical status of *Aspergillus* species from South Africa (Visagie and Houbraken, 2020). The type strain *A. terreus* CVS503 65 (EU147717), was the most closely related species with our isolate (**Figure 7**), known to be a notorious fungus, playing a key role in human opportunistic pathogenesis, unpublished (Balajee et al., 2009). While *A. aureoterreus* CMV010F6 displayed 99.37% identity

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

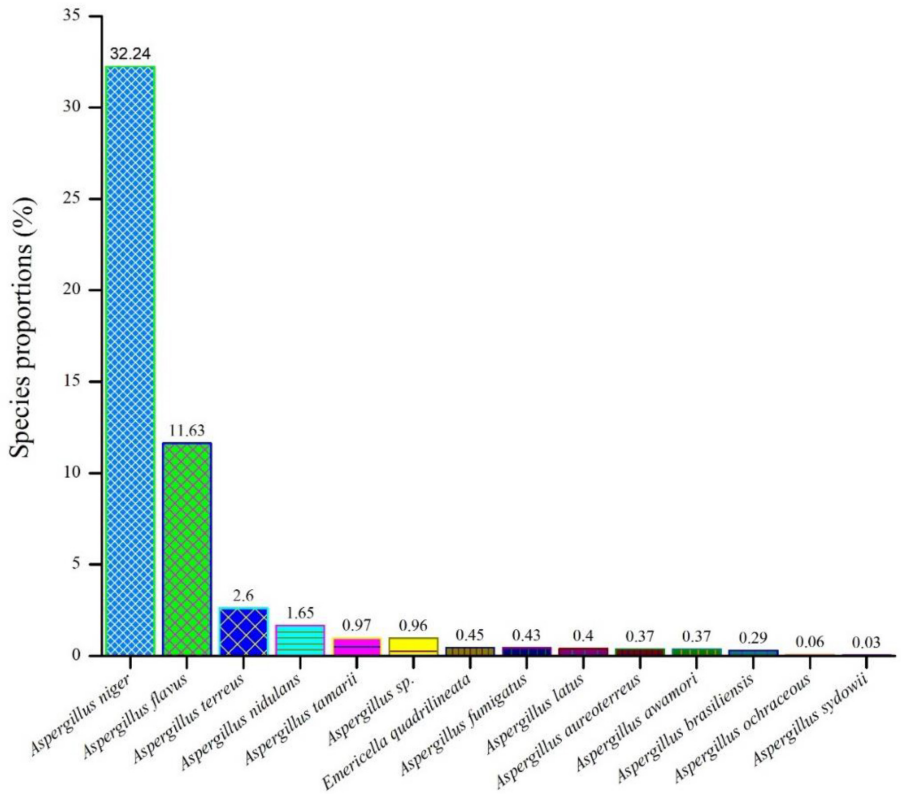

FIGURE 1 | An assessment of the relative abundances of several *Aspergillus* species in the *F. vulgare* samples.

with our isolate *A. aureoterreus* FOEVPRB18, *A. terreus* CVS503 65 showed 96.32% identity.

Similarly, the clades *A. flavus* and *A. tamarii* were placed under the *Flavi* Section. *A. flavus* FOEVPRB61 occurred as a separate taxon although it was closely related to previously described clades of the same species (**Figure 11**). Partial ITS and β-tubulin gene sequences were used to arrive at species identity of the 27 *Aspergillus* isolates. Sequences from both the genetic markers gave almost the same identification. However, a doubtful analysis was experienced with *Aspergillus* isolate (ITS Barcode: MN294688; Section Unassigned), in determining a species name (**Figure 12**). Nevertheless, a detailed macro- and micro-morphological analysis of the unassigned fungal isolate is available. Although it emerged as a separate taxon during phylogenetic analysis, it showed 100% bs value (ML method) with *Aspergillus* sp. SS 30 1 studied by another research team from India, Direct Submission, Genetics, Agharkar Research Institute, Maharashtra, India (Chavan, 2020). This species was studied with *Penicillium* species clone 1 4 TSS as an outgroup. In most instances, our isolates were closely related to type strains which have been previously studied and described by providing appreciable support values.

In the Section *Fumigati* our isolate namely *A. fumigatus* FOEVPRB31 (**Figure 13**) was found to be a close relative of *A. fumigatus* 3, Direct Submission, Department of Dermatology, Nanjing University, China (Chen et al., 2018) and *A. fumigatus* CMXY2075, Direct Submission, Ecology of Clinical Fungi, Fungal Biodiversity Institute, Netherlands (Chen and Xu, 2018).

A. nidulans, *A. quadrilineatus* and *A. latus* were conveniently assigned in their clades under the *Nidulante*s Section. *A. nidulans* is phenotypically very similar to *A. quadrilineatus* except for differences in ascospore morphology (Chen et al., 2016). With respect to their phylogeny too they were situated in closely related clades, emerging from that of *E. nidulans* which is the sexual morph of *A. nidulans* (**Figure 14**). The sexual stages were appreciably documented in our isolate also, as realized by micro-morphological analyses. However, the concept of naming *Aspergillus* species with their sexual morphs is gradually becoming obsolete, due to major decisions taken at the meeting of the International Commission for *Penicillium* and *Aspergillus* (ICPA) in 2012.[6] *A. quadrilineatus* and *A. latus*, which shared 99.76% identity with the type strain *Emericella quadrilineata* IFM 42006 (AB248335.1) and 99.55% identity with the type strain DTO 047 H2 (KU866810.1) respectively, were classified into well-defined clades with *A. quadrilineatus*, which formed discrete and conspicuous sexual stages and therefore is referred to as *Emericella quadrilineata* (Teleomorph). This was also in support of data generated through phylogenetic analysis of β-tubulin gene

[6]http://www.aspergilluspenicillium.org/

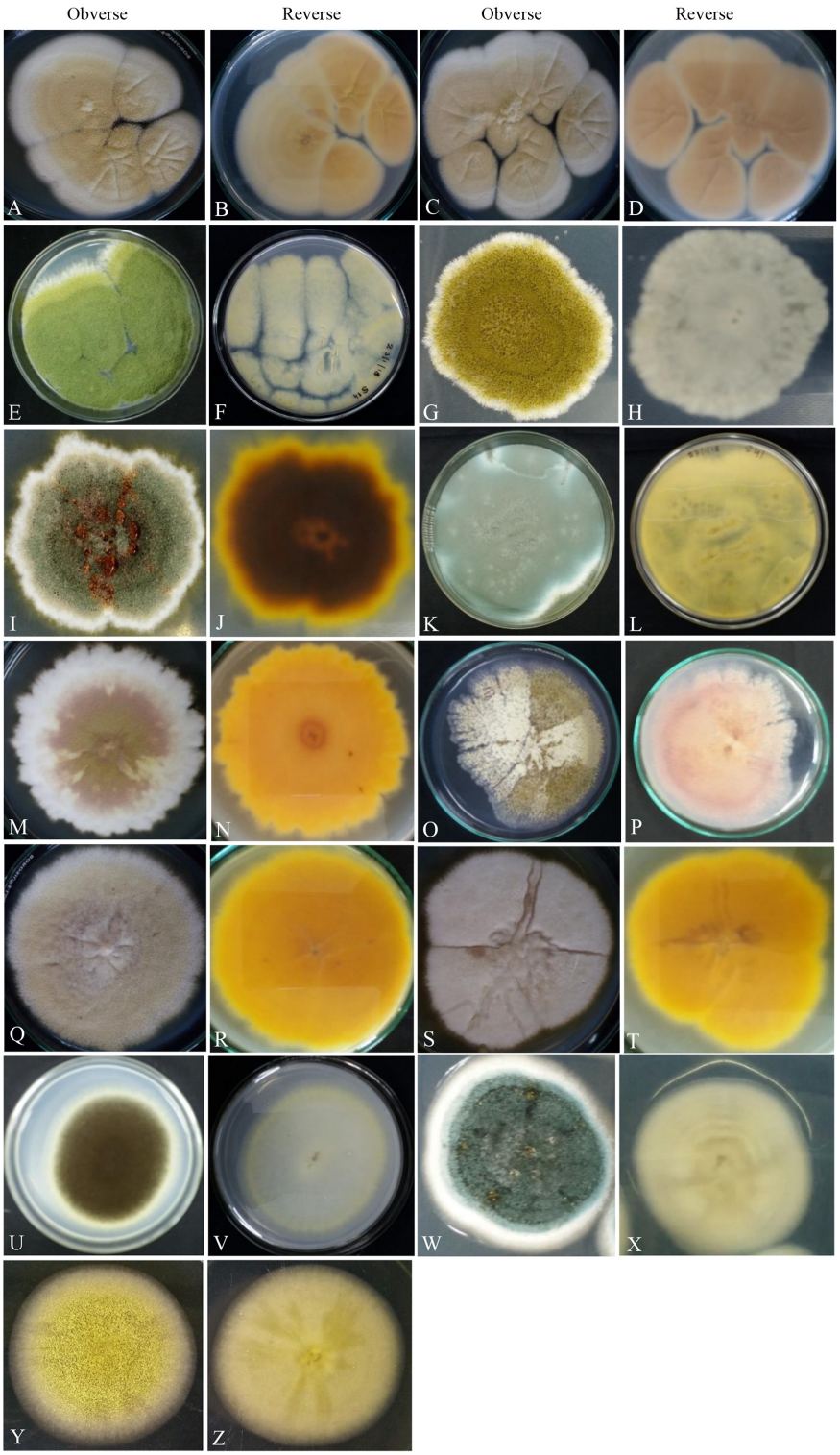

FIGURE 2 | Photographs illustrating the macromorphology of strains of *Aspergillus aureoterrus* **(A,B)** (Acc. No. MN791096); *Aspergillus terreus* **(C,D)** (Acc. No. MN791095); *Aspergillus flavus* **(E,F)** (Acc. No. MN791106); *Aspergillus tamarii* **(G,H)** (Acc. No. MN791098); *Aspergillus fumigatus* **(I,J)** (Acc. No. MN791103); *Aspergillus fumigatus* **(K,L)** (Acc. No. MN791100); *Aspergillus nidulans* **(M–P)** (Acc. No. MN791101); *Aspergillus quadrilineatus/Emericella quadrilineata* **(Q,R)** (Acc. No. MN791105); *Aspergillus latus* **(S,T)** (Acc. No. MN791110); *Aspergillus awamori* **(U,V)** (Acc. No. MN791114); *Aspergillus sydowii* **(W,X)** (Acc. No. MN298848), and *Aspergillus* species **(Y,Z)** (Acc. No. MN294688) are shown. Images **(O,P)** of *Aspergillus nidulans* were incubated in CDA for 7 days, whereas all other isolates were kept in PDA.

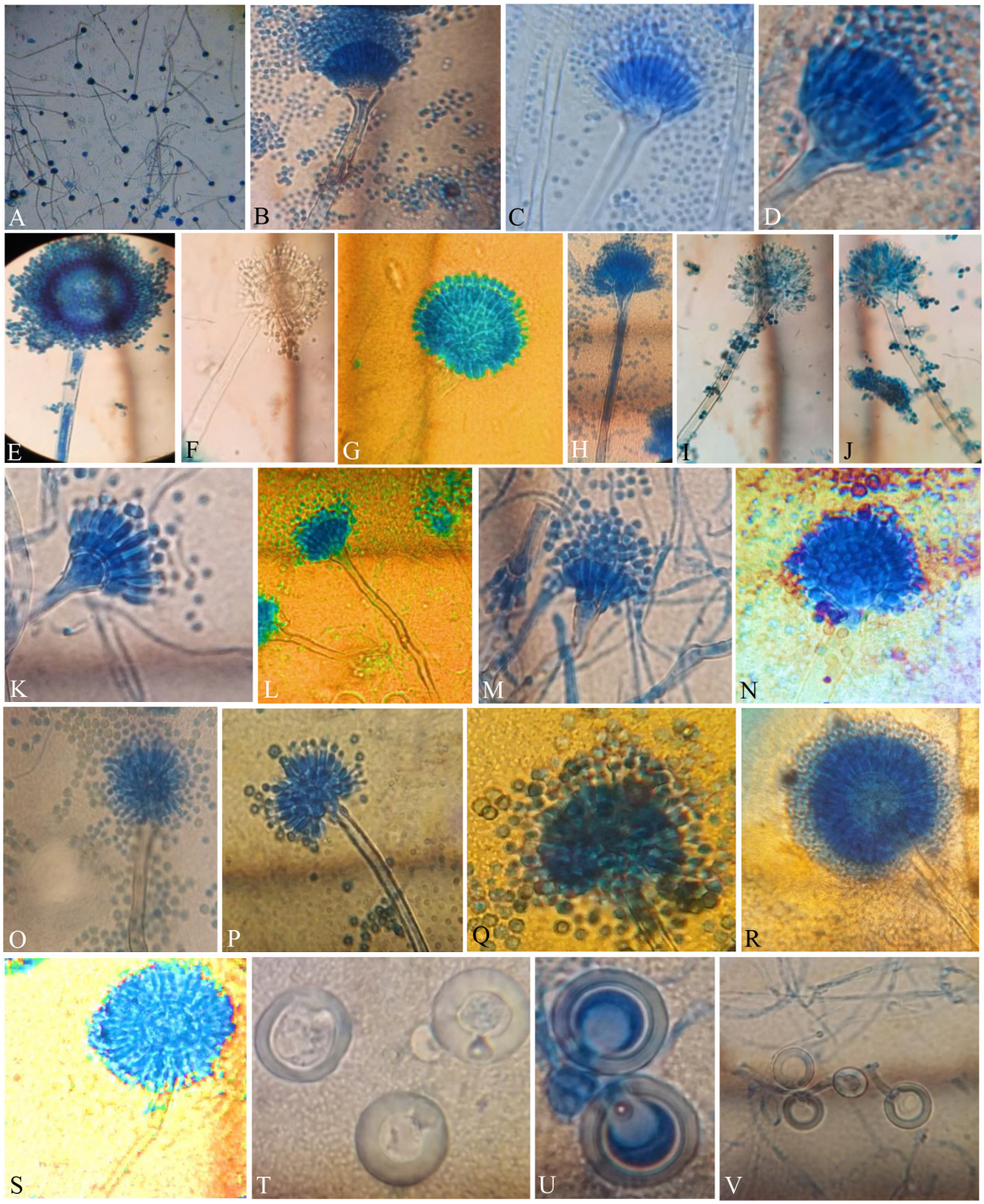

FIGURE 3 | Compound Light Microscope (CLM) images of the hyphae, conidiophores, and Hülle cells, **(A,B)** hyphae and elongated conidiophores of *A. aureoterreus*; **(C,D)** enlarged conidiophores of *A. terreus*; **(E,F)** stained and unstained conidiophores of *A. flavus*; **(G)** enlarged conidial head of *A. tamarii*; **(H–J)** conidiophores with bottle-shaped vesicles, uniseriate sterigmata, and conidia of *A. fumigatus*; **(K)** enlarged conidiophores of *A. nidulans*; **(L)** conidiophores of *E. quadrilineata*; **(M)** conidiophores of *A. latus*; **(N,O)** *A. awamori* conidiophores; **(P,Q)** *A. sydowii* conidiophores; **(R,S)** *Aspergillus* species conidiophores, and **(T–V)** *A. nidulans*, *E. quadrilineata*, and *A. latus* Hülle cells, respectively.

(**Figure 8**) being closely related to *Emericella quadrilineata* IFM 42006 (AB248335.1) studied previously (Matsuzawa et al., 2006). *A. latus* was found to emerge as a separate taxon (**Figure 9**) from *A. quadrilineatus* DTO 048 A8, Direct Submission, CBS-KNAW Fungal Biodiversity Centre (Chen et al., 2016) and displayed 100% bs value with *Emericella foveolata* IFM 54285

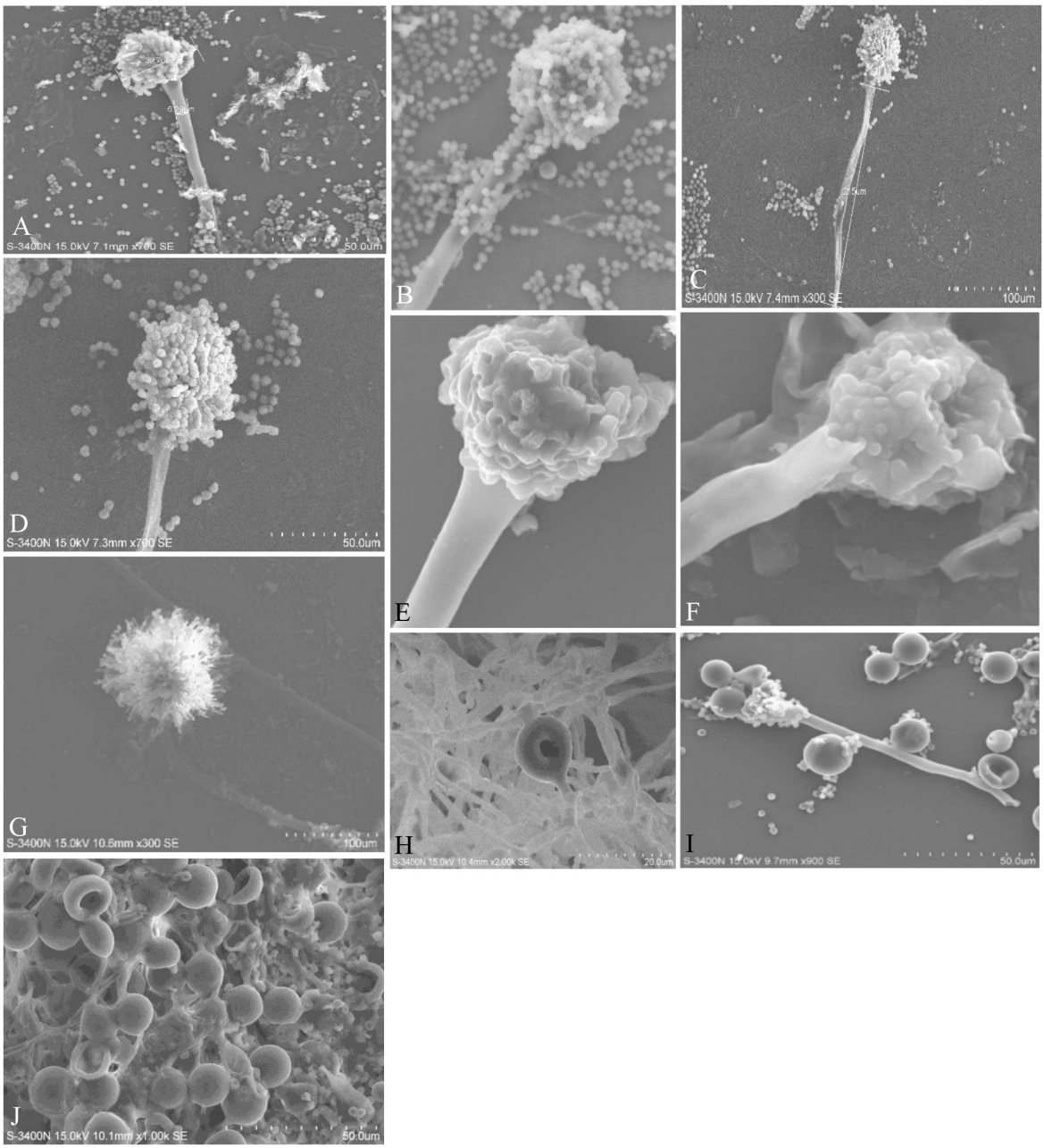

FIGURE 4 | Conidiophores, vesicles, metulae, phialides, conidia, and Hülle cells of the strains are shown in Scanning Electron Microscope (SEM) images. **(A)** Conidiophores of *A. aureoterreus*; **(B)** conidiophores of *A. terreus*; **(C,D)** conidial head with chains of conidia of *A. tamarii*; **(E)** conidiophores of *A. nidulans*; **(F)** conidiophores of *A. latus*; **(G)** globose conidial heads, with metulae, phialides, chains of conidia and smooth conidiophores in *A. awamori*; **(H,I)** mycelium and Hülle cell development and Hülle cells of *A. nidulans* **(J)** Hülle cells formation in *E. quadrilineata.*

and *E. foveolata* IFM 42015 (Matsuzawa et al., 2006), which has also been studied under the Section *Nidulantes* while discerning evolutionary relatedness.

Aspergillus awamori S24 BD02 isolated from the fennel samples diverged as a clade closely related to *A. awamori* Mal02, Direct Submission, Department of Biotechnology, University of Verona, Italy (Andreolli et al., 2018), which in turn were related to *A. niger* studied collectively as black *Aspergilli*. Only one isolate namely *A. sydowii* FOEV83 was studied under the Section *Versicolores*. Studies inferred by phylogenetic analyses showed its close relatedness to *A. sydowii* DUCC5715, Direct Submission, Department of Microbiology, Dankook University, South Korea (Ahn et al., 2020), and a 100% bs value with *A. versicolor* M4 C4, Direct Submission, Postgraduate in Biological Sciences, University of Nayarit, Mexico (Bobadilla-Carrillo et al., 2020). This was especially true with cultural and morphological studies,

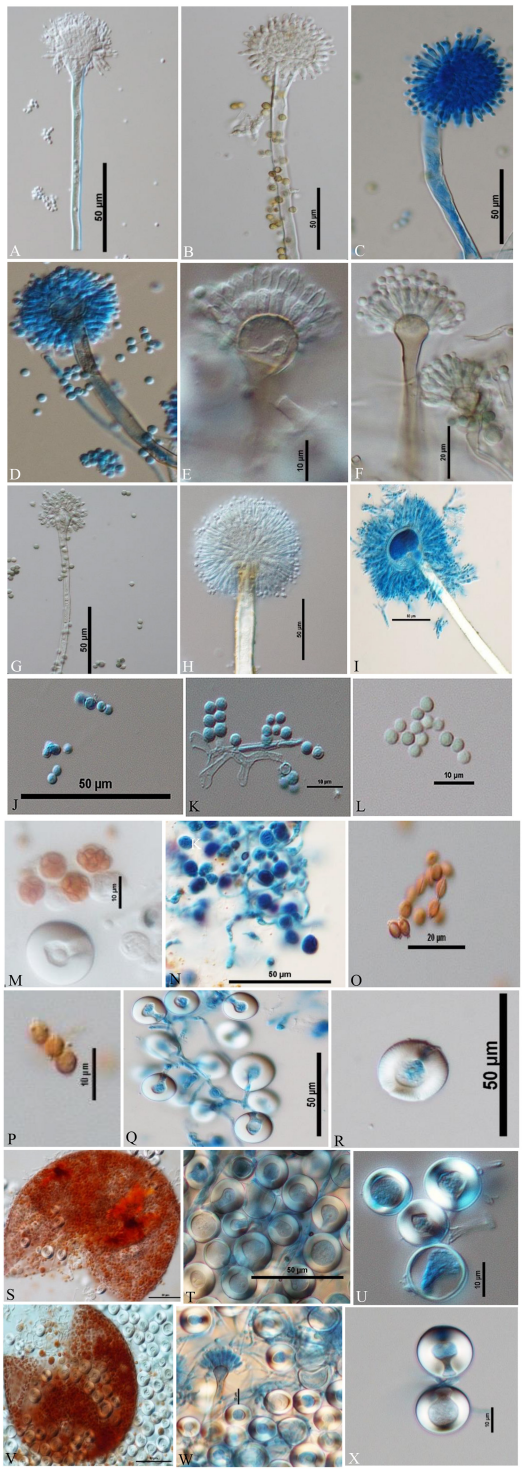

FIGURE 5 | Conidiophores, conidia, asci, ascospores, cleistothecia, and Hülle cells Differential Interference Contrast (DIC) images of *Aspergilli* **(A)** *A. aureoterreus* smooth, long, colourless conidiophores and columnar, biseriate conidial heads with hemispherical vesicles, metulae phialides, and conidia; **(B,C)** *A. tamarii* conidiophores containing globose, radiating and rough conidia, globose vesicles with uniseriate sterigmata; **(D)** *A. nidulans* smooth-walled, sinuate conidiophores and columnar conidial heads with small *(Continued)*

FIGURE 5 | hemispherical vesicles, metulae, phialides, and conidia; **(E)** *E. quadrilineata* smooth, sinuate conidiophores and short, columnar conidial heads with hemispherical vesicles, metulae, phialides, and conidia; **(F)** *A. latus* smooth-walled, sinuate conidiophores and columnar conidial heads with small hemispherical vesicles, metulae, phialides, and conidia; **(G)** *A. sydowii* smooth, sinuous conidiophores and hemispherical conidial heads with globose to elliptical vesicles, metulae, phialides, and conidia; **(H,I)** *Aspergillus* species conidiophores with rough, pitted surfaces, globose conidial heads with globose, thinner vesicles, metulae, phialides, and conidia; **(J–L)** globose conidia of *A. nidulans*, *E. quadrilineata*, and *A. latus*, respectively; **(M,N)** asci of *A. nidulans* and *A. latus*; **(O)** lenticular ascospores of *A. latus*; **(P)** lenticular, smooth walled (unstained) ascospores of *E. quadrilineata*; **(Q,R)** Hülle cells of *A. nidulans*; **(S–U)** ruptured cleistothecium and Hülle cells of *E. quadrilineata*, and **(V)** in *A. latus*, ruptured cleistothecium showing asci, surrounded by Hülle cells, **(W,X)** Hülle cells separated from *A. latus*.

where *A. sydowii* FOEV83 appeared very similar to *A. versicolor* by the characteristic pigmentation.

Single Nucleotide Polymorphisms

In order to study the variants in the β-tubulin genes, the sequences obtained from Sanger sequencing were aligned to the reference genome sequences of respective species. The alignment was carried out using BLASTn algorithm. The variants such as single nucleotide variants, substitutions, insertions, and deletions were identified manually. Based on the regions of coding sequence (CDS) of β-tubulin gene provided at the protein feature annotations, the effect of these mutations at protein level were manually evaluated. The site of variations at the genome and protein level are tabulated (**Table 2**). Once these sequences for each species had been aligned, we observed that they all share a high degree of identity and displayed few variants with the exceptions of MN791095, MN791096, MN791109, and MN791093 sequences. We found that five species MN791107: *A. nidulans* 4, MN791105: *E. quadrilineata* 2, MN791112: *A. latus* 3, MN791108: *A. tamarii* 2, and MN791099: *A. fumigatus* 1 did not exhibit variations with regard to the reference genome in their respective sections. Hence, these five species have been excluded (**Table 2**). Our SNP analysis showed a high number of intronic variances and protein coding synonymous variants (changes without coding sequences) that result in synonymous mutations. On the contrary, four non-synonymous (substituting amino acids with new ones) variants were identified- MN791102: *A. nidulans* 3, MN791116: *E. nidulans* 5, MN791104: *E. quadrilineata* 1, and MN791110: *A. latus* 1, all contain single nucleotide variations (SNVs) that alter protein sequences whose functional consequences has not been studied.

The SNP profiling analyses revealed five mis-sense mutations in the CDS (coding DNA sequence) regions of beta-tubulin gene in four *Aspergillus* isolates. Histidine was replaced by tyrosine at the 105th position in *A. nidulans* MN791102, leucine to proline and threonine to serine at positions 30 and 107, respectively, in *E. nidulans* MN791116. In *E. quadrilineata* MN791104 leucine was replaced to proline at the 10th position. Furthermore, analysis of *A. latus* MN791110 indicated that the amino acid asparagine was substituted with isoleucine at the 87th position.

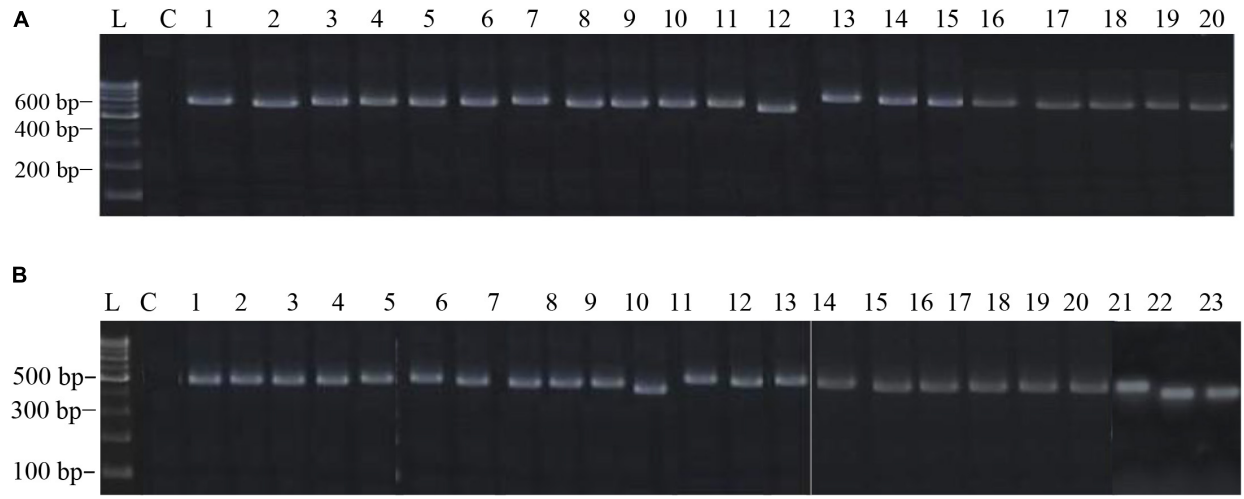

FIGURE 6 | (A) The electrophoretogram shows the positive PCR amplification of *Aspergillus* ITS genes with species-specific amplicons. Here, L, 100-bp MBT049 HIMEDIA DNA ladder; C, PCR with a negative control; 1, 49-*Aspergillus nidulans*; 2, 58-*Aspergillus nidulans*; 3, 63-*Aspergillus tamarii*; 4, 31-*Aspergillus fumigatus*; 5, 37-*Aspergillus fumigatus*; 6, 45-*Aspergillus nidulans*; 7, 51-*Aspergillus terreus*; 8, 52-*Aspergillus terreus*; 9, 71-*Aspergillus nidulans*; 10, 75-*Aspergillus tamarii*; 11, 83-*Aspergillus sydowii*; 12, 17-*Aspergillus miyajii*; 13, 16-*Aspergillus terreus*; 14, 22-*Aspergillus tamarii*; 15, 29-*Aspergillus sp.;* 16, 4-*Aspergillus quadrilineatus*; 17, 2-*Aspergillus rugulosus*; 18, 1-*Aspergillus terreus*; 19, 11-*Aspergillus nidulans*; 20, 12-*Aspergillus nidulans*. **(B)** The electrophoretogram shows the positive PCR amplification of *Aspergillus* β-tubulin genes with species-specific amplicons. Here, L, 100-bp MBT049 HIMEDIA DNA ladder; C, PCR with a negative control; 1, 49-*Emericella quadrilineata*; 2, 58-*Emericella quadrilineata*; 3, 61-*Aspergillus flavus*; 4, 31-*Aspergillus fumigatus*; 5, 37-*Aspergillus fumigatus*; 6, 45-*Aspergillus nidulans*; 7, 51-*Aspergillus aureoterreus*; 8, 71-*Aspergillus nidulans*; 9, 75-*Aspergillus tamarii*; 10, 18-*Aspergillus aureoterreus*; 11, 17-*Aspergillus nidulans*; 12, 16-*Aspergillus terreus*; 13, 9-*Aspergillus nidulans*; 14, 24-*Aspergillus fumigatus*; 15, 22- *Aspergillus tamarii;* 16, 4-*Aspergillus latus*; 17, 2-*Emericella nidulans*; 18, 1-*Aspergillus aureoterreus*; 19, 11- *Aspergillus latus*; 20, 12-*Aspergillus latus*; 21, S8-*Aspergillus awamori*; 22, S24-*Aspergillus awamori*; 23, S44-*Aspergillus tamarii*.

DISCUSSION

Spices occupy a very prominent role in the culinary preparations in India (Siruguri and Bhat, 2015). Based on the limited information and scientific evidence available through literature, fennel continues to be the most widely used herb in traditional medicine. *F. vulgare* is known to exhibit an array of pharmacological properties such as antimicrobial, antiviral, anti-inflammatory, anti-mutagenic, antipyretic, anti-spasmodic, anti-thrombotic, apoptotic, cardiovascular, chemomodulatory, anti-tumor, hepatoprotective, and hypoglycemic. Also, they've been linked to memory enhancement (Badgujar et al., 2014), which implies that they could be useful in the field of pharmaceutical biology for the development of new drugs to treat a wide range of ailments.

The use of spices in food has been practiced since time immemorial. Black pepper, cardamom, clove, coriander, cumin, ginger and other spices have been explored for their intrinsic antibacterial effects when used in foods, in addition to the flavors and aromas that they produce. In spite of the antibacterial compounds that they harbor, they are vulnerable to mycological deterioration, constantly being colonized by mycoflora like the *Aspergilli*, *Penicillia,* and *Fusaria* during pre- and post-harvest operations at the field.

Accurate identification of *Aspergillus* isolates up to the species level has become ecologically, epidemiologically, and pathologically significant because of the extensive damage that they cause in food crops (Battilani et al., 2016; Alshannaq and Yu, 2017), feed crops (Ráduly et al., 2020), and spices (Ali et al., 2015; Ssepuuya et al., 2018). Specific studies focused on mycological analyses have been conducted by research groups on spices like cumin, clove, black pepper, ginger, cardamom, coriander and cinnamon with dominancy of *Aspergillus flavus* and *A. niger,* with cumin being the most contaminated sample (Elshafie et al., 2002). Notable contamination by *A. flavus*, *A. fumigatus*, *A. alutaceus*, *A. niger*, and *A. sulphureus* has been observed in yet another study conducted using aniseed, rosemary and spice products (Ahene et al., 2011). Another research group (Hammami et al., 2014) examined fourteen distinct spice samples and reported the highest fungal contamination by *Aspergillus flavus*, *A. nomius,* and *A. niger*. The studies conducted using fennel seeds from India (Kulshrestha et al., 2014) have been sparse, with reports on *Aspergillus* contamination being investigated from fennel seeds, cumin and coriander seeds which are extensively used in culinary preparations in India. Two other research studies (Azzoune et al., 2015; Garcia et al., 2018) have been conducted in order to evaluate the mycotoxigenic potential of select *Aspergilli*, from nearly about two hundred spice samples of rosemary, fennel, cinnamon, clove, pepperoni, black and white pepper and oregano. Our study is the first of its kind to reveal a significant degree of *Aspergillus* contamination in the Indian subcontinent, with *A. niger* being the most frequently isolated species, followed by *Mycelia sterilia*, *A. flavus*, *Mucor* species, *Penicillium* species, and Yeasts, among others.

The *Aspergilli* are an all-pervasive and species substantial genus, containing more than three hundred thread-like filiform fungi (Anonymous, 2021). By far, the majority of *Aspergilli*

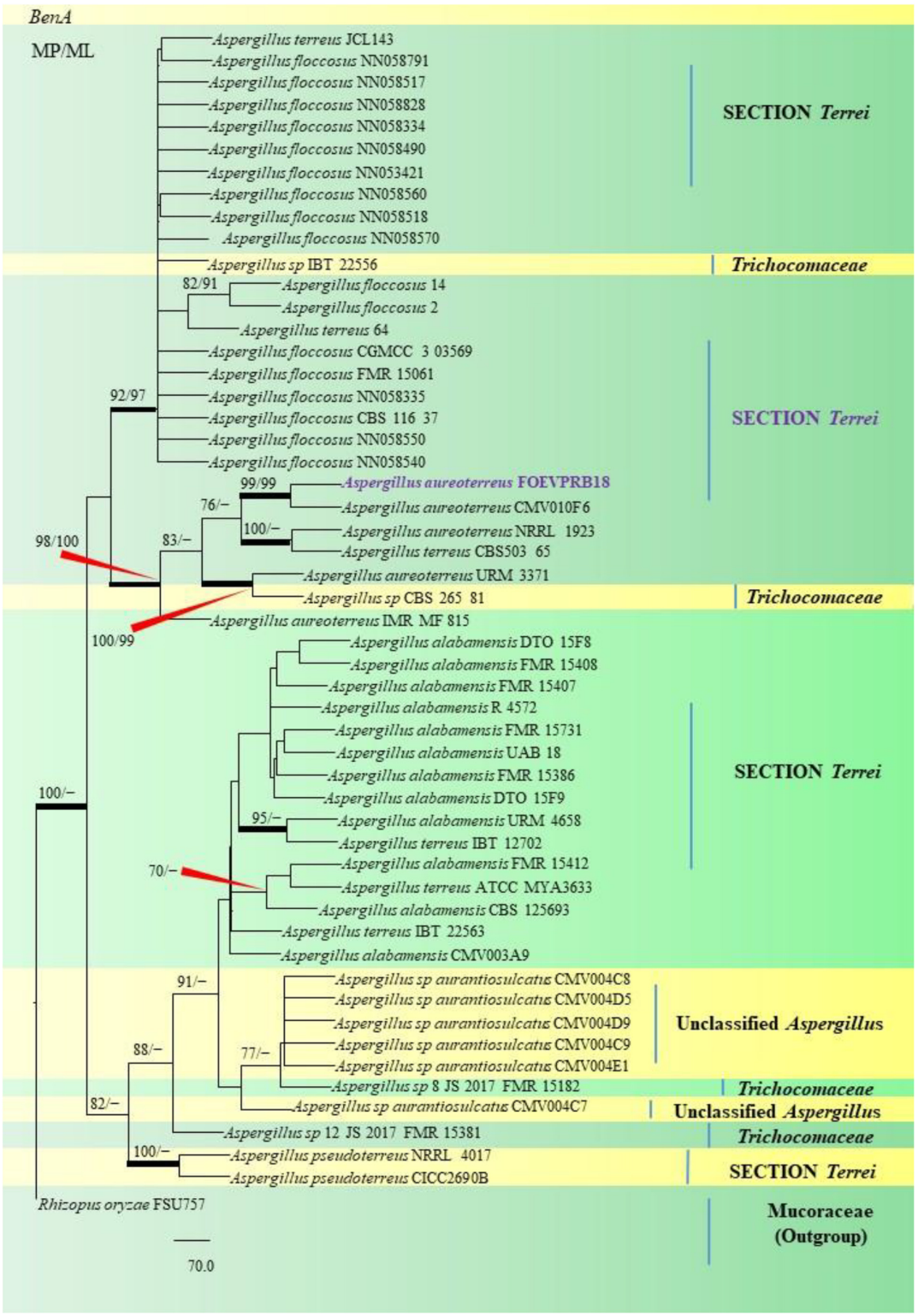

FIGURE 7 | A phylogenetic tree (taxon branching diagram) presents the genetic analysis of the *Aspergillus aureoterreus* FOEVPRB18 clade belonging to the Section *Terrei*, exhibiting a maximum parsimony analysis of partial β-tubulin concatenated sequences, acquired. Bootstrap (bs) values ≥ 70% are designated at the nodes for maximum parsimony (MP) and maximum likelihood (ML) methods; branches with values ≥ 95% bs are bold black. A specific isolate code was mentioned as a suffix to the species name. Hyphen (-) represents the range of support values for parsimony and RAxML analyses of below 70%. Our isolate, which is marked in purple, was rooted using *Rhizopus oryzae* FSU757.

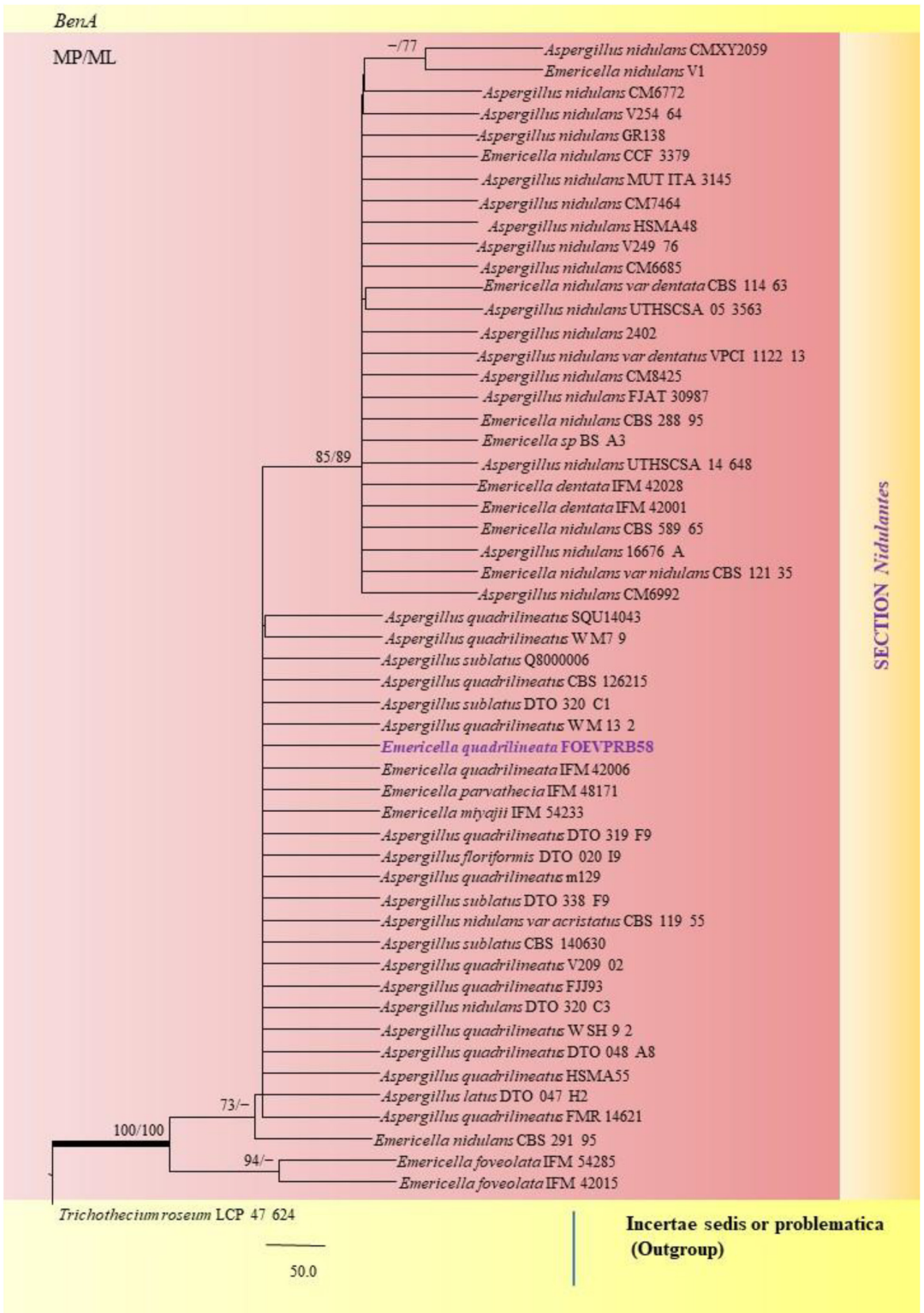

FIGURE 8 | A phylogenetic tree (taxon branching diagram) which presents the genetic analysis of the *Emericella quadrilineata* FOEVPRB58 clade belonging to the Section *Nidulantes,* exhibiting a maximum parsimony analysis of partial β-tubulin concatenated sequences, acquired. Bootstrap (bs) values ≥ 70% are designated at the nodes for maximum parsimony (MP) and maximum likelihood (ML) methods; branches with values ≥ 95% bs are bold black. A specific isolate code was mentioned as a suffix to the species name. Hyphen (-) represents the range of support values for parsimony and RAxML analyses of below 70%. Our isolate, which is marked in purple, was rooted using *Trichothecium roseum* LCP 47 624.

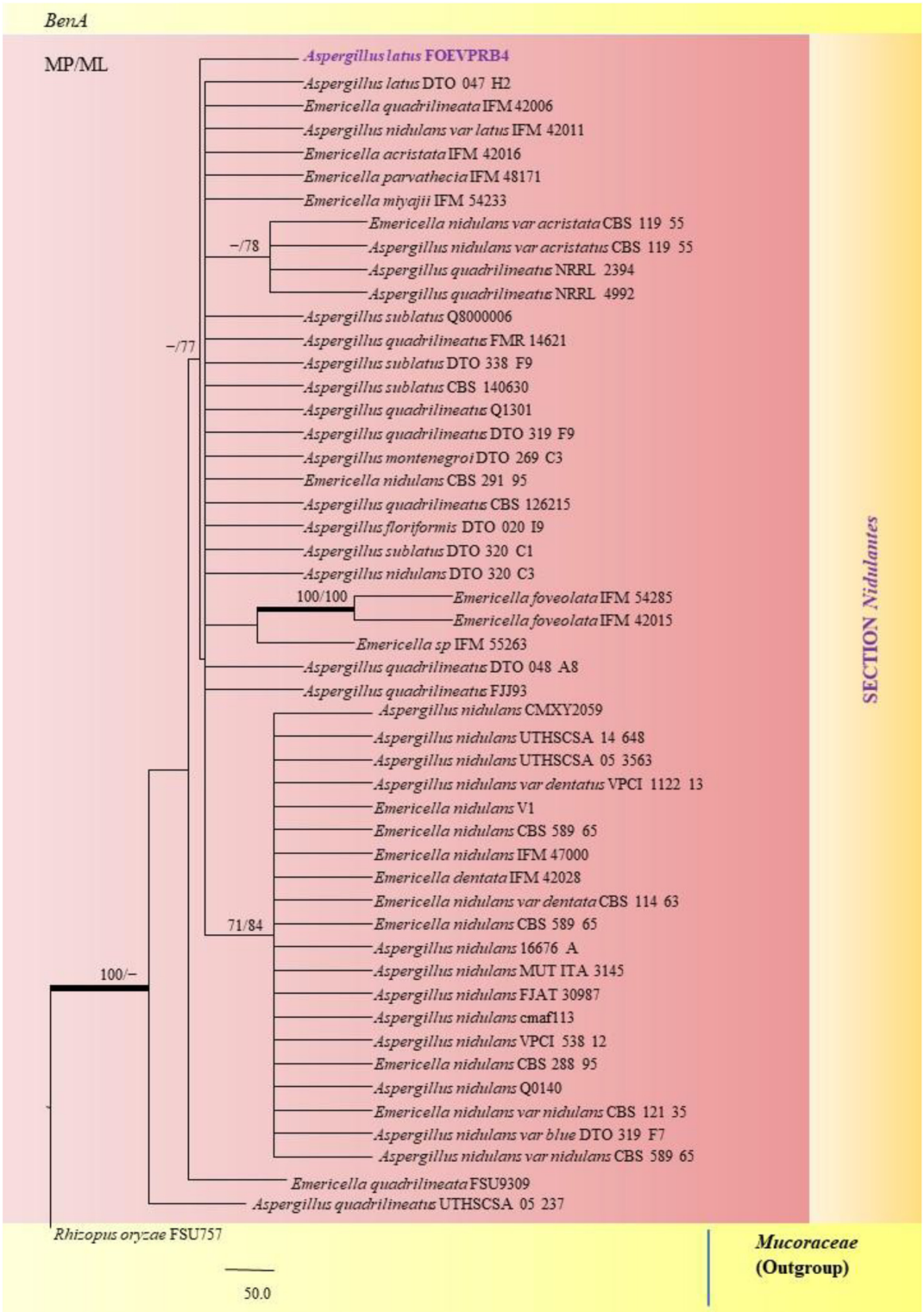

FIGURE 9 | A phylogenetic tree (taxon branching diagram) which presents the genetic analysis of the *Aspergillus latus* FOEVPRB4 clade belonging to the Section *Nidulantes,* exhibiting a maximum parsimony analysis of partial β-tubulin concatenated sequences, acquired. Bootstrap (bs) values ≥ 70% are designated at the nodes for maximum parsimony (MP) and maximum likelihood (ML) methods; branches with values ≥ 95% bs are bold black. A specific isolate code was mentioned as a suffix to the species name. Hyphen (-) represents the range of support values for parsimony and RAxML analyses of below 70%. Our isolate, which is marked in purple, was rooted using *Rhizopus oryzae* FSU757.

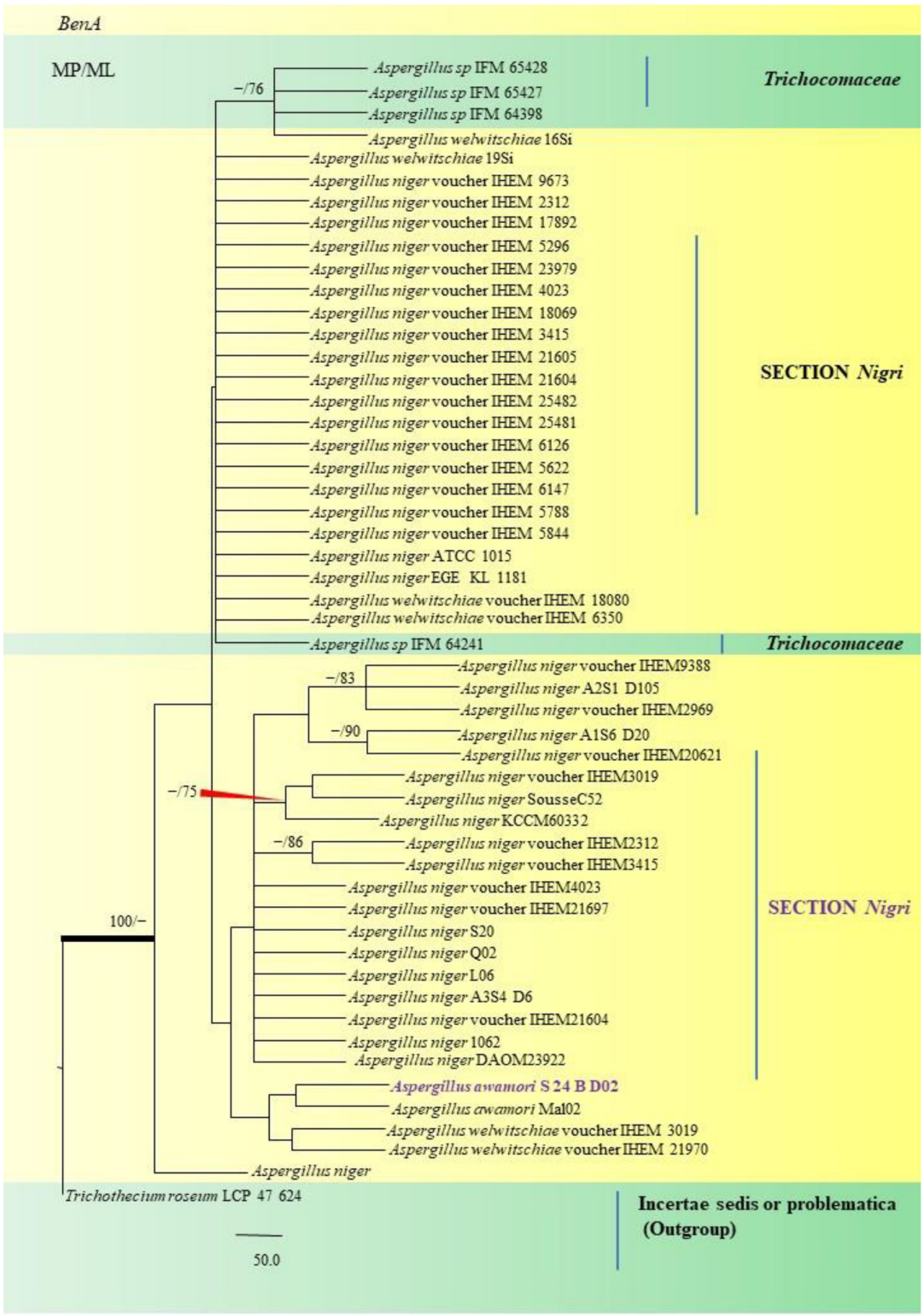

FIGURE 10 | A phylogenetic tree (taxon branching diagram) which presents the genetic analysis of the *Aspergillus awamori* S 24 B D02 clade belonging to the Section *Nigri,* exhibiting a maximum parsimony analysis of partial β-tubulin concatenated sequences, acquired. Bootstrap (bs) values ≥ 70% are designated at the nodes for maximum parsimony (MP) and maximum likelihood (ML) methods; branches with values ≥ 95% bs are bold black. A specific isolate code was mentioned as a suffix to the species name. Hyphen (-) represents the range of support values for parsimony and RAxML analyses of below 70%. Our isolate, which is marked in purple, was rooted using *Trichothecium roseum* LCP 47 624.

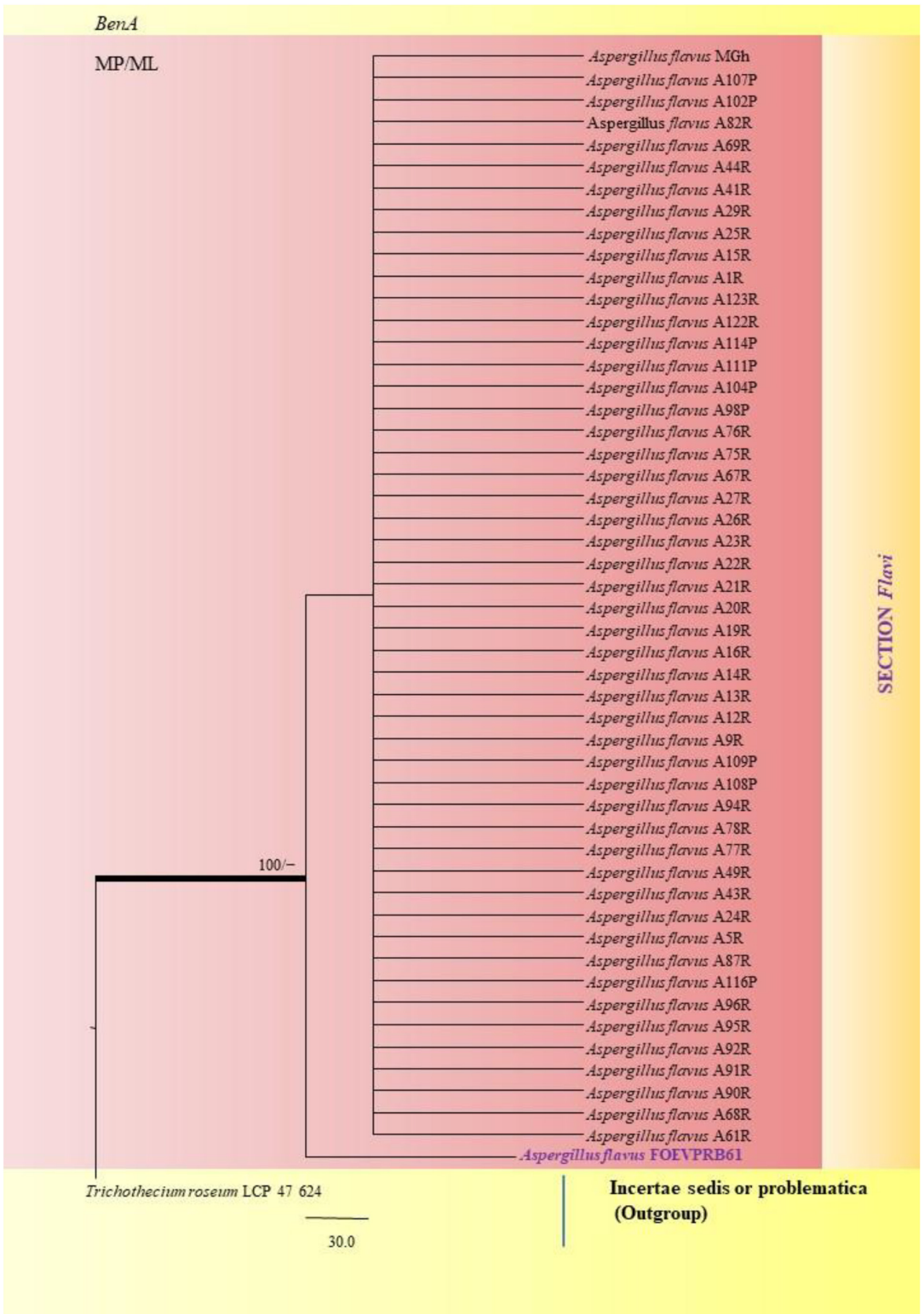

FIGURE 11 | A phylogenetic tree (taxon branching diagram) which presents the genetic analysis of the *Aspergillus flavus* FOEVPRB61 clade belonging to the Section *Flavi*, exhibiting a maximum parsimony analysis of partial β-tubulin concatenated sequences, acquired. Bootstrap (bs) values ≥ 70% are designated at the nodes for maximum parsimony (MP) and maximum likelihood (ML) methods; branches with values ≥ 95% bs are bold black. A specific isolate code was mentioned as a suffix to the species name. Hyphen (-) represents the range of support values for parsimony and RAxML analyses of below 70%. Our isolate, which is marked in purple, was rooted using *Trichothecium roseum* LCP 47 624.

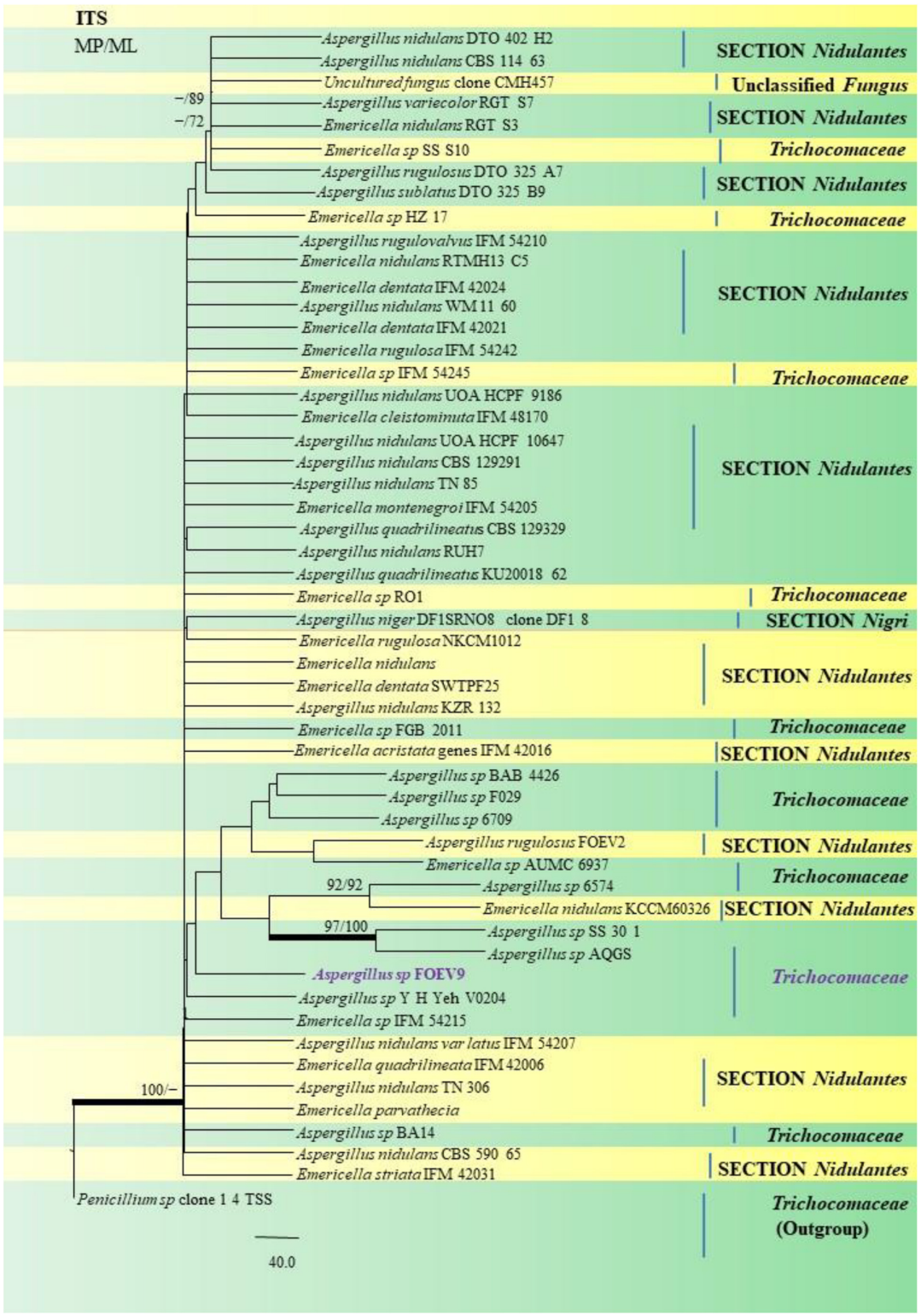

FIGURE 12 | A phylogenetic tree (taxon branching diagram) which presents the genetic analysis of the *Aspergillus* sp. FOEV9 clade belonging to the family *Trichocomaceae*, exhibiting a maximum parsimony analysis of partial ITS 1 and ITS 4 concatenated sequences, acquired. Bootstrap (bs) values ≥ 70% are designated at the nodes for maximum parsimony (MP) and maximum likelihood (ML) methods; branches with values ≥ 95% bs are bold black. A specific isolate code was mentioned as a suffix to the species name. Hyphen (-) represents the range of support values for parsimony and RAxML analyses of below 70%. Our isolate, which is marked in purple, was rooted using *Penicillium species* clone 1 4 TSS.

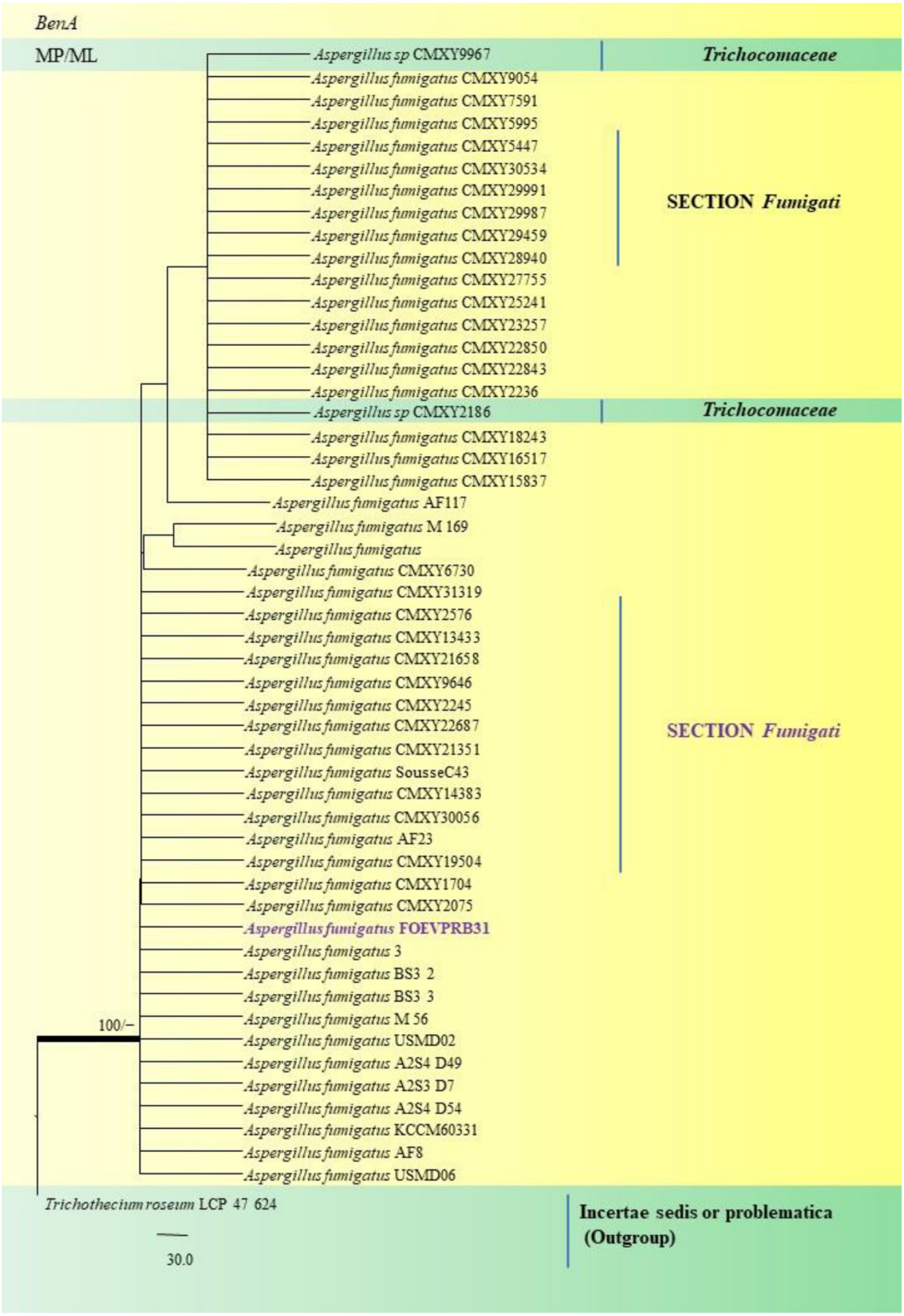

FIGURE 13 | A phylogenetic tree (taxon branching diagram) which presents the genetic analysis of the *Aspergillus fumigatus* FOEVPRB31 clade belonging to the Section *Fumigati* exhibiting a maximum parsimony analysis of partial β-tubulin concatenated sequences, acquired. Bootstrap (bs) values ≥ 70% are designated at the nodes for maximum parsimony (MP) and maximum likelihood (ML) methods; branches with values ≥ 95% bs are bold black. A specific isolate code was mentioned as a suffix to the species name. Hyphen (-) represents the range of support values for parsimony and RAxML analyses of below 70%. Our isolate, which is marked in purple, was rooted using *Trichothecium roseum* LCP 47 624.

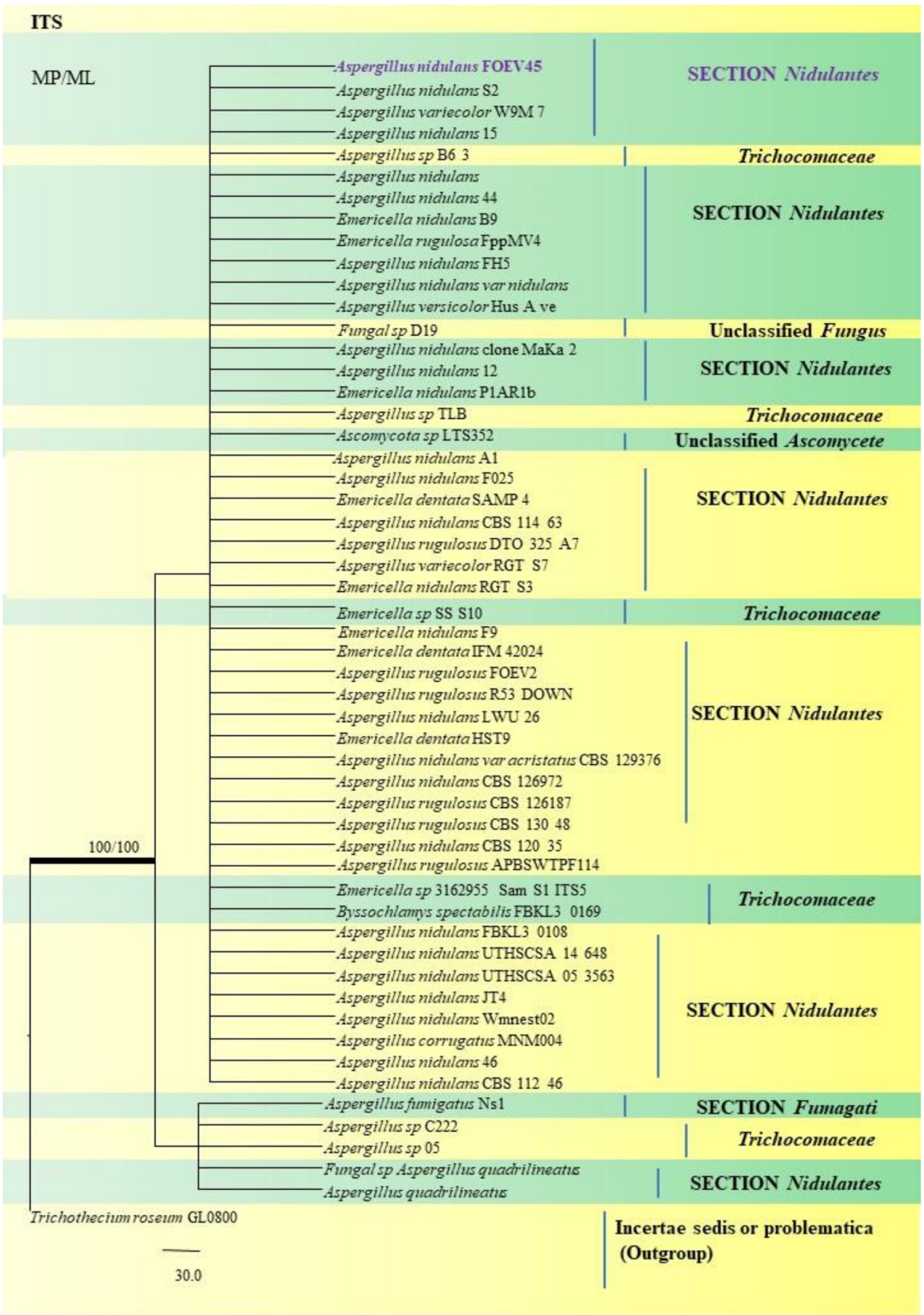

FIGURE 14 | A phylogenetic tree (taxon branching diagram) which presents the genetic analysis of the *Aspergillus nidulans* FOEV45 clade belonging to the Section *Nidulantes*, exhibiting a maximum parsimony analysis of partial ITS 1 and ITS 4 concatenated sequences, acquired. Bootstrap (bs) values ≥ 70% are designated at the nodes for maximum parsimony (MP) and maximum likelihood (ML) methods; branches with values ≥ 95% bs are bold black. A specific isolate code was mentioned as a suffix to the species name. Hyphen (-) represents the range of support values for parsimony and RAxML analyses of below 70%. Our isolate, which is marked in purple, was rooted using *Trichothecium roseum* GL0800.

from *F. vulgare* were found in the *Nigri* section in this study, which was collected from distinct geographical locations across India, including Delhi, Pondicherry, Rajasthan, Uttar Pradesh, and West Bengal. Isolates studied here, are predominantly predisposed with a toxigenic potential, known to cause a wide range of human and animal diseases, in addition to being plant pathogens.

Two *Aspergilli*, namely *Aspergillus aureoterreus* and *A. terreus*, were isolated from fennel samples classified under the *Terrei* Section. *A. aureoterreus* differs from *Aspergillus terreus* by the sequence analyses of the regions of the beta-tubulin genes (Samson et al., 2011). In the phylogenetic studies, the *A. aureoterreus* species formed a distinct lineage different from the *A. terreus* clade (Samson et al., 2011). The type-strain was isolated from *Foeniculum vulgare* (Sample No.18, place-Midnapore, West Bengal) and has been deposited in the National Center for Biotechnology Information (NCBI) GenBank with Accession No. MN791096.

A. sydowii has also been recognized as *Emericella sydowii*, a soil fungus, that could adulterate foodstuffs. It is also a pathogen known to cause human diseases like aspergillosis, onychomycosis, and keratomycosis (de Hoog et al., 2000). Apart from its disease-causing and toxigenic potential, *A. sydowii* has been the source of anticancer alkaloids. This species generates alkaloids, namely, 6-methoxyspirotryprostatin B, 18-oxotryprostatin A and 14-hydroxyterezine D (Zhang et al., 2008) and have potential cytotoxic activity against a549 cells (Nadumane et al., 2016).

Aspergillus flavus has received continual attention ever since its aflatoxigenic potential was discovered. The mold placed under the *Flavi* section, is known to be air-borne, thermotolerant and popularly as a saprophytic soil fungus. The fungus has been found all over the world (Ramírez-Camejo et al., 2012) and is abundant in temperate soil (Thom and Raper, 1945). *A. flavus* was found to be the most frequently occurring species in the fennel samples analyzed in the present study after *A. niger* and *Mycelia sterilia*, is known to be highly infectious with a broad range of infectivity and is able to contaminate agricultural crops during pre-harvest and post-harvest procedures (Bignell, 2010). In certain regions in Africa and Asia, AFs are appraised to be a major cause of severe acute sickness and loss of several lives per year (Wood, 1992; Alshannaq and Yu, 2017). The fungus is the most common source of aflatoxins in contaminated agricultural crops, can grow well in order to release aflatoxins frequently in almost any crop seed. The aflatoxins (AFs) are highly carcinogenic and immunosuppressive in nature, that has threatened global food safety, and has proven to be hazardous to animals, insects, and humans. It is an opportunistic fungal pathogen to humans as well as animals causing aspergillosis in immune-compromised hosts. In mammals, the pathogen can induce cancer of the liver (Amaike and Keller, 2011). Globally, over 4.5 billion people have been exposed to unmonitored levels of AFs (Alshannaq et al., 2018). Aflatoxin B_1 (AFB_1) is a highly toxic and the most potent natural compound to cause hepatocellular carcinoma in the universe. AFs are severely toxic, mutagenic, carcinogenic, teratogenic, immunosuppressants and are classified as Group1 carcinogens in human beings (Ostry et al., 2017) and hence are of special concern to human, animal and poultry health. *A. tamarii* is an epidemiological agent of human mycotic keratitis in India (Homa et al., 2019). Although it does not produce aflatoxins, it can induce elaboration of other toxic compounds like cyclopiazonic acid (Dorner et al., 1983; Ito, 1998), hence, posing a threat to human health.

Aspergillus fumigatus was studied under the *Fumigati* section and was isolated from fennel samples. With regard to the formation of sexual structures, sclerotia were not produced in the isolates studied by us. *A. fumigatus* has long been thought of reproduce only by asexual means until a detailed study was published by two researchers (Dyer and Paoletti, 2005). They elaborated on the possibility of the occurrence of sexual reproduction in *A. fumigatus*. Also, the detailed investigations and opinions of several mycologists have led to the findings, that point toward evidence that sexual reproduction may occur within populations of *A. fumigatus*. Teleomorphs of *A. fumigatus* are yet to be detected. However, it would be very significant to note that, the presence of sexual cycles would lead to an impending understanding of their biology and life cycles (Dyer and Paoletti, 2005). Other perspectives to understanding this species is its role as a human pathogen. *A. fumigatus* has been the highest prevailing aerial fungal parasite in immunodeficient hosts and is known to be associated with lethal systemic septicemia in multiple organs. It causes incursive infection in the lungs or long-term infection affecting the lungs, hypersensitive bronchopneumonia aspergillosis or allergic disorders in immune deficiencies (Hohl and Feldmesser, 2007; Segal, 2009). The disease can spread especially in individuals suffering from certain types of leukemia (Parahym et al., 2014) and those who are at the advanced stages of AIDS (Kaur et al., 2017), ailing with bronchial asthma (Kosmidis and Denning, 2015), fibrocystic disease of the pancreas (Düesberg et al., 2020), and in patients who have recovered from an influenza infection (Van De Veerdonk et al., 2017).

A. nidulans is a deuteromycetous member, which falls under the category of polyphyletic group of fungi whose sexual stages (teleomorph) are rare or unknown. In spite of this, mutants of *A. nidulans* are known to produce copious amounts of Hülle cells in the absence of sexual stages (Dyer and O'Gorman, 2012). A large number of species in the section *Nidulantes* have a sexual state whose name is *Emericella*, in the dual nomenclature system of fungi. If sexual stages are known and observed, the teleomorph has taxonomic precedence over the asexual stage, generally referred to as anamorph. Therefore, this species is most properly known as *Emericella nidulans*. Nevertheless, this fungal species will continue to be known in literature and the scientific or commercial industry for its inclusiveness as *A. nidulans*, which may be used here for both the anamorph and teleomorph interconvertibly. Asci contained eight-spored cells ascospores which were reddish/cherry-brown and lens-shaped and correlated with the original description by Thom and Raper (1945). A detailed microscopic description is available (Chen et al., 2016) for ascospore morphology of *Aspergilli* of the *Nidulantes* section. The ascospore color, ornamentation, shape and size are of particular value for differentiating species.

TABLE 2 | Sanger's sequencing data to provide a list of protein-altering mutations.

Sl. No.	Species	Our sequence ID	Genome references	Reference genome position	Mutation variant	Nucleotide substitution	Protein ID	Protein alteration	Mutation site
1	*Aspergillus nidulans*	MN791097.1	AACD01000016.1	426,219	Insertion	17insT	EAA66300.1	Nil	Intronic
2	*Aspergillus nidulans*	MN791097.1	AACD01000016.1	426,091	SNV	146C > A	EAA66300.1	–	Intronic
3	*Aspergillus nidulans*	MN791097.1	AACD01000016.1	426,075	SNV	162T > A	EAA66300.1	–	Intronic
4	*Aspergillus nidulans*	MN791097.1	AACD01000016.1	426,027	SNV	210A > G	EAA66300.1	p.43Q > Q	Protein coding: Synonymous
5	*Aspergillus nidulans*	MN791097.1	AACD01000016.1	425,990	SNV	138T > G	EAA66300.1	–	Intronic
6	*Aspergillus nidulans*	MN791097.1	AACD01000016.1	425,839	Double substitution	290AC > TG	EAA66300.1	p.89N > M	Protein coding: Non-synonymous
7	*Aspergillus nidulans*	MN791101.1	AACD01000016.1	425,867	SNV	261T > C	EAA66300.1	p.79G > G	Protein coding: Synonymous
8	*Aspergillus nidulans*	MN791102.1	AACD01000016.1	426,218	Insertion	19insT	EAA66300.1	–	Intronic
9	*Aspergillus nidulans*	MN791102.1	AACD01000016.1	426,075	SNV	174T > A	EAA66300.1	–	Intronic
10	*Aspergillus nidulans*	MN791102.1	AACD01000016.1	426,091	SNV	148C > A	EAA66300.1	–	Intronic
11	*Aspergillus nidulans*	MN791102.1	AACD01000016.1	426,028	SNV	212A > G	EAA66300.1	p.38G > G	Protein coding: Synonymous
12	*Aspergillus nidulans*	MN791102.1	AACD01000016.1	425,792	SNV	447C > T	EAA66300.1	p.105H > Y	**Protein coding: Non-synonymous**
13	*Aspergillus nidulans*	MN791116.1	AACD01000016.1	426,238	Double substitution	109TT > AC	EAA66300.1	p.16I > N	Protein coding: Non-synonymous
14	*Aspergillus nidulans*	MN791116.1	AACD01000016.1	425,859	SNV	489C > T	EAA66300.1	p.82G > G	Protein coding: Synonymous
15	*Aspergillus nidulans*	MN791116.1	AACD01000016.1	426,218	Insertion	20insT	EAA66300.1	–	Intronic
16	*Aspergillus nidulans*	MN791116.1	AACD01000016.1	426,132	SNV	107T > C	EAA66300.1	p.30L > P	**Protein coding: Non-synonymous**
17	*Aspergillus nidulans*	MN791116.1	AACD01000016.1	426,075	SNV	164T > A	EAA66300.1	–	Intronic
18	*Aspergillus nidulans*	MN791116.1	AACD01000016.1	426,091	SNV	148C > A	EAA66300.1	–	Intronic
19	*Aspergillus nidulans*	MN791116.1	AACD01000016.1	426,021	SNV	218G > A	EAA66300.1	p,45E > E	Protein coding: Synonymous
20	*Aspergillus nidulans*	MN791116.1	AACD01000016.1	426,027	SNV	212A > C	EAA66300.1	p.43Q > Q	Protein coding: Synonymous
21	*Aspergillus nidulans*	MN791116.1	AACD01000016.1	425,970	SNV	269G > A	EAA66300.1	–	Intronic
22	*Aspergillus nidulans*	MN791116.1	AACD01000016.1	425,786	SNV	453A > T	EAA66300.1	p.107T > S	**Protein coding: Non-synonymous**
23	*Aspergillus quadrilineatus*	MN791104.1	JAAXYA010000001.1	628,223	SNV	107T > C	ABW72458.1	p.10L > P	**Protein coding: Non-synonymous**
24	*Aspergillus quadrilineatus*	MN791104.1	JAAXYA010000001.1	628,112	SNV	218G > A	ABW72458.1	–	Intronic
25	*Aspergillus quadrilineatus*	MN791104.1	JAAXYA010000001.1	628,061	SNV	269G > A	ABW72458.1	–	Intronic
26	*Aspergillus quadrilineatus*	MN791104.1	JAAXYA010000001.1	627,877	SNV	453A > T	–	–	–
27	*Aspergillus latus*	MN791110.1	VCRL01000022.1	334,420	SNV	422T > A	QJS39736.1	p.83G > G	Protein coding: Synonymous
28	*Aspergillus latus*	MN791110.1	VCRL01000022.1	334,409	SNV	433A > T	QJS39736.1	p.87N > I	**Protein coding: Non-synonymous**
29	*Aspergillus latus*	MN791111.1	VCRL01000022.1	334,764	SNV	78C > G	QJS39736.1	–	Intronic
30	*Aspergillus tamarii*	MN791098.1	ML738700.1	20,713	SNV	10C > T	KAE8158230.1	p.21W > W	Protein coding: Synonymous
31	*Aspergillus tamarii*	MN791115.1	ML738700.1	20,700	Double substitution	12TA > AC	KAE8158230.1	p.16I > N	Protein-coding: Non-synonymous
32	*Aspergillus fumigatus*	MN791100.1	NC_007194.1	2,849,872	SNV	6T > C	XP_752456.1	–	Non-protein coding
33	*Aspergillus fumigatus*	MN791103.1	NC_007194.1	2,849,872	SNV	6T > C	XP_752456.1	–	Non-protein coding
34	*Aspergillus fumigatus*	MN791103.1	NC_007194.1	2,849,337	SNV	541C > T	XP_752456.1	–	Intronic
35	*Aspergillus awamori*	MN791113.1	BDHI01000014.1	2,536,570	SNV	13T > C	GCB22540.1	–	Intronic
36	*Aspergillus awamori*	MN791113.1	BDHI01000014.1	2,536,901	SNV	344T > G	GCB22540.1	–	Intronic
37	*Aspergillus awamori*	MN791113.1	BDHI01000014.1	2,536,908	SNV	351A > T	GCB22540.1	–	Intronic
38	*Aspergillus awamori*	MN791114.1	BDHI01000014.1	2,536,570	SNV	13T > C	GCB22540.1	–	Intronic
39	*Aspergillus awamori*	MN791114.1	BDHI01000014.1	2,536,901	SNV	344T > G	GCB22540.1	–	Intronic
40	*Aspergillus awamori*	MN791114.1	BDHI01000014.1	2,536,908	SNV	351A > T	GCB22540.1	–	Intronic
41	*Aspergillus awamori*	MN791114.1	BDHI01000014.1	2,536,658	SNV	108T > A	GCB22540.1	–	Intronic

Bold values refers to non-synonymous SNV mutations that we found. Such mutations may have an effect on the protein coding functional outcomes if any.

A. latus and *A. quadrilineatus* are being reported from *F. vulgare* for the very first time. Also, *A. latus* has been found to be phylogenetically similar to *A. sublatus* and *A. montenegroi*, thus considered synonymous. Ascospore morphology, which is a major and distinguishing feature while studying the sexual stages of different *Aspergilli*, were also documented in the current micromorphological analyses and were found to be orange/reddish brown with two pleated equatorial crests as studied recently and reported by Chen et al. (2016). The ascospores usually have two equatorial crests, whereas four crests can be seen in *A. quadrilineatus* by the aid of Scanning Electron Microscopy (SEM) and can sometimes be low or inconspicuous (Hubka et al., 2016). It has also been opined by Chen et al. (2016) that *A. nidulans* resembles *A. quadrilineatus* and is distinguishable in terms of its ascospore morphology. Phylogenetically *A. miyajii* has been found to be identical with *A. quadrilineatus*, *A. parvathecius A. acristatus,* and *A. floriformis* (Chen et al., 2016). *A. awamori* is being reported for the first and foremost time in fennel samples and is regarded as a domesticated form of *A. niger* (Samson et al., 2006) and its economic importance in Koji fermentation has found to be noteworthy (Kitamoto, 2002).

Nevertheless, the fungal phylogeny of *Aspergillus* species has undergone radical modifications over the course of over a hundred years. Phylogenetic analysis was especially useful when it was used to authenticate our morphological identification. In most instances, morphological identification and phylogenetic analyses were in agreement and helped in ascertaining species identity in *A. awamori* from the *Nigri* Section because of the striking phenotypic similarities to *A. niger* exhibited by our isolates. Phylogenetic studies were particularly useful while studying *A. aureoterreus*, *A. latus* and *Emericella quadrilineata* from Sections *Terrei*, and *Nidulantes,* respectively, along with *A. awamori,* which are being distinctly reported for the first time in *F. vulgare*. As a result, we were able to successfully distinguish them from closely related species and identify their Sections using the two genetic markers, which was precisely what we set out to do when this research was conceived.

Through phylogenetic analysis and tree construction, we were able to study twenty-two (22) *Aspergillus* species for the SNP analysis (**Figure 15**), classifying them into the Sections *Nidulantes*, *Flavi*, *Nigri*, *Terrei*, and *Fumigati*. Under the *Nidulantes* section, we could classify two *A. latus* (Isolate Nos. 1 and 2), five *A. nidulans* (Isolate Nos. 1, 2, 3, 4 and 5) and two *E. quadrilineata* (*E. quadrilineata* Isolate No. 1 and *E. quadrilineata* Isolate No. 2). Thus, the species of *E. quadrilineata* (Isolate No. 1) is rather closely connected to *A. latus* (Isolate No. 1) on one end, and *A. nidulans* (Isolate No. 3) is closely related to *E. quadrilineata* (Isolate No. 2) on the other end. *F. vulgare* samples from Pondicherry showed the presence of *E. quadrilineata* (Isolate No. 1) and *A. latus* (Isolate No. 1), and *A. nidulans* (Isolate No. 3) and *E. quadrilineata* (Isolate No. 2) were obtained from Rajasthan. While speaking about the diversity of fungi in the *F. vulgare* samples, three isolates of *A. latus* were studied: *A. latus* (Isolate Nos. 1 and 2) were obtained from Pondicherry, while *A. latus* (Isolate No. 3) was isolated from Jhargram, West Bengal, India. In the case of the occurrence of *A. nidulans* species, *A. nidulans* (Isolate Nos. 1, 2, 4, and 5) were from Pondicherry, whereas *A. nidulans* (Isolate No. 3) had been obtained from Rajasthan. This phylogenetic tree (**Figure 15**) of *Aspergilli* from *F. vulgare* samples from several regions demonstrate how closely related strains are distributed geographically. Additionally, in the section *Flavi*, three *A. tamarii* species appear, two of which, specifically (*A. tamarii* Isolate Nos. 1 and 2) were found to be more closely related and were isolated from Pondicherry spice samples. The *A. tamarii* (Isolate No. 3) from New Delhi emerged as a separate clade. Under the *Nigri* section, *A. awamori* species has been studied: *A. awamori* (FOEVPS8 1) was isolated from Pondicherry, while *A. awamori* (GLPL) was isolated from West Bengal. Three *A. aureoterreus* species and one *A. terreus* were grouped in the *Terrei* section. *A. aureoterreus* (Isolate No. 1) and *A. terreus* (Isolate No. 1) was recovered from a *F. vulgare* sample collected from New Delhi, *A. aureoterreus* (Isolate Nos. 2 and 3) were isolated from Pondicherry fennel seeds. *A. fumigatus* (Isolate Nos. 1 and 2) from Section *Fumigati*, were isolated from a Pondicherry fennel sample, whereas *A. fumigatus* (Isolate No. 3) which emerged as a separate branch was obtained from Lucknow. Strains are likely to represent a diverse set of features within the section under investigation because of the divergence between strains within clades.

The SNV analysis reveals that several mutations were identified exclusively in the *Nidulantes* section. Strain *E. nidulans* 5 displayed most SNPs (**Figure 15**) suggesting that the *Aspergilli* identified under this section may contain an isolate that is distinct. This further divulges information that, there was considerable genetic heterogeneity between *A. nidulans*, *E. quadrilineata* and *A. latus* species. Additionally, the genetic divergence between distant isolates supports distribution of *Aspergilli* across geographic regions. Furthermore, as demonstrated in the *Flavi*, *Fumigati*, and *Nigri* sections, there is no variance since the genetic sequences employed as markers in this study are highly conserved in terms of SNV mutations, showing that most SNPs have a small effect on the protein structure. The non-synonymous SNPs that were identified in our study, may have significant favorable or negative outcomes and needs to be investigated. The variants may be unique and could be used as biological markers to study specific species. However, the presence of SNPs in *Aspergillus* species may help us understand the rationale behind genetic diversity with implications of functional consequences if any, through further explorations on their mycotoxigenic potential and related prospective studies. According to the present study, the highlights include *Aspergillus* diversity from fennel seeds, a polyphasic approach to reappraise *Aspergilli* into six sections using the official fungal barcode and functional genetic markers which have been performed in adequate detail, with β-tubulin coding sequences being the focus for SNP analysis. Substitutions, insertions and SNV-kind of mutations have been identified among the *Aspergillus* isolates considered in this study. The data represented is the first report to the best of our knowledge. The occurrence of *Aspergilli* may help us determine the health status of the spices or cereals as indicator organisms, to adopt suitable strategies for avoidance of contamination at the field and during storage.

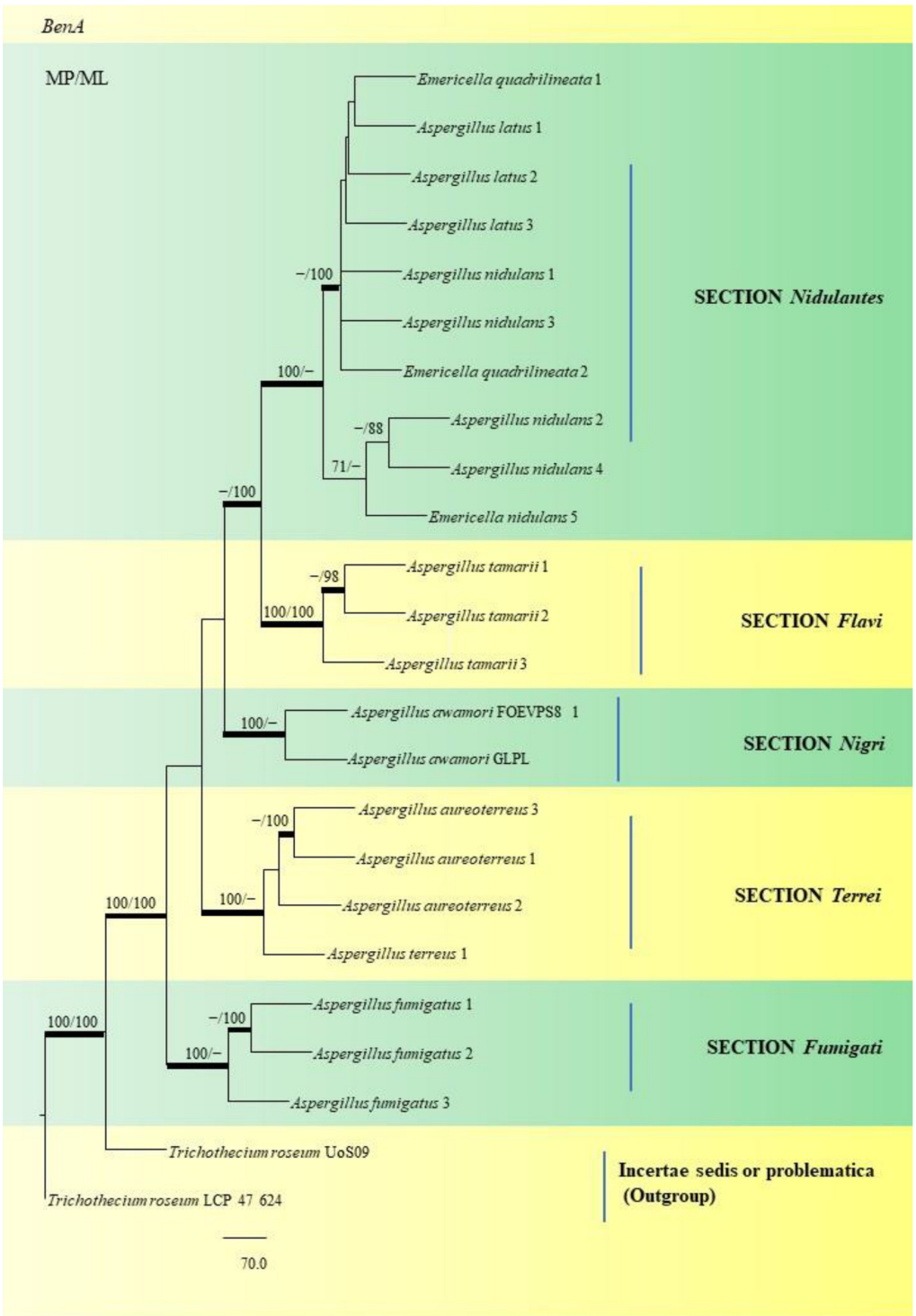

FIGURE 15 | This phylogram depicts the beta-tubulin phylogenetic tree of 22 *Aspergillus species* in sections, and subsequent analysis of their single nucleotide polymorphisms is discussed below. Maximum parsimony analysis was used to create this tree, which incorporates partial β-tubulin concatenated sequences. Boot strap (bs) values ≥ 70% are designated at the nodes for maximum parsimony (MP) and maximum likelihood (ML) methods; branches with values ≥ 95% bs are bold black. Hyphen (-) indicates support values below 70% for both parsimony and RAxML analyses. *Trichothecium roseum* Uo S09 and *Trichothecium roseum* LCP 47 624 were used to root the isolates. The Fungal Genetics and Mycotoxicology (FGM) laboratory codes for the isolates are listed here. 49: *E. quadrilineata* 1, 4: *A. latus* 1, 11: *A. latus* 2, 12: *A. latus* 3: 17: *A. nidulans* 1, 9: *A. nidulans* 3, 58: *E. quadrilineata* 2, 45: *A. nidulans* 2, 71: *A. nidulans* 4, 2: *E. quadrilineata*, 22: *A. tamarii* 1, 75: *A. tamarii* 2, S44: *A. tamarii* 3, S8: *A. awamori* S8, S24: *A. awamori* GLPL, 51: *A. aureoterreus* 3, 18: *A. aureoterreus* 1, 1: *A. aureoterreus* 2, 16: *A. terreus* 1, 24: *A. fumigatus* 1, 37: *A. fumigatus* 2 and 31: *A. fumigatus* 3.

CONCLUSION

The study was an elaborate investigation on the isolation and phenotypic characterization of *Aspergilli* from an Indian spice namely *F. vulgare*. While nearly 27 isolates have been adequately described with substantive morphological descriptions, attempts have been made to characterize these *Aspergilli* phylogenetically as well. A key strength of this endeavor is the number of fennel samples analyzed and *Aspergilli* being reported in a "first of it' s kind" investigation. The research study constitutes a key milestone in analyses of evolutionary relationships of fungi, detection of single nucleotide polymorphisms in the target fungal isolates. This validates genetic information, which serves as an essentiality for existing and further research. Although, the study may have its limitations, we propose to expand on the number of isolates to be studied for SNPs and their possible impacts on protein structure and functional consequences. Studies on their mycotoxigenic potentials is currently underway. Clearly, with respect to the diversity of *Aspergillus* species *F. vulgare*, the study is the largest till date. The extensive use of partial beta-tubulin gene analyses to evaluate the association between SNPs in five *Aspergillus* species sections is one of the highlights. Our research outcomes suggest that the presence of *Aspergilli* in food stuffs may pose a considerable hazard and threat to human consumption. Hence, identification and characterization may serve as a key factor in designing strategies to control post-harvest contamination and elaboration of mycotoxins. Fungal genomics serves as a useful molecular tool for inquiries into fungal evolution, by detecting gene differences and gene structure, to deduce the genetic basis of fungal evolution. However, blending morphological characteristics with molecular phylogeny is considered equally important for taxonomic studies and is the norm today.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/genbank/, MN791093; https://www.ncbi.nlm.nih.gov/genbank/, MN791096; https://www.ncbi.nlm.nih.gov/genbank/, MN791109; https://www.ncbi.nlm.nih.gov/genbank/, MN791095; https://www.ncbi.nlm.nih.gov/genbank/, MN791106; https://www.ncbi.nlm.nih.gov/genbank/, MN791098; https://www.ncbi.nlm.nih.gov/genbank/, MN791108; https://www.ncbi.nlm.nih.gov/genbank/, MN791115; https://www.ncbi.nlm.nih.gov/genbank/, MN791099; https://www.ncbi.nlm.nih.gov/genbank/, MN791100; https://www.ncbi.nlm.nih.gov/genbank/, MN791103; https://www.ncbi.nlm.nih.gov/genbank/, MN791097; https://www.ncbi.nlm.nih.gov/genbank/, MN791101; https://www.ncbi.nlm.nih.gov/genbank/, MN791102; https://www.ncbi.nlm.nih.gov/genbank/, MN791107; https://www.ncbi.nlm.nih.gov/genbank/, MN791116; https://www.ncbi.nlm.nih.gov/genbank/, MN791104; https://www.ncbi.nlm.nih.gov/genbank/, MN791105; https://www.ncbi.nlm.nih.gov/genbank/, MN791110; https://www.ncbi.nlm.nih.gov/genbank/, MN791111; https://www.ncbi.nlm.nih.gov/genbank/, MN791112; https://www.ncbi.nlm.nih.gov/genbank/, MN791113; https://www.ncbi.nlm.nih.gov/genbank/, MN791114; https://www.ncbi.nlm.nih.gov/genbank/, MN298848; https://www.ncbi.nlm.nih.gov/genbank/, MN309877; https://www.ncbi.nlm.nih.gov/genbank/, MN264637; https://www.ncbi.nlm.nih.gov/genbank/, MN326529; and https://www.ncbi.nlm.nih.gov/genbank/, MN39290.

AUTHOR CONTRIBUTIONS

PM and RD engaged in the research project's planning and design. PM carried out all the experiments, documented the findings, and wrote the manuscript. RD authored the results, drafted the manuscript, corrected the errors, and assisted in writing the manuscript. AP undertook the phylogenetic analysis, while BM performed the SNP analysis. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The Central Instrumentation Facility, Pondicherry University, is acclaimed for the services it provides. We sincerely acknowledge V. Venkateswara Sarma, Fungal Biotechnology Laboratory, Department of Biotechnology, School of Life Sciences, Pondicherry University, for providing the Differential Interference Contrast (DIC) microscope facility. In appreciation of Pondicherry University's Non-NET University Fellowship, PKM extends his gratitude to the University.

REFERENCES

Ahene, R. E., Odamtten, G. T., and Owusu, E. (2011). Fungal and bacterial contaminants of six spices and spice products in Ghana. *Afr. J. Environ. Sci. Technol.* 5, 633–640.

Ahmad, M. M., Ahmad, M., Ali, A., Hamid, R., Javed, S., and Abdin, M. Z. (2014). Detection of *Aspergillus flavus* and *Aspergillus parasiticus* from aflatoxin-contaminated peanuts and their differentiation using PCR-RFLP. *Ann. Microbiol.* 64, 1597–1605. doi: 10.1007/s13213-014-0803-5

Ahn, G., Kim, S., Lee, J.-E., Lee, S., Kim, S. H., and Kim, S.-R. (2020). *Penicillium is Dominant Airborne Contaminant in Winter Seasons in Asthma Patients Houses: A Case Study in Korea*. Available online at: https://www.ncbi.nlm.nih.gov/nuccore/MT582755 (accessed January 21, 2021).

Ali, N., Hashim, N. H., and Shuib, N. S. (2015). Natural occurrence of aflatoxins and ochratoxin A in processed spices marketed in Malaysia. *Food Addit. Contam. Part A* 32, 518–532. doi: 10.1080/19440049.2015.1011712

Alshannaq, A., and Yu, J. H. (2017). Occurrence, toxicity, and analysis of major mycotoxins in food. *Int. J. Environ. Res. Public Health* 14:632. doi: 10.3390/ijerph14060632

Alshannaq, A. F., Gibbons, J. G., Lee, M. K., Han, K. H., Hong, S. B., and Yu, J. H. (2018). Controlling aflatoxin contamination and propagation of *Aspergillus flavus* by a soy-fermenting *Aspergillus oryzae* strain. *Sci. Rep.* 8, 1–14. doi: 10.1038/s41598-018-35246-1

Alshehri, B., and Palanisamy, M. (2020). Evaluation of molecular identification of *Aspergillus* species causing fungal keratitis. *Saudi J. Biol. Sci.* 27, 751–756. doi: 10.1016/j.sjbs.2019.12.030

Amadi, J. E., and Adeniyi, D. O. (2009). Mycotoxin production by fungi isolated from stored grains. *Afr. J. Biotechnol.* 8, 1219–1221.

Amaike, S., and Keller, N. P. (2011). *Aspergillus flavus*. *Annu. Rev. Phytopathol.* 49, 107–133. doi: 10.1146/annurev-phyto-072910-095221

Andreolli, M., Milanesi, C., Faleri, C., Cresti, M., Lampis, S., Vallini, G., et al. (2018). *Well Preserved Seeds of Apple Found in An Amphora Excavated From a Burned Roman Villa Despite the Deteriorating Microorganisms*. Available online at: https://www.ncbi.nlm.nih.gov/nuccore/mh447369 (accessed March 22, 2021).

Anonymous (2020). *Connect2india*. Available online at: https://connect2india.com/global/Fennel-Seed-export-from-india/1 (accessed February 18, 2021).

Anonymous (2021). *MycoCosm*. Available online at: https://mycocosm.jgi.doe.gov/Aspquag1/Aspquag1.home.html (Accessed March 8, 2021).

Arné, P., Risco-Castillo, V., Jouvion, G., Le Barzic, C., and Guillot, J. (2021). *Aspergillosis* in wild birds. *J. Fungi* 7:241. doi: 10.3390/jof7030241

Ayliffe, M. A., Dodds, P. N., and Lawrence, G. J. (2001). Characterisation of a β-tubulin gene from Melampsora lini and comparison of fungal β-tubulin genes. *Mycol. Res.* 105, 818–826. doi: 10.1017/S0953756201004245

Azzoune, N., Mokrane, S., Riba, A., Bouras, N., Verheecke, C., Sabaou, N., et al. (2015). Contamination of common spices by aflatoxigenic fungi and aflatoxin B1 in Algeria. *Qual. Assur. Saf. Crop. Foods* 8, 137–144. doi: 10.3920/qas2014.0426

Badgujar, S. B., Patel, V. V., and Bandivdekar, A. H. (2014). Foeniculum vulgare Mill: a review of its botany, phytochemistry, pharmacology, contemporary application, and toxicology. *Biomed. Res. Int.* 2014:842674. doi: 10.1155/2014/842674

Balajee, S. A., Hurst, S. F., Baddley, J. W., Peterson, S. W., Nickle, D., Lass-Florl, C., et al. (2009). *Molecular Phylogeny of the Opportunistic Fungal Pathogen Aspergillus terreus*. Available online at: https://www.ncbi.nlm.nih.gov/nuccore/EU147717 (Accessed April 21, 2021).

Battilani, P., Toscano, P., Van der Fels-Klerx, H. J., Moretti, A., Leggieri, M. C., Brera, C., et al. (2016). Aflatoxin B 1 contamination in maize in Europe increases due to climate change. *Sci. Rep.* 6, 1–7. doi: 10.1038/srep24328

Bennett, J. W. (2010). "An overview of the genus *Aspergillus*," in *Aspergillus: Molecular Biology and Genomics*, eds M. Machida and K. Gomi (Norfolk: Caister Academic Press), 1–17.

Bhatnagar-Mathur, P., Sunkara, S., Bhatnagar-Panwar, M., Waliyar, F., and Sharma, K. K. (2015). Biotechnological advances for combating *Aspergillus flavus* and aflatoxin contamination in crops. *Plant Sci.* 234, 119–132. doi: 10.1016/j.plantsci.2015.02.009

Bignell, E. (2010). *Aspergillus*: molecular biology and genomics. *Biotechnol. J.* 5, 336–337. doi: 10.1002/biot.201000025

Bobadilla-Carrillo, G. I., Magallon-Servin, P., Lopez-Vela, M., Ramirez-Ramirez, J. C., Gutierrez-Leyva, R., Palomino-Hermosillo, Y. A., et al. (2020). *Characterization and Proliferation Capacity of Potentially Pathogenic Fungi in Marine and Freshwater Fish Commercial Feeds*. Available online at: https://www.ncbi.nlm.nih.gov/nuccore/1799637099 (accessed April 21, 2021).

Casas López, J. L., Sánchez Pérez, J. A., Fernández Sevilla, J. M., Acién Fernández, F. G., Molina Grima, E., and Chisti, Y. (2004). Fermentation optimization for the production of lovastatin by *Aspergillus terreus*: use of response surface methodology. *J. Chem. Technol. Biotechnol.* 79, 1119–1126. doi: 10.1002/jctb.1100

Chavan, S. (2020). *Molecular Identification of Fungi Causing Superficial Mycoses*. Available online at: https://www.ncbi.nlm.nih.gov/nuccore/MT497447 (Accessed March 20, 2021).

Chen, A. J., Frisvad, J. C., Sun, B. D., Varga, J., Kocsubé, S., Dijksterhuis, J., et al. (2016). *Aspergillus* section Nidulantes (formerly Emericella): polyphasic taxonomy, chemistry and biology. *Stud. Mycol.* 84, 1–118. doi: 10.1016/j.simyco.2016.10.001

Chen, M., and Xu, Y. (2018). *Phylogenetic Diversity and Antifungal Susceptibility of Clinical Aspergillus Isolates in Shanghai, China*. Available online at: https://www.ncbi.nlm.nih.gov/nuccore/MG991348 (accessed January 20, 2021).

Chen, P., Liu, M., Zeng, Q., Zhang, Z., Kong, Q., Liu, W., et al. (2018). *Uncovering New Mutations in the cyp51A Gene in Aspergillus Fumigatus Conferring Azole Resistance*. Available online at: https://www.ncbi.nlm.nih.gov/nuccore/MH536090 (accessed February 26, 2021).

Cleveland, D. W., and Sullivan, K. F. (1985). Molecular biology and genetics of tubulin. *Annu. Rev. Biochem.* 54, 331–366. doi: 10.1146/annurev.bi.54.070185.001555

de Hoog, G. S., Guarro, J., Gené, J., and Figueras, M. J. (2000). *Atlas of Clinical Fungi*. Utrecht: Centraalbureau voor Schimmelcultures (CBS).

Dorner, J. W., Cole, R. J., Lomax, L. G., Gosser, H. S., and Diener, U. L. (1983). Cyclopiazonic acid production by *Aspergillus flavus* and its effects on broiler chickens. *Appl. Environ. Microbiol.* 46, 698–703. doi: 10.1128/aem.46.3.698-703.1983

dos Santos-Ciscon, B. A., van Diepeningen, A., da Cruz Machado, J., Dias, I. E., and Waalwijk, C. (2019). *Aspergillus* species from Brazilian dry beans and their toxigenic potential. *Int. J. Food Microbiol.* 292, 91–100. doi: 10.1016/j.ijfoodmicro.2018.12.006

Düesberg, U., Wosniok, J., Naehrlich, L., Eschenhagen, P., and Schwarz, C. (2020). Risk factors for respiratory *Aspergillus fumigatus* in German Cystic Fibrosis patients and impact on lung function. *Sci. Rep.* 10, 1–9. doi: 10.1038/s41598-020-75886-w

Dyer, P. S., and O'Gorman, C. M. (2012). Sexual development and cryptic sexuality in fungi: insights from *Aspergillus* species. *FEMS Microbiol. Rev.* 36, 165–192. doi: 10.1111/j.1574-6976.2011.00308.x

Dyer, P. S., and Paoletti, M. (2005). Reproduction in *Aspergillus fumigatus*: sexuality in a supposedly asexual species? *Med. Mycol.* 43, S7–S14. doi: 10.1080/13693780400029015

Edgcomb, V. P., Roger, A. J., Simpson, A. G., Kysela, D. T., and Sogin, M. L. (2001). Evolutionary relationships among "jakobid" flagellates as indicated by alpha-and beta-tubulin phylogenies. *Mol. Biol. Evol.* 18, 514–522. doi: 10.1093/oxfordjournals.molbev.a003830

Elshafie, A. E., Al-Rashdi, T. A., Al-Bahry, S. N., and Bakheit, C. S. (2002). Fungi and aflatoxins associated with spices in the Sultanate of Oman. *Mycopathologia* 155, 155–160. doi: 10.1023/A:1020427527963.pdf

Garcia, M. V., Parussolo, G., Moro, C. B., Bernardi, A. O., and Copetti, M. V. (2018). Fungi in spices and mycotoxigenic potential of some *Aspergilli* isolated. *Food Microbiol.* 73, 93–98. doi: 10.1016/j.fm.2018.01.013

Goto, T., Wicklow, D. T., and Ito, Y. (1996). Aflatoxin and cyclopiazonic acid production by a sclerotium-producing *Aspergillus tamarii* strain. *Appl. Environ. Microbiol.* 62, 4036–4038. doi: 10.1128/aem.62.11.4036-4038.1996

Hall, T. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* 41, 95–98.

Hammami, W., Fiori, S., Al Thani, R., Kali, N. A., Balmas, V., Migheli, Q., et al. (2014). Fungal and aflatoxin contamination of marketed spices. *Food Control* 37, 177–181. doi: 10.1016/j.foodcont.2013.09.027

Hamzah, T. N. T., Lee, S. Y., Hidayat, A., Terhem, R., Faridah-Hanum, I., and Mohamed, R. (2018). Diversity and characterization of endophytic fungi isolated from the tropical mangrove species, *Rhizophora mucronata*, and identification of potential antagonists against the soil-borne fungus, *Fusarium solani*. *Front. Microbiol.* 9:1707. doi: 10.3389/fmicb.2018.01707

Han, X., Jiang, H., Xu, J., Zhang, J., and Li, F. (2017). Dynamic Fumonisin B2 production by *Aspergillus niger* intented used in food industry in China. *Toxins* 9:217. doi: 10.3390/toxins9070217

Hillis, D. M., and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* 42, 182–192. doi: 10.1093/sysbio/42.2.182

Hiort, J., Maksimenka, K., Reichert, M., Perović-Ottstadt, S., Lin, W. H., Wray, V., et al. (2004). New natural products from the sponge-derived fungus *Aspergillus niger*. *J. Nat. Prod.* 67, 1532–1543. doi: 10.1021/np030551d

Hohl, T. M., and Feldmesser, M. (2007). *Aspergillus fumigatus*: principles of pathogenesis and host defense. *Eukaryot. Cell* 6, 1953–1963. doi: 10.1128/EC.00274-07

Homa, M., Manikandan, P., Szekeres, A., Kiss, N., Kocsubé, S., Kredics, L., et al. (2019). Characterization of *Aspergillus tamarii* strains from human keratomycoses: molecular identification, antifungal susceptibility patterns and cyclopiazonic acid producing abilities. *Front. Microbiol.* 10:2249. doi: 10.3389/fmicb.2019.02249

Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res.* 9, 868–877. doi: 10.1101/gr.9.9.868

Hubka, V., Nováková, A., Peterson, S. W., Frisvad, J. C., Sklenář, F., and Matsuzawa, T. (2016). A reappraisal of *Aspergillus* section Nidulantes with descriptions of two new sterigmatocystin-producing species. *Plant Syst. Evol.* 302, 1267–1299. doi: 10.1007/s00606-016-1331-5

Ito, Y. (1998). Properties of *Aspergillus tamarii*, A. caelatus and related species from acidic tea field soils in Japan. *Mycopathologia* 144, 169–175. doi: 10.1023/A:1007021527106.pdf

Javidnia, K., Dastgheib, L., Samani, S. M., and Nasiri, A. (2003). Anti-hirsutism activity of fennel (fruits of Foeniculum vulgare) extract–a double-blind placebo-controlled study. *Phytomedicine* 10, 455–458. doi: 10.1078/094471103322331386

Juuti, J. T., Jokela, S., Tarkka, M. T., Paulin, L., and Lahdensalo, J. (2005). Two phylogenetically highly distinct β-tubulin genes of the basidiomycete *Suillus bovinus*. *Curr. Genet.* 47, 253–263. doi: 10.1007/s00294-005-0564-6

Kashyap, S. (2020). Approach to allergic bronchopulmonary aspergillosis (ABPA): quick review. *EC Pul. Res. Med.* 9, 90–101.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kaur, R., Mehra, B., Dhakad, M. S., Goyal, R., and Dewan, R. (2017). Pulmonary aspergillosis as opportunistic mycoses in a cohort of human immunodeficiency virus-infected patients: report from a tertiary care hospital in North India. *Int. J. Health Sci.* 11:45.

Keeling, P. J., Luker, M. A., and Palmer, J. D. (2000). Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. *Mol. Biol. Evol.* 17, 23–31. doi: 10.1093/oxfordjournals.molbev.a026235

Kirk, K. E., and Morris, N. R. (1991). The tubB alpha-tubulin gene is essential for sexual development in *Aspergillus nidulans*. *Genes Dev.* 5, 2014–2023. doi: 10.1101/gad.5.11.2014

Kitamoto, K. (2002). Molecular biology of the Koji molds. *Adv. Appl. Microbiol.* 51, 129–154. doi: 10.1016/s0065-2164(02)51004-2

Koppula, S., and Kumar, H. (2013). Foeniculum vulgare Mill (Umbelliferae) attenuates stress and improves memory in wister rats. *Trop. J. Pharm. Res.* 12, 553–558. doi: 10.4314/tjpr.v12i4.17

Kosmidis, C., and Denning, D. W. (2015). The clinical spectrum of pulmonary aspergillosis. *Thorax* 70, 270–277. doi: 10.1136/thoraxjnl-2014-206291

Kulshrestha, P., Singh, C., Gupta, A., Mahajan, S., and Sharma, R. (2014). Mycoflora associated with spices. *Int. J. Curr. Microbiol. Appl. Sci.* 3, 741–746.

Mages, W., Cresnar, B., Harper, J. F., Brüderlein, M., and Schmitt, R. (1995). Volvox carteri α2-and β2-tubulin-encoding genes: regulatory signals and transcription. *Gene* 160, 47–54. doi: 10.1016/0378-1119(95)00178-9

Magnoli, C. E., Astoreca, A. L., Chiacchiera, S. M., and Dalcero, A. M. (2007). Occurrence of ochratoxin A and ochratoxigenic mycoflora in corn and corn-based foods and feeds in some South American countries. *Mycopathologia* 163, 249–260. doi: 10.1007/s11046-007-9005-z

Makhlouf, J., Carvajal-Campos, A., Querin, A., Tadrist, S., Puel, O., Lorber, S., et al. (2019). Morphologic, molecular and metabolic characterization of *Aspergillus* section Flavi in spices marketed in Lebanon. *Sci. Rep.* 9, 1–11. doi: 10.1038/s41598-019-41704-1

Malathi, S., and Chakraborty, R. (1991). Production of alkaline protease by a new *Aspergillus flavus* isolate under solid-substrate fermentation conditions for use as a depilation agent. *Appl. Environ. Microbiol.* 57, 712–716. doi: 10.1128/aem.57.3.712-716.1991

Matsuzawa, T., Yaguchi, T., Horie, Y., and Nshimura, K. (2006). *Molecular Phylogenetics of the Genus Emericella and the Surface Structure of Ascospores*. Available online at: https://www.ncbi.nlm.nih.gov/nuccore/AB248335 (accessed October 13, 2020).

Moloney, C. (2019). *First Post*. Available online at: https://www.firstpost.com/business/indias-major-agricultural-produce-losses-estimated-at-rs-92000-cr-2949002.html (accessed February 16, 2021).

Mukherjee, M., Hadar, R., Mukherjee, P. K., and Horwitz, B. A. (2003). Homologous expression of a mutated beta-tubulin gene does not confer benomyl resistance on *Trichoderma virens*. *J. Appl. Microbiol.* 95, 861–867. doi: 10.1046/j.1365-2672.2003.02061.x

Nadumane, V. K., Venkatachalam, P., and Gajaraj, B. (2016). "*Aspergillus* applications in cancer research," in *New and Future Developments in Microbial Biotechnology and Bioengineering*, ed. V. K. Gupta (Amsterdam: Elsevier), 243–255.

Nurtjahja, K., Zuhra, C. F., Sembiring, H., Bungsu, A., Simanullang, J., Silalahi, J. E., et al. (2019). Fungal contamination spices from Indonesia with emphasis on *Aspergillus flavus*. *Czech J. Food Sci.* 37, 338–344. doi: 10.17221/18/2019-CJFS

Op De Beeck, M., Lievens, B., Busschaert, P., Declerck, S., Vangronsveld, J., and Colpaert, J. V. (2014). Comparison and validation of some ITS primer pairs useful for fungal metabarcoding studies. *PLoS One* 9:e97629. doi: 10.1371/journal.pone.0097629

Ostry, V., Malir, F., Toman, J., and Grosse, Y. (2017). Mycotoxins as human carcinogens—the IARC Monographs classification. *Mycotoxin Res.* 33, 65–73. doi: 10.1007/s12550-016-0265-7

Pal, M., Dave, P., and Manna, A. K. (2014). Emerging role of *Aspergillus flavus* in human and animal disorders. *J. Mycopathol. Res.* 52, 211–216.

Parahym, A. M. R. D. C., Neto, P. J. R., Silva, C. M. D., Gonçalves, S. S., and Motta, C. M. D. S. (2014). Fatal invasive Aspergillosis in acute lymphoblastic leukemia patient. *J. Clin. Case Rep.* 4:2. doi: 10.4172/2165-7920.1000368

Ráduly, Z., Szabó, L., Madar, A., Pócsi, I., and Csernoch, L. (2020). Toxicological and medical aspects of *Aspergillus*-derived mycotoxins entering the feed and food chain. *Front. Microbiol.* 10:2908. doi: 10.3389/fmicb.2019.02908

Raja, H. A., Miller, A. N., Pearce, C. J., and Oberlies, N. H. (2017). Fungal identification using molecular tools: a primer for the natural products research community. *J. Nat. Products* 80, 756–770. doi: 10.1021/acs.jnatprod.6b01085

Rambaut, A., and Drummond, A. J. (2009). *FigTree v. 1.3.1. Computer Program and Documentation Distributed by the Author*. Available online at: http://tree.bio.ed.ac.uk/software/ (accessed February 11, 2020).

Ramírez-Camejo, L. A., Zuluaga-Montero, A., Lázaro-Escudero, M., Hernández-Kendall, V., and Bayman, P. (2012). Phylogeography of the cosmopolitan fungus *Aspergillus flavus*: is everything everywhere? *Fungal Biol.* 116, 452–463. doi: 10.1016/j.funbio.2012.01.006

Rather, M. A., Dar, B. A., Sofi, S. N., Bhat, B. A., and Qurishi, M. A. (2016). Foeniculum vulgare: a comprehensive review of its traditional use, phytochemistry, pharmacology, and safety. *Arab. J. Chem.* 9, S1574–S1583. doi: 10.1016/j.arabjc.2012.04.011

Reddy, K. R. N., Reddy, C. S., and Muralidharan, K. (2009). Detection of *Aspergillus* spp. and aflatoxin B1 in rice in India. *Food Microbiol.* 26, 27–31. doi: 10.1016/j.fm.2008.07.013

Refai, M., El-Yazid, H. A., and Hassan, A. (eds) (2014). "Monograph on *Aspergillus* and aspergillosis in man, animals and birds," in *A Guide for Classification and Identification of Aspergilli, Diseases Caused by Them, Diagnosis and Treatment* (Cambridge, MA: Academia Press).

Samadi-Noshahr, Z., Hadjzadeh, M. A. R., Moradi-Marjaneh, R., and Khajavi-Rad, A. (2021). The hepatoprotective effects of fennel seeds extract and trans-Anethole in streptozotocin-induced liver injury in rats. *Food Sci. Nutr.* 9, 1121–1131. doi: 10.1002/fsn3.2090

Samson, R. A., Hong, S. B., and Frisvad, J. C. (2006). Old and new concepts of species differentiation in *Aspergillus*. *Med. Mycol.* 44, S133–S148. doi: 10.1080/13693780600913224

Samson, R. A., Peterson, S. W., Frisvad, J. C., and Varga, J. (2011). New species in *Aspergillus* section terrei. *Stud. Mycol.* 69, 39–55. doi: 10.3114/sim.2011.69.04

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., et al. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U.S.A.* 109, 6241–6246. doi: 10.1073/pnas.1117018109

Schütze, J., Krasko, A., Custodio, M. R., Efremova, S. M., Müller, I. M., and Müller, W. E. (1999). Evolutionary relationships of Metazoa within the eukaryotes based on molecular data from Porifera. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 266, 63–73. doi: 10.1098/rspb.1999.0605

Segal, B. H. (2009). "Invasive aspergillosis in chronic granulomatous disease," in *Aspergillosis: From Diagnosis to Prevention*, ed. A. C. Pasqualotto (London: Springer), 527–543.

Siruguri, V., and Bhat, R. V. (2015). Assessing intake of spices by pattern of spice use, frequency of consumption and portion size of spices consumed from routinely prepared dishes in southern India. *Nutr. J.* 14, 1–9. doi: 10.1186/1475-2891-14-7

Soni, P., Gangurde, S. S., Ortega-Beltran, A., Kumar, R., Parmar, S., Sudini, H. K., et al. (2020). Functional biology and molecular mechanisms of host-pathogen interactions for aflatoxin contamination in groundnut (*Arachis hypogaea* L.) and maize (*Zea mays* L.). *Front. Microbiol.* 11:227. doi: 10.3389/fmicb.2020.00227

Ssepuuya, G., Van Poucke, C., Ediage, E. N., Mulholland, C., Tritscher, A., Verger, P., et al. (2018). Mycotoxin contamination of sorghum and its contribution to

human dietary exposure in four sub-Saharan countries. *Food Addit. Contam. A* 35, 1384–1393. doi: 10.1080/19440049.2018.1461253

Stamatakis, A., Aberer, A. J., Goll, C., Smith, S. A., Berger, S. A., and Izquierdo-Carrasco, F. (2012). RAxML-Light: a tool for computing terabyte phylogenies. *Bioinformatics* 28, 2064–2066. doi: 10.1093/bioinformatics/bts309

Streit, E., Schatzmayr, G., Tassis, P., Tzika, E., Marin, D., Taranu, I., et al. (2012). Current situation of mycotoxin contamination and co-occurrence in animal feed—Focus on Europe. *Toxins* 4, 788–809. doi: 10.3390/toxins4100788

Swofford, D. L., and Sullivan, J. (2003). "Phylogeny inference based on parsimony and other methods using PAUP*," in *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny*, eds M. Salemi and A.-M. Vandamme (Cambridge, MA: Cambridge University Press).

Tam, E. W., Chen, J. H., Lau, E. C., Ngan, A. H., Fung, K. S., Lee, K. C., et al. (2014). Misidentification of *Aspergillus nomius* and *Aspergillus tamarii* as *Aspergillus flavus*: characterization by internal transcribed spacer, β-tubulin, and calmodulin gene sequencing, metabolic fingerprinting, and matrix-assisted laser desorption ionization–time of flight mass spectrometry. *J. Clin. Microbiol.* 52, 1153–1160. doi: 10.1128/JCM.03258-13

Thom, C., and Raper, K. B. (1945). *A Manual of the Aspergilli*. Philadelphia: LWW.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882. doi: 10.1093/nar/25.24.4876

Van De Veerdonk, F. L., Kolwijck, E., Lestrade, P. P., Hodiamont, C. J., Rijnders, B. J., Van Paassen, J., et al. (2017). Influenza-associated aspergillosis in critically ill patients. *Am. J. Respir. Crit. Care Med.* 196, 524–527. doi: 10.1164/rccm.201612-2540LE

Varga, J., and Samson, R. A. (eds) (2008). *Aspergillus in the Genomic Era*. Wageningen: Wageningen Academic Publishers.

Visagie, C. M., and Houbraken, J. (2020). Updating the taxonomy of *Aspergillus* in South Africa. *Stud. Mycol.* 95, 253–292. doi: 10.1016/j.simyco.2020.02.003

Wood, G. E. (1992). Mycotoxins in foods and feeds in the United States. *J. Anim. Sci.* 70, 3941–3949. doi: 10.2527/1992.70123941x

Xie, G. F., Li, W. J., Lu, J., Cao, Y., Fang, H., Zou, H. J., et al. (2007). Isolation and identification of representative fungi from Shaoxing rice wine wheat Qu using a polyphasic approach of culture-based and molecular-based methods. *J. Inst. Brew.* 113, 272–279. doi: 10.1002/j.2050-0416.2007.tb00287.x

Zhang, M., Wang, W. L., Fang, Y. C., Zhu, T. J., Gu, Q. Q., and Zhu, W. M. (2008). Cytotoxic alkaloids and antibiotic nordammarane triterpenoids from the marine-derived fungus *Aspergillus sydowi*. *J. Nat. Prod.* 71, 985–989. doi: 10.1021/np700737g

Zhao, Z., Liu, H., Luo, Y., Zhou, S., An, L., Wang, C., et al. (2014). Molecular evolution and functional divergence of tubulin superfamily in the fungal tree of life. *Sci. Rep.* 4, 1–13. doi: 10.1038/srep06746

Conflict of Interest: BM was employed by the International Tech Park.

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.

frontiers in Microbiology

METHODS
published: 29 March 2022
doi: 10.3389/fmicb.2022.875347

Rapid and Routine Molecular Typing Using Multiplex Polymerase Chain Reaction and MinION Sequencer

Yu-Chieh Liao[1†], Han-Chieh Wu[2†], Ci-Hong Liou[2], Tsai-Ling Yang Lauderdale[2,3], I-Wen Huang[2], Jui-Fen Lai[2] and Feng-Jui Chen[2,3*]*

[1]Institute of Population Health Sciences, National Health Research Institutes, Zhunan Town, Miaoli County, Taiwan, [2]National Institute of Infectious Diseases and Vaccinology, National Health Research Institutes, Zhunan Town, Miaoli County, Taiwan, [3]Department of Biological Science and Technology, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

Molecular typing is an essential tool that has been extensively applied in laboratories as well as in clinical settings. Next-generation sequencing technologies promise high-throughput and cost-effective molecular applications; however, the accessibility of these technologies is limited due to the high capital cost. Oxford Nanopore Technologies (ONT) offers a MinION device with the advantages of real-time data analysis, rapid library preparation, and low cost per test. However, the advantages of the MinION device are often overshadowed by its lower raw accuracy. Herein, we present a concise multilocus sequence typing protocol of *Staphylococcus aureus* using multiplex polymerase chain reaction and Rapid Barcoding Kit for barcoding and MinION device for sequencing. Moreover, to clarify the effects of carryover DNA on tasks that require high sequence accuracy, we used the MinION flow cell in successive runs of washing and reusing. Our results revealed that the MinION flow cell could achieve accurate typing of a total of 467 samples with 3,269 kilobase-long genes within a total of 5 runs. This thus demonstrates the effectiveness of a portable nanopore MinION sequencer in providing accurate, rapid, and routine molecular typing.

Keywords: nanopore sequencing, molecular typing, multiplex polymerase chain reaction, multilocus sequence typing, MinION

OPEN ACCESS

Edited by:
Zheng Zhang,
Shandong University, China

Reviewed by:
Farzaneh Firoozeh,
Alborz University of Medical Sciences, Iran
Nicolás Tomasini,
Universidad Nacional de Salta, Argentina

***Correspondence:**
Yu-Chieh Liao
jade@nhri.edu.tw
Feng-Jui Chen
frchen@nhri.edu.tw

†These authors have contributed equally to this work

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

Received: *14 February 2022*
Accepted: *09 March 2022*
Published: *29 March 2022*

Citation:
Liao Y-C, Wu H-C, Liou C-H, Lauderdale T-LY, Huang I-W, Lai J-F and Chen F-J (2022) Rapid and Routine Molecular Typing Using Multiplex Polymerase Chain Reaction and MinION Sequencer. Front. Microbiol. 13:875347. doi: 10.3389/fmicb.2022.875347

INTRODUCTION

Molecular diagnostics is widely used in clinical microbiology for routine detection and epidemiological analysis of infectious microorganisms (Chen et al., 2018). The invention of polymerase chain reaction (PCR) has led to remarkable developments in clinical molecular diagnostics because the use of PCR-based technologies requires relatively simple instrumentation and only small amounts of biological material (Maheaswari et al., 2016). PCR-based molecular diagnostic methods are important in studies of infectious diseases. For examples, 16S rRNA gene/internal transcribed spacer region sequencing is a well-established method for bacterial and fungal identification (Raja et al., 2017; Peker et al., 2019), and multilocus sequence typing (MLST) has become a commonly applied technique in molecular evolution studies of numerous microbial species (Jolley et al., 2018). Sanger sequencing is commonly used to obtain sequences of interest; however, the cost of hundreds of samples is prohibitive (Kircher and Kelso, 2010).

Next-generation sequencing (NGS) technologies (e.g., PacBio and Illumina) have been used to achieve high-throughput and cost-effective molecular diagnostics (Chen et al., 2015, 2018; Perez-Losada et al., 2018; Zhang et al., 2018; Peker et al., 2019), which has greatly affected clinical microbiology. However, the large costs associated with installing NGS instrumentation limit the accessibility of rapid and routine molecular typing in small- to medium-sized laboratories.

Oxford Nanopore Technologies (ONT) currently offer an inexpensive, pocket-sized MinION device that produces long sequences; however, the raw reads from this device are of lower accuracy in comparison with Illumina platform (Lin et al., 2021). In conjunction with consensus sequence generation and homopolymer correction, accurate molecular sequences can be obtained through nanopore sequencing (Liou et al., 2020). MinION sequencer provides the advantages of real-time data analysis, low capital cost, and highly accurate consensus sequence generation, all of which are adequately suited to the constraints of clinical settings (Sheka et al., 2021). Therefore, MinION sequencer has been used in numerous applications of clinical microbiology and infectious diagnostics (Ma et al., 2013; Benitez-Paez et al., 2016; Liou et al., 2020; Baldan et al., 2021; Ben et al., 2021; Ferreira et al., 2021; Sheka et al., 2021; Snell et al., 2021; Urban et al., 2021). However, the sample size for a MinION flow cell cannot exceed 96 due to the limitations of the barcoding kits. Although several studies have used tailored primers (Currin et al., 2019) or proposed dual-barcode systems (Liou et al., 2020) to address the problem of sample size, a requirement of additional efforts or costs remains inevitable. Furthermore, although ONT provides a Flow Cell Wash Kit, the influence of repeated washing and use of a MinION flow cell on the accuracy of a consensus sequence has yet to be comprehensively studied. With the recent release of the Rapid Barcoding Kit 96 (SQK-RBK110.96, released on March 2021) and the increasing demand for routine molecular diagnostics, evaluating the capabilities of a single MinION flow cell for accurate, timely, and routine molecular typing has become imperative.

Accordingly, in this study, we proposed a rapid protocol entailing the use of multiplex PCR of seven housekeeping genes and rapid barcoding of 392 *Staphylococcus aureus* isolates in conjunction with a MinION flow cell for sequencing to obtain a total of 3,269 kilobase-long consensus sequences. In addition to using Krocus (Page and Keane, 2018) for rapid MLST of *S. aureus*, we implemented nanoMLST2, which was modified from our previously proposed nanoMLST (Liou et al., 2020), for consensus sequence generation. Sixteen new alleles were identified by nanoMLST2 and validated with Sanger sequencing. The study results suggest that MinION nanopore sequencing of multiplex PCR amplicons could be a cost-effective method for rapid and routine molecular typing.

MATERIALS AND METHODS

Bacterial Isolates and DNA Extraction

A total of 392 *S. aureus* isolates were used in this study (designated as Sau 1–392). The isolates were collected from the Taiwan Surveillance of Antimicrobial Resistance program, a national surveillance program in Taiwan (Ho et al., 1999). Bacterial DNA templates from pure cultures were prepared using DNAzol Direct (Molecular Research Center, Inc. Cincinnati, OH, United States), according to the manufacturer's instructions. Of the 392 *S. aureus* isolates, 88 had been subjected to MinION nanopore sequencing to determine sequencing types (STs) in a previous study of 96 isolates (Liou et al., 2020) and were used in the second and the fourth runs as references to validate the accuracy of the workflow. Furthermore, 50 of these 88 isolates had DNA templates (i.e., 350 alleles), and two alleles, namely *pta*_664 and *glpF*_732, had been subjected to Sanger sequencing (Liou et al., 2020).

Multiplex PCR

Seven housekeeping genes were subjected to multiplex PCR using the Thermo Scientific Phusion High-fidelity DNA Polymerase kit (Thermo Fisher Scientific, Waltham, MA, United States) in a total volume of 25 µl (5 µl of 5× HF buffer, 2 µl of 2.5 mM dNTP, 10 µl of primer mix, 0.25 µl of Phusion enzyme, 1 µl of DNA template, and 6.75 µl of nuclease-free water). The primer sequences are listed in **Supplementary Table**. The primer mix included 10 µM each of forward and reverse primers of carbamate kinase (*arcC*), shikimate dehydrogenase (*aroE*), glycerol kinase (*glpF*), guanylate kinase (*gmk*), phosphate acetyltransferase (*pta*), triosephosphate isomerase (*tpi*), and acetyl coenzyme A acetyltransferase (*yqiL*) in a balanced ratio. The PCR program was set as follows: initial denaturation at 98°C for 30 s followed by 35 cycles of denaturation at 98°C for 10 s, annealing at 65°C for 30 s, and extension at 72°C for 1 min; and then a single final extension at 72°C for 10 min.

Library Preparation and Sequencing

The newly released Rapid Barcoding Kit 96 (SQK-RBK110.96) was used for the rapid barcoding of the 96 samples. Each sample was mixed with 5 µl of multiplexing PCR product, 2.5 µl of nuclease-free water, and 2.5 µl of one rapid barcode. The mixture was incubated at 30°C for 2 min, followed by incubation at 80°C for 2 min. All 96 barcoded DNA samples were pooled, and 120 µl of the pooled DNA was sampled and mixed with an equal volume of solid phase reversible immobilization beads (SPRI). After 5 min of incubation on a hula mixer, the barcoded DNA was cleaned twice with 240 µl of 80% ethanol and eluted with 30 µl of elution buffer (EB). An aliquot of 800 ng of barcoded DNA was used to make up a total volume of 11 µl with EB. One microliter of rapid adaptor F was added to the barcoded DNA, and the mixture was incubated at room temperature for 10 min. A pre-sequencing mix (PSM) was prepared by adding 37.5 µl of Sequencing Buffer II and 25.5 µl of loading beads to a 12 µl DNA library. The PSM was loaded *via* the SpotON port into a primed Flow cell (FLO-106MIN) for sequencing. Basecalling and de-multiplexing were performed in real time through MinKNOW GUI (v4.3.4) implemented with GPU Guppy (v5.0.11) on a desktop PC with an NVIDIA RTX 3090 graphics card with 24-GB RAM to produce high-accuracy reads in FASTQ format. The default output set for MinKNOW was a FASTQ file containing 4,000 reads.

Data Analysis

A sequencing run was stopped when individual FASTQ output files were obtained for each barcoded sample; this is because the derivation of such output files suggested more than 4,000 reads had been obtained for the sample. The FASTQ files obtained for each sample were collected and analyzed using Krocus 1.0.1 (Page and Keane, 2018) with a prepared directory named "Staphylococcus_aureus" which contained *S. aureus* MLST alleles and allelic profiles downloaded from PubMLST (Jolley et al., 2018). For each sample, consensus sequences were generated using Medaka v1.4.3[1] along with the FASTQ file and the reference sequences of seven housekeeping genes of *S. aureus* NCTC8325 (Liou et al., 2020). Samples with gene reads fewer than 40 were identified and labeled as "LSD (low sequencing depth)" by aligning the sequencing reads against the reference sequences using Minimap2 (v2.20; Li, 2018). Homopolymer errors registered for the consensus sequences were corrected, if necessary, to assign MLST alleles and to profile sequence types using a modified script, namely runtyping.py, in nanoMLST (Liou et al., 2020); this updated workflow was denoted as nanoMLST2, and it is available at https://github.com/jade-nhri/nanoMLST2.

Flow Cell Wash and Reuse

When 4,000 reads had been collected for each of the 96 samples, the sequencing experiment was stopped. The flow cell was left in the device. A flow cell wash mix was prepared by mixing 398 µl of wash diluent (DIL) and 2 µl of wash mix (WMX) from the Flow Cell Wash Kit (EXP-WSH004); this mixture was then loaded into the flow cell through the priming port. After 1 h of incubation at room temperature, 500 µl of storage buffer (S) was added through the priming port. The priming port was then closed to allow for the removal of all fluid from the waste channel through the waste port. The washed flow cell was stored at 4°C for reuse. The same cell was used five times on a total of 480 samples (96 samples per run, a total of 392 isolates). Notably, 88 PCR amplicons in the fourth run were aliquots from the second run but were barcoded with different barcodes. Another flow cell was used to ensure the reproducibility of this study.

Sanger Sequencing

Allele types, determined through Krocus and nanoMLST2 were compared to identify inconsistencies. The inconsistent alleles were further subjected to Sanger sequencing with conventional *S. aureus* MLST primers (Jolley et al., 2018).

RESULTS AND DISCUSSION

Rapid Library Preparation and Real-Time MinION Sequencing

In a previous study, a dual-barcoding system was established to multiplex 96 *S. aureus* isolates for seven housekeeping genes using 12 native barcodes in combination with 8×7 pairs of primers (Liou et al., 2020). The throughput of MinION nanopore sequencing, with careful electrophoresis and quantification processes, was estimated to be sufficient for 1,000 samples (Liou et al., 2020). However, the study used a labor-intensive and time-consuming process (**Figure 1A**); furthermore, ordering 96×7 pairs of primers solely for *S. aureus* MLST is impractical and cost prohibitive. Accordingly, we devised a new process involving multiplex PCR and the rapid barcoding of 96 isolates based on the newly released Rapid Barcoding Kit 96 (SQK-RBK110.96, released in March 2021) in 3 h (**Figure 1B**), as a preparation for MinION nanopore sequencing. A sequencing run was conducted with the objective of achieving 4,000 reads per sample; the five successive runs required 3.6, 3.9, 5, 9.8, and over 48 h (**Table 1**). A reduced pass rate was observed, which might have been engendered by the impairment of the integrity of the reused pores; the reduced pass rate along with the decrease in available pores may have increased unclassified rates and run times. Nevertheless, sufficient reads were available for the molecular typing of the samples in all five runs. As illustrated in **Figure 1B**, to simplify the process, the quantification steps were omitted before the pooling of the 96 samples. This rapid protocol requires minimal effort for quantification. To execute PCR, a DNA template (1 µl) was applied through a single-tube multiplex PCR assay (a total volume of 25 µl containing seven pairs of primers for *S. aureus* MLST). To achieve rapid barcoding, a 5 µl multiplex PCR product was mixed with a barcode. After the pooled DNA was sampled, only cleanup and quantification were required prior to the preparation of a PSM. Despite this simplification, among the 392 multiplex PCR products from the five runs, an extremely high (380/392 = 97%) success rate was observed for the amplification process in our protocol, with only 12 samples being labeled as low sequencing depth (LSD; exclusive of 88 samples—barcode01-barcode72 and barcode81-barcode96—in Run4 of **Figure 2A**); this can be attributed to the presence of samples with gene reads fewer than 40. Notably, 88 PCR products in Run4 (highlighted with background colors in **Figure 2A**) were aliquots of amplicons in Run2 but were barcoded with different barcodes. Among the 88 PCR products, 50 had been previously Sanger sequenced (Liou et al., 2020). They were used as references to evaluate whether carryover reads influence typing accuracy. Besides, the rapid protocol provided stable read counts for nearly every gene, except for those with high amounts of *yqi* (**Figure 3**). Although 88 PCR products in Run4 were aliquots of amplicons in Run2 but with different barcodes, the read counts of the 88 counterparts in Run2 and Run4 were not correlated ($R^2 = 0.017$); conversely, the read counts of the barcoded samples between runs were moderately correlated (R^2 for the correlation between R1 and Run2–Run5: from 0.289 to 0.555 and from 0.355 to 0.460 for the two flow cells, respectively). This suggests that some barcodes tend to have high or low read counts; for example, barcode30 and barcode56 had high read counts, but barcode42 and barcode89 had low read counts. This may be useful for executing ratio adjustment in order to obtain even distributions. Future research should focus on the refinement of multiplex primers to provide a narrow distribution of read counts among all genes. Our rapid protocol successfully amplified 380 out of 392 *S. aureus* isolates and required less than 3 h of library preparation per 96 samples, indicating that this protocol can facilitate rapid and routine molecular typing and can be easily adapted to different applications.

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

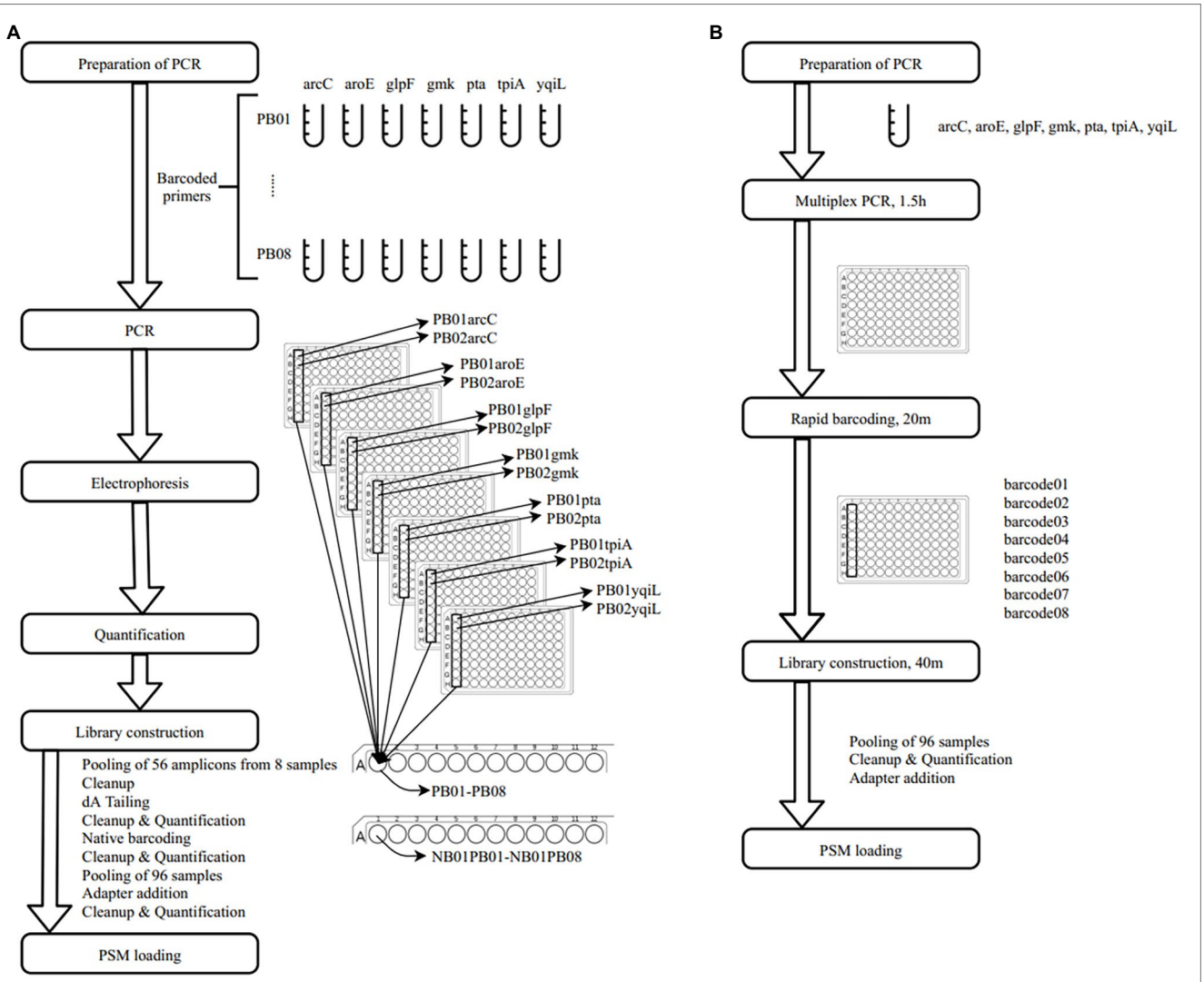

FIGURE 1 | Schematic workflows of **(A)** dual-barcode system (Liou et al., 2020) and **(B)** rapid multiplex polymerase chain reaction (PCR) and barcoding protocol.

TABLE 1 | Summary of MinION nanopore sequencing results.

Run	Available pores	Run time	Read counts	Passed reads	Passed (%)	Barcoded reads	Unclassified (%)	Average barcoded reads
Run1	830	3 h 39 m 9 s	934,114	829,160	88.8	786,375	5.2	8191.4 ± 1963.2
Run2	572	3 h 56 m 24 s	983,673	849,178	86.3	803,037	5.4	8365.0 ± 1911.2
Run3	510	5 h 3 m 47 s	1,021,596	843,838	82.6	795,398	5.7	8285.4 ± 1932.1
Run4	379	9 h 47 m 53 s	998,293	765,018	76.6	716,831	6.3	7467.0 ± 1552.1
Run5	325	48 h 2 m 24 s	1,092,802	764,021	69.9	707,887	7.3	7373.8 ± 1828.7

Rapid Sequence Typing Using Krocus

Through real-time basecalling and de-multiplexing in MinKnow, FASTQ files of the 96 barcoded amplicons were produced in only 4 h in the first two runs. Through parallel processing, Krocus was able to directly identify STs using uncorrected long reads within 10 min for all of the 96 barcoded samples simultaneously. This thus indicates that our proposed protocol can execute molecular typing for 96 multiplexed PCR amplicons within 8 h. **Table 2** presents the results obtained after executing Krocus on 4,000 reads per sample (as detailed in **Supplementary Table**). Among the 480 sequencing samples, Krocus predicted 454 STs with >99% coverage and labeled only 26 samples as "ND," indicating that they were untypable due to (1) low sequencing depth of genes, (2) novel combinations

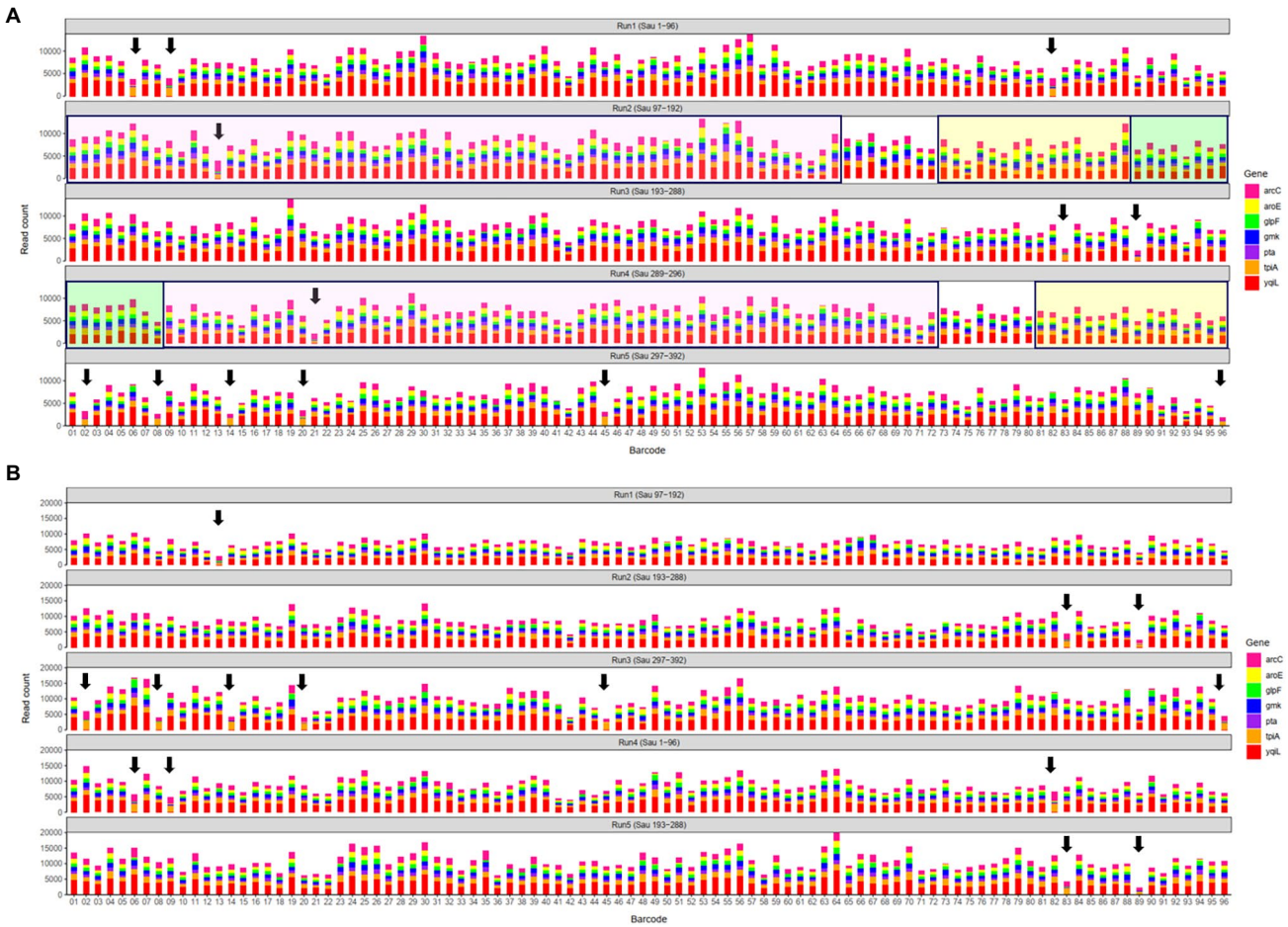

FIGURE 2 | Distributions of sequencing reads on the first flow cell **(A)** and the second flow cell **(B)**. Arrows indicate samples labeled as low sequencing depth (LSD). In **(A)**, 88 polymerase chain reaction (PCR) products in Run4 highlighted with background colors to indicate aliquots of amplicons in Run2 that were rapid barcoded with different barcodes; the other 8 PCR products in Run4 were from 8 isolates (Sau 289–296). Run1–Run5 in **(B)** containing identical PCR products in Run2, Run3, Run5, Run1, and Run3 in **(A)**, respectively.

of alleles, or (3) new alleles. Of the 26 samples labeled as ND, 13 with <99% coverage had previously been labeled as LSD. Because a total of 352 alleles had been previously Sanger sequenced (Liou et al., 2020), and these alleles ($50 \times 7 + 2$) were conducted separately in Run2 and Run4. Among the 704 alleles (352×2) subjected to Sanger sequencing, all were correctly predicted by Krocus, except one allele in barcode46 in Run2 was wrongly predicted to be *tpi*_58 rather than *tpi*_26. Krocus incorrectly predicted this sample (barcode46, ST398) to be ND with 100% coverage. Nevertheless, MinION nanopore sequencing coupled with Krocus provided a 98.86% accuracy (703/704 = 99.86%) in allele typing. In addition, Krocus predicted two other samples to be ND with 100% coverage in Run2 and in Run4 (**Table 2**); this could be attributed to a submission of new alleles (*pta*_664 and *glpF*_732; Liou et al., 2020) without corresponding ST information to PubMLST (Jolley et al., 2018). Finally, of the 26 samples labeled as ND, the other 13 were predicted by Krocus to be ND with ≧99% coverage; these samples were further evaluated to identify the presence of either a novel combination of alleles or new alleles (**Table 3**).

New Allele Types Identified by NanoMLST2

The sequencing reads of 467 samples, excluding those labeled LSD, were analyzed for consensus sequence generation and MLST typing using nanoMLST2. The results obtained through nanoMLST2 were similar to those obtained using Krocus, and 100% accuracy was observed in the 704 Sanger-sequenced alleles in Run2 and Run4 (**Supplementary Table**). This perfect value not only suggests the accuracy of nanoMLST2, but also means no effects of the carryover reads on the MLST typing of the following runs. Through a comparison of the results obtained using Krocus and nanoMLST2, 29 samples containing 24 alleles were identified to be inconsistent between the methods (**Table 3**). Nevertheless, nanoMLST2 had exceptionally good agreement with Krocus (3,245/3,269 = 99.27%). The inconsistent alleles were later sequenced using Sanger. As indicated in **Table 3**, the 16 new alleles identified through nanoMLST2

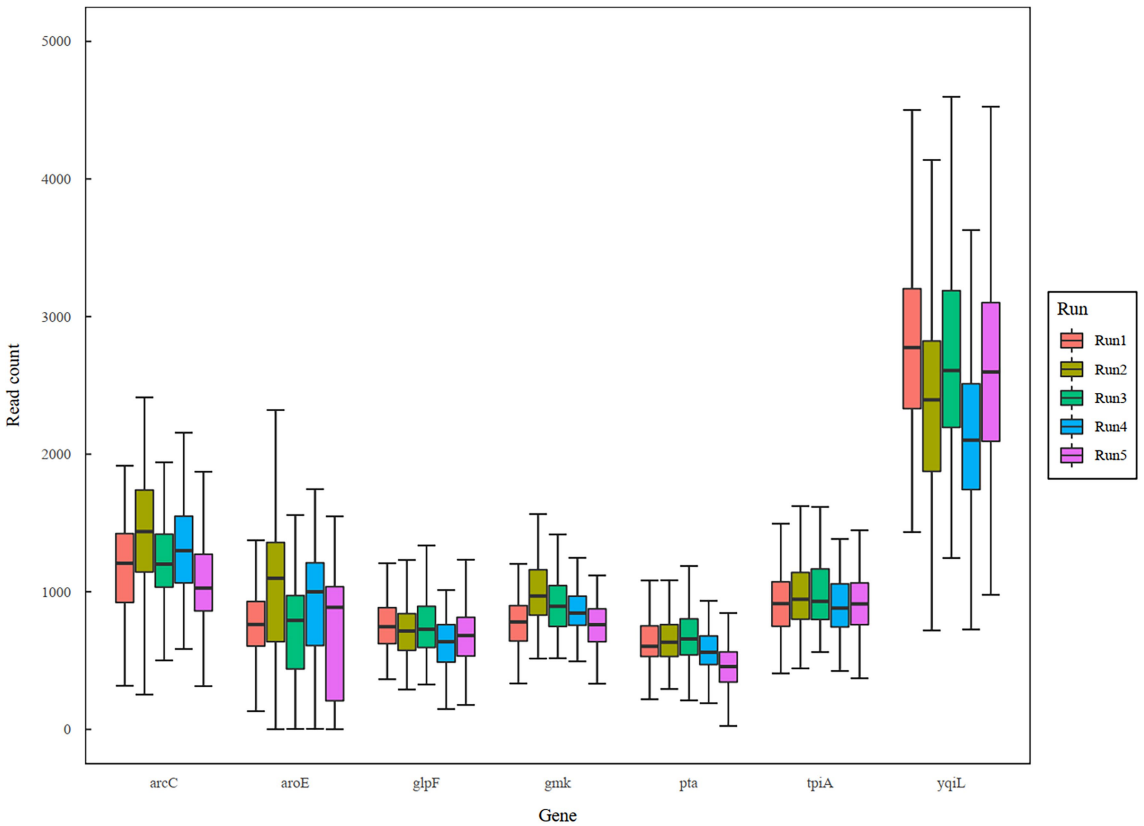

FIGURE 3 | Boxplot of sequencing reads across genes and runs.

TABLE 2 | Krocus sequence type (ST) prediction for *Staphylococcus aureus*.

Krocus result		Number of samples[2]				
Coverage	**Prediction[1]**	**Run1**	**Run2**	**Run3**	**Run4**	**Run5**
100	ST	83	91	89	91	85
≧99	ST	8	1	2	2	2
100	ND	2	3	3	2	1
≧99	ND	0	0	0	0	2
<99	ND	**3**	**1**	**2**	**1**	**6**

[1]ST: sequence type predicted by Krocus; ND: sequence type not determined by Krocus.

[2]Boldface indicates the number of samples labeled as low sequencing depth (LSD).

were all validated using Sanger sequencing to ensure accuracy. In addition, the consensus sequences generated by nanoMLST2 were full-length genes ranging from 1,067 to 1,489 bp.

Krocus was used to predict *S. aureus* STs directly from uncorrected reads (Page and Keane, 2018). However, in addition to the expected predictive failures in 22 samples (**Table 3**) with 16 new alleles and 6 new STs, Krocus wrongly classified eight alleles (indicated by strikethroughs) in seven samples. As listed in **Table 3**, Krocus frequently predicted *glpF*_344 to *glpF*_19 and *tpi*_26 to *tpi*_58, while these two pairs differ by one nucleotide. In contrast to Krocus, nanoMLST2 could generate consensus sequences for new alleles, and all of the sequences were later validated through Sanger. Therefore, the findings of this study indicate that the benefits gained through consensus sequence generation of MinION nanopore sequencing may address the needs of a wide range of kilobase-long molecular typing.

Capacity of MinION for Routine Molecular Typing

In this study, more than 3,000 kilobase-long consensus sequences conducted in five batches (Run1 to Run5) were obtained using a single MinION flow cell. Although the flow cell was washed and reused for the successive runs (Run2 to Run5), the variation between the STs in the different batches indicates that the effects of the cell reuse were marginal. For example, the STs of barcode10 in Run1, Run2, Run3, Run4, and Run5 were 15, 6, 188, 254, and 59, respectively. However, consistent STs were obtained through Sanger (STs of 6 and 254 in Run2 and Run4), demonstrating that the STs were not influenced by the preceding runs. Regarding rapid and routine molecular typing, the results of the first three runs of the 96 samples (Run1 to Run3) were obtained 4–6 h after the initiation of the sequencing process (3.6–5 h run time plus 0.5 h data analysis; **Table 1**). The remaining two runs (Run4 and Run5) required a longer sequencing period because of the reduced number of available pores (<400) in

TABLE 3 | Alleles prediction inconsistencies between Krocus and nanoMLST2.

Run	BC	Krocus			nanoMLST2[2]	Sanger[3]
		ST	Cov	Allele[1]		
1	03	15	99.68	*arcC*(13)*	**New allele**	*arcC*_826
1	16	623	99.65	*pta*(4)*	**New allele**	*pta*_842
1	32	7	99.71	*tpi*(6)*	**New allele**	*tpi*_786
1	36	59	100	*arcC*(19)	**New allele**	*arcC*_834
1	43	ND	100	~~*glpF*(19), *tpi*(58)~~	***glpF*_344, *tpi*_26**	*glpF*_344, *tpi*_26
1	59	1	99.71	*tpi*(1)*	**New allele**	*tpi*_787
1	65	25	99.71	*glpF*(4)*	**New allele**	*glpF*_890
1	70	22	99.77	*aroE*(6)*	**New allele**	*aroE*_1016
1	72	239	99.77	*arcC*(2)*	**New allele**	*arcC*_820
1	93	ND	100		New ST	
1	94	5,535	99.87	*gmk*(438)*	**New allele**	*gmk*_554
2	46	ND	100	~~*tpi*(58)~~	***tpi*_26**	*tpi*_26[#]
2	77	ND	100	***pta*(664)**	**New ST**	*pta*_664[#]
2	89	ND	100	***glpF*(732)**	**New ST**	*glpF*_732[#]
3	13	59	100	*tpi*(20)	**New allele**	*tpi*_781
3	29	188	100	*tpi*(1)	**New allele**	*tpi*_790
3	39	398	100	~~*glpF*(19)~~	***glpF*_344**	*glpF*_344
3	41	ND	100		New ST	
3	42	ND	100	~~*tpi*(58)~~	***tpi*_26**	*tpi*_26
3	48	291	100	*tpi*(26)	**New allele**	*tpi*_791
3	51	ND	100	~~*tpi*(58)~~	***tpi*_26**	*tpi*_26
3	80	4,567	99.65	*aroE*(35)*	**New allele**	*aroE*_1022
4	01	ND	100	***glpF*(732)**	**New ST**	*glpF*_732[#]
4	85	ND	100	***pta*(664)**	**New ST**	*pta*_664[#]
5	04	188	99.68	*pta*(1)*	**New allele**	*pta*_844
5	35	ND	99.94	~~*tpi*(58)*~~	***tpi*_26**	*tpi*_26
5	36	ND	100	~~*tpi*(58)~~	***tpi*_26**	*tpi*_26
5	37	ND	99.74	*glpF*(213)*	**New allele**	*glpF*_891
5	59	7	100	*arcC*(5)	**New allele**	*arcC*_830

[1]*Asterisks represent alleles partially covered using Krocus (Page and Keane, 2018), strikethroughs represent wrong predictions, and boldface represents allele prediction consistency between Krocus and Sanger sequencing.*

[2]*Boldface represents consistency in consensus sequences between nanoMLST2 and Sanger sequencing.*

[3]*Underlines represent new Sanger-sequenced alleles identified in this study, and # represents alleles that were Sanger sequenced previously (Liou et al., 2020).*

the used MinION; nevertheless, the generated consensus sequences remained accurate for routine molecular typing. Another flow cell containing more than 1,500 available pores in the beginning was also used in this study, shorter sequencing periods (1.6–3.5 h) were required for the five runs. As shown in **Figure 2**, the aliquots of amplicons from the first flow cell Run2, Run3, Run5, Run1, and Run3 were used in the second flow cell Run1–Run5, respectively. Although the sequencing components in the preceding run varied, all the STs obtained from the second flow cell were identical to that of the corresponding samples in the first flow cell (**Supplementary Table**), which again reveals that the effects of the cell reuse were marginal and the STs were not influenced by the preceding runs.

Compared with the approach used in our previous study (Liou et al., 2020), the proposed protocol in this study not only reduced the primer cost significantly but also reduced the PCR reagents and amounts of manual effort by seven times. Library preparation in this protocol cost US$110 for a total of 96 samples; by contrast, the cost incurred by ligation-based procedures (Liou et al., 2020) for such a run is US$148, regardless of whether third-party consumables such as AmpureXP beads and NEB End Repair/dA-Tailing enzymes are required (Liou et al., 2020). The MinION flow cell used in our protocol could produce accurate typing results for a total of 3,269 kilobase-long genes, in addition to affording less expensive and more rapid multiplexing PCR and library preparation. This protocol allows for a more efficient and cost-effective method for routine molecular typing at an estimated cost of US$4 per sample (**Supplementary Table**). Specifically, through the protocol, the cost of a kilobase-long gene would be less than US$1, which is substantially less than that in Sanger sequencing. Overall, our results demonstrate the effectiveness of the portable MinION sequencer in providing accurate, rapid, and routine molecular typing.

CONCLUSION

The features of friendly access (USD$1,000 for a starter pack), portability and the ability to monitor real-time output and reuse of a flow cell of the ONT MinION sequencer remove the barriers of accessing accurate, rapid, and routine molecular typing in small- to medium-sized laboratories. To the best of our knowledge, our study is the first to investigate the reusability of a MinION flow cell and to provide the evidence of the sequencing accuracy of a reused flow cell. In this study, a workflow was designed entailing the use of one universal primer

set with seven primer pairs of housekeeping genes to amplify full-length target genes simultaneously for *S. aureus* MLST and rapid barcoding in conjunction with ONT nanopore sequencing on a portable MinION platform followed by the nanoMLST2 analysis. Our results indicate that the benefits gained through consensus sequence generation using this workflow may address the needs of a wide range of kilobase-long molecular typing.

DATA AVAILABILITY STATEMENT

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

https://doi.org/10.6084/m9.figshare.19179107
https://doi.org/10.6084/m9.figshare.19179110
https://doi.org/10.6084/m9.figshare.19179116
https://doi.org/10.6084/m9.figshare.19179122
https://doi.org/10.6084/m9.figshare.19179125
https://doi.org/10.6084/m9.figshare.19179128.

AUTHOR CONTRIBUTIONS

Y-CL, H-CW, and F-JC conceived the study. Y-CL and H-CW designed the methodology. Y-CL implemented the pipeline. H-CW and C-HL validated data. Y-CL, H-CW, C-HL, and F-JC investigated. T-LL, I-WH, and J-FL provided the strains and participated in discussion of the study. Y-CL, H-CW, T-LL, and F-JC wrote the manuscript. Y-CL and F-JC supervised and acquired funding. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by intramural grants from National Health Research Institutes (IV-110-PP-06 to F-JC and PH-110-PP-05 to Y-CL) and Ministry of Science and Technology (MOST 110-2314-B-400-038).

ACKNOWLEDGMENTS

This manuscript was edited by Wallace Academic Editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.875347/full#supplementary-material

REFERENCES

Baldan, R., Cliff, P. R., Burns, S., Medina, A., Smith, G. C., Batra, R., et al. (2021). Development and evaluation of a nanopore 16S rRNA gene sequencing service for same day targeted treatment of bacterial respiratory infection in the intensive care unit. *J. Infect.* 83, 167–174. doi: 10.1016/j.jinf.2021.06.014

Ben, C. S., Filloux, D., Fernandez, E., Moubset, O., Hoareau, M., Julian, C., et al. (2021). Nanopore sequencing is a credible alternative to recover complete genomes of Geminiviruses. *Microorganisms* 9:903. doi: 10.3390/microorganisms9050903

Benitez-Paez, A., Portune, K. J., and Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the min ION portable nanopore sequencer. *Gigascience* 5:4. doi: 10.1186/s13742-016-0111-z

Chen, Y., Frazzitta, A. E., Litvintseva, A. P., Fang, C., Mitchell, T. G., Springer, D. J., et al. (2015). Next generation multilocus sequence typing (NGMLST) and the analytical software program MLSTEZ enable efficient, cost-effective, high-throughput, multilocus sequencing typing. *Fungal Genet. Biol.* 75, 64–71. doi: 10.1016/j.fgb.2015.01.005

Chen, J. W., Lau, Y. Y., Krishnan, T., Chan, K. G., and Chang, C. Y. (2018). Recent advances in molecular diagnosis of *Pseudomonas aeruginosa* infection by state-of-the-art genotyping techniques. *Front. Microbiol.* 9:1104. doi: 10.3389/fmicb.2018.01104

Currin, A., Swainston, N., Dunstan, M. S., Jervis, A. J., Mulherin, P., Robinson, C. J., et al. (2019). Highly multiplexed, fast and accurate nanopore sequencing for verification of synthetic DNA constructs and sequence libraries. *Synth. Biol.* 4:ysz025. doi: 10.1093/synbio/ysz025

Ferreira, F. A., Helmersen, K., Visnovska, T., Jorgensen, S. B., and Aamot, H. V. (2021). Rapid nanopore-based DNA sequencing protocol of antibiotic-resistant bacteria for use in surveillance and outbreak investigation. *Microb. Genom.* 7:000557. doi: 10.1099/mgen.0.000557

Ho, M., McDonald, L. C., Lauderdale, T. L., Yeh, L. L., Chen, P. C., and Shiau, Y. R. (1999). Surveillance of antibiotic resistance in Taiwan, 1998. *J. Microbiol. Immunol. Infect.* 32, 239–249. PMID: 10650488

Jolley, K. A., Bray, J. E., and Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* 3:124. doi: 10.12688/wellcomeopenres.14826.1

Kircher, M., and Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. *BioEssays* 32, 524–536. doi: 10.1002/bies.200900181

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

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

Liou, C. H., Wu, H. C., Liao, Y. C., Yang Lauderdale, T. L., Huang, I. W., and Chen, F. J. (2020). nanoMLST: accurate multilocus sequence typing using Oxford Nanopore Technologies MinION with a dual-barcode approach to multiplex large numbers of samples. *Microb Genom* 6:e000336. doi: 10.1099/mgen.0.000336

Ma, L., Lu, P. L., Siu, L. K., and Hsieh, M. H. (2013). Molecular typing and resistance mechanisms of imipenem-non-susceptible *Klebsiella pneumoniae* in Taiwan: results from the Taiwan surveillance of antibiotic resistance (TSAR) study, 2002-2009. *J. Med. Microbiol.* 62, 101–107. doi: 10.1099/jmm.0.050492-0

Maheaswari, R., Kshirsagar, J. T., and Lavanya, N. (2016). Polymerase chain reaction: A molecular diagnostic tool in periodontology. *J. Indian Soc. Periodontol.* 20, 128–135. doi: 10.4103/0972-124X.176391

O.N. Medaka Technologies. Available at: https://github.com/nanoporetech/medaka (Accessed 2021).

Page, A. J., and Keane, J. A. (2018). Rapid multi-locus sequence typing direct from uncorrected long reads using Krocus. *PeerJ* 6:e5233. doi: 10.7717/peerj.5233

Peker, N., Garcia-Croes, S., Dijkhuizen, B., Wiersma, H. H., van Zanten, E., Wisselink, G., et al. (2019). A comparison of three different bioinformatics analyses of the 16S-23S rRNA encoding region for bacterial identification. *Front. Microbiol.* 10:620. doi: 10.3389/fmicb.2019.00620

Perez-Losada, M., Arenas, M., and Castro-Nallar, E. (2018). Microbial sequence typing in the genomic era. *Infect. Genet. Evol.* 63, 346–359. doi: 10.1016/j.meegid.2017.09.022

Raja, H. A., Miller, A. N., Pearce, C. J., and Oberlies, N. H. (2017). Fungal identification using molecular tools: A primer for the natural products

research community. *J. Nat. Prod.* 80, 756–770. doi: 10.1021/acs.jnatprod.6b01085

Sheka, D., Alabi, N., and Gordon, P. M. K. (2021). Oxford nanopore sequencing in clinical microbiology and infection diagnostics. *Brief. Bioinform.* 22:bbaa403. doi: 10.1093/bib/bbaa403

Snell, L. B., Cliff, P. R., Charalampous, T., Alcolea-Medina, A., Ebie, S. A. R. T., Sehmi, J. K., et al. (2021). Rapid genome sequencing in hospitals to identify potential vaccine-escape SARS-CoV-2 variants. *Lancet Infect. Dis.* 21, 1351–1352. doi: 10.1016/s1473-3099(21)00482-5

Urban, L., Holzer, A., Baronas, J. J., Hall, M. B., Braeuninger-Weimer, P., Scherm, M. J., et al. (2021). Freshwater monitoring by nanopore sequencing. *elife* 10:e61504. doi: 10.7554/eLife.61504

Zhang, N., Wheeler, D., Truglio, M., Lazzarini, C., Upritchard, J., McKinney, W., et al. (2018). Multi-locus next-generation sequence typing of DNA extracted From pooled colonies detects multiple unrelated *Candida albicans* strains in a significant proportion of patient samples. *Front. Microbiol.* 9:1179. doi: 10.3389/fmicb.2018.01179

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: 29 March 2022
doi: 10.3389/fmicb.2022.829152

Response of Fungal Sub-Communities in a Maize-Wheat Rotation Field Subjected to Long-Term Conservation Tillage Management

Cunzhi Zhang[1†], Hao Liu[1†], Senlin Liu[1], Sarfraz Hussain[1,2], Liting Zhang[1], Xiaowei Yu[1], Kaixun Cao[1], Xiuli Xin[3], Hui Cao[1] and Anning Zhu[3]**

[1] Key Laboratory of Agricultural Environmental Microbiology, Ministry of Agriculture and Rural Affairs, College of Life Sciences, Nanjing Agricultural University, Nanjing, China, [2] Key Laboratory of Integrated Regulation and Resource Development on Shallow Lakes of Ministry of Education, College of Environment, Hohai University, Nanjing, China, [3] Fengqiu Agro-Ecological Experimental Station, State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing, China

Conservation tillage is an advanced agricultural technology that seeks to minimize soil disturbance by reducing, or even eliminating tillage. Straw or stubble mulching in conservation tillage systems help to increase crop yield, maintain biodiversity and increase levels of exogenous nutrients, all of which may influence the structure of fungal communities in the soil. Currently, however, the assembly processes and co-occurrence patterns of fungal sub-communities remain unknown. In this paper, we investigated the effects of no-tillage and straw mulching on the composition, assembly process, and co-occurrence patterns of soil fungal sub-communities in a long-term experimental plot (15 years). The results revealed that combine straw mulching with no-tillage significantly increased the richness of fungi but not their diversity. Differential abundance analysis and principal component analysis (PCA) indicated that tillage management had a greater effect on the fungal communities of abundant and intermediate taxa than on the rare taxa. Available phosphorus (AP) and total nitrogen (TN) were the major determinants of fungal sub-communities in NT treatment. The abundant fungal sub-communities were assembled by deterministic processes under medium strength selection, while strong conservation tillage strength shifts the abundant sub-community assembly process from deterministic to stochastic. Overall, the investigation of the ecological network indicated that no-tillage and straw mulching practices decreased the complexity of the abundant and intermediate fungal networks, while not significantly influencing rare fungal networks. These findings refine our knowledge of the response of fungal sub-communities to conservation tillage management techniques and provide new insights into understanding fungal sub-community assembly.

Keywords: conservation tillage, fungal sub-communities, assembly process, community structure, co-occurrence pattern

OPEN ACCESS

Edited by:
Pengfei Ding,
University of Maryland, Baltimore County, United States

Reviewed by:
Weiyan Wang,
Northwest A&F University, China
Hao Wang,
Northwest A&F University, China
Xiaojing Hu,
Northeast Institute of Geography and Agroecology (CAS), China

***Correspondence:**
Hui Cao
hcao@njau.edu.cn
Anning Zhu
anzhu@issas.ac.cn

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

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

Received: *05 December 2021*
Accepted: *08 February 2022*
Published: *29 March 2022*

Citation:
Zhang C, Liu H, Liu S, Hussain S, Zhang L, Yu X, Cao K, Xin X, Cao H and Zhu A (2022) Response of Fungal Sub-Communities in a Maize-Wheat Rotation Field Subjected to Long-Term Conservation Tillage Management. Front. Microbiol. 13:829152. doi: 10.3389/fmicb.2022.829152

INTRODUCTION

Conservation tillage management, such as no-tillage (NT) or reduced tillage (RT) and straw retention, has been a widely-used practice in agriculture systems. Indeed, it is estimated that an area of 155 million hectares is subject to no-tillage management, accounting for 11% of all arable farmland worldwide (Kassam et al., 2014). Conservation tillage can conserve plant-available water and reduce soil erosion caused by rain, wind, and feed soil biota by increasing soil nutrient (Balesdent et al., 2000; Hobbs et al., 2008). On the other hand, soil microbes are crucial for nitrogen cycling and soil fertility enhancement, and are influenced by NT and straw mulching practices (Levy-Booth et al., 2014; Joergensen and Wichern, 2018).

Conservation tillage (e.g., straw mulching and NT) accumulates more C and N sources in the soil, and has the potential to minimize soil disturbance and enhance soil aggregation, thus creating a favorable soil nutrition condition for soil microbial communities (Guo et al., 2015; Wang et al., 2019). Long-term straw mulching and NT have been shown to significantly increase soil pH, total carbon and the C:N ratio (Wang et al., 2020a; Zhou et al., 2021); these environmental factors, in turn, can significantly influence microbial diversity. For example, soil pH is a strong predictor of microbial community diversity and structure (Hartmann et al., 2015). Whereas conservation tillage effects on soil properties have been well studied, however, there are still lack of effects of environmental factors on fungal community. In this study, we investigated the key regulatory factors of fungal sub-communities under different conservation tillage strategies.

In addition, the assembly process of the soil microbial community is crucial for understanding the response of ecosystems to environmental changes. In that regard, both stochastic processes and deterministic processes influence community assembly (Vellend, 2010; Tripathi et al., 2018). On the one hand, changes in the environmental conditions influence biotic and abiotic filtering and the structure of the microbial community deterministically (Chesson, 2000; Vellend, 2010). On the other hand, community assembly patterns arising from processes dispersal and ecological drift occur stochastically (Chave, 2004). The fungal communities in agricultural soils were strongly affected by stochastic processes (Jiao et al., 2021). In different tillage managements, assembly processes have been investigated in rhizosphere microorganism, including diazotrophic community (Li et al., 2021), arbuscular mycorrhizal fungi (Wang et al., 2020b), bacterial community (Wang et al., 2020c). However, stochastic and deterministic processes of fungal sub-communities under long-term conservation tillage management have not yet been clarified. Therefore, we sought to identify the assembly processes of fungal sub-communities across four different treatments.

Co-occurrence network analysis has been recently used to elucidate the potential complex interaction among different taxonomic group associated with patterns of assembly process (Li et al., 2021). Recent study has found that agricultural intensification decreased the complexity of fungal network, and the abundance of mycorrhizal fungi was highest under organic farming rather than no-tillage and conventional practices (Banerjee et al., 2019). Straw mulching has also been shown to reduce the complexity of fungal network, and increased the risk of root rot by increasing the abundance of the soil-borne pathogens *F. graminearum* and *F. moniliforme* (Wang et al., 2020a). Previous study reported that NT practices had higher densities of fungal mycelia than CT treatment (Beare et al., 1997). NT practices, meanwhile, may result in soil compaction, and plow tillage strengthens the fungal-fungal interactions and reduced tillage (RT) induces a more stable network structure than NT (Hu et al., 2021). The soil fungal sub-community co-occurrence patterns in long-term conservation tillage field remain unknown. In this paper, we compared the changes in the co-occurrence patterns of rare, intermediate and abundant sub-communities under different tillage and straw mulching practices.

Overall, while the effects of conservation tillage on fungal community diversity and functional group have been well documented (Degrune et al., 2016; Schmidt et al., 2019), there is still limited knowledge about fungal sub-communities. Furthermore, the research needs to pay more attention to the fact that microbial communities tend to consist of a few highly abundant taxa and numerous intermediate and rare taxa. It is therefore incomplete to analyze the microbial community in broad groups (e.g., at a domain or kingdom level); both abundant and rare groups should be considered if the community dynamics are to be fully understood (Jiang et al., 2019).

In this research, therefore, a field experiment applying fungal ITS region sequencing was conducted in a long-term conservation trial field at the Fengqiu National agro-ecological experiment station. We hypotheses that (i) conservation tillage managements create different environments for fungi, which increase the correlation between environment factors and fungal sub-communities, and deeply change fungi community structure and composition, (ii) the different ecological environments significantly alter the assembly processes of fungal sub-communities along with conservation tillage strength, (iii) different tillage and straw managements change the network structure for fungi sub-communities. The findings may provide a theoretical and practical foundation for sustainable agriculture development from a microbial ecology perspective.

MATERIALS AND METHODS

Site Description and Soil Sampling

The study site is situated in Fengqiu National Agro-Ecological Experimental Station (35°00′N, 114°24′E), Chinese Academy of Sciences, Henan province, Central China. This area has a typical temperate continental monsoon climate, with an average annual temperature of 13.9°C and an average annual rainfall of 615 mm. The test soil is classified as Aquic Inceptisol, which is developed from the alluvial sediments of the Yellow River.

The long-term conservation field was commenced in 2006 and based on a completely randomized block design with three replications. The current experiment was set up in a maize-wheat crop rotation with four treatments: (1) tillage for wheat and no-tillage for corn (conventional tillage, CT), (2) tillage for wheat and

no-tillage for corn plus straw mulching (CTS), (3) no-tillage for wheat and corn (NT), (4) no-tillage for wheat and corn coupling straw mulching (NTS). Regarding tillage practice, soils were plowed to 20–22 cm depth with a moldboard plow. Regarding straw mulching, residues were crushed into 2–3 cm pieces for maize and 6–7 cm pieces for wheat, and then were spread evenly on the soil surface as mulch. The amount of straw was related to the yield of the plot. As for no-straw mulching treatments, all residues were removed from the plots. Each experimental plot was 14 m × 6.5 m in size. Three soil samples were randomly collected from each plot from the surface layer of soil (0–20 cm) using a sterile soil driller. The three samples were immediately mixed to form a composite soil sample. The composite samples were sieved through a 2 mm sieve so as to homogenize them and remove plant roots and stones, before being transferred to the laboratory for storage at 4°C for measuring soil physicochemical properties, and at −80°C for DNA extraction.

Analysis of Soil Physical and Chemical Properties

The physicochemical properties of all twelve soil samples were determined. Soil pH was determined using a pH meter (FE20-Five Easy PlusTM, Switzerland) with a 1:2.5 soil/water mixture. In addition, we measured the total organic carbon (TOC), total phosphorus (TP), alkaline nitrogen (AN), total nitrogen (TN), total potassium (TK), available phosphorus (AP), and available potassium (AK), following the method in Bao (2000).

DNA Extraction and ITS Sequencing

Genomic DNA was extracted from approximately 0.5 g of fresh soil using a Fast DNA^{TM} Spin Kit (MP Biomedicals, United States), following the manufacturer's instructions. The quality of the extracted DNA was determined using a NanoDrop 2000 Spectrophotometer (Thermo Scientific, Wilmington, DE, United States). The fungal-specific ITS1 region was amplified with the ITS1F (CTTGGTCATTTAGGAAGTAA) and ITS2 (GCTGCGTTCTTCATCGATGC) primer sets (Mueller et al., 2014). The following PCR conditions were used: initial denaturation at 94°C for 30 s, 35 cycles consisting of 15 s of denaturation at 98°C, 30 s of annealing at 55°C, followed by 45 s at 72°C, and a final extension for 10 min at 72°C (Schmidt et al., 2019). Pooled PCR products were purified using the GeneJET^{TM} Gel Extraction Kit (Thermo Scientific, United States). Finally, Personalbio Biotechnology Institute (Shanghai, China) sequenced the purified products using an Illumina Miseq platform (Illumina, United States).

Processing of Sequence Analysis

The microbiome sequences was processed using QIIME pipeline (Version 1.9.0) (Caporaso et al., 2012), and low-quality sequencing reads with a length shorter than 150 bp, and with an average base quality score < 20 were discarded from further analysis. Operational taxonomic units (OTUs) were clustered at a ≥ 97% similarity level using the UCLUST feature in QIIME 1 (Edgar et al., 2011). Taxonomic identification of representative sequences was performed using the BLAST database. Genomic sequencing data has been deposited in the NCBI Sequence Read Archive (BioProject ID PRJNA764374, submission ID SUB10399701).

Statistical and Bioinformatics Analysis

We defined rare, intermediate and abundant fungal sub-communities to support the understanding of fungal community variation. OTUs with relative abundance above 0.5% were defined as "abundant," while those below 0.01% were defined as "rare." Those taxa with relative abundance between 0.5 and 0.01% OTUs were defined as "intermediate." We calculated the relative abundance of these sub-communities across all samples. This definition was based on previous research (Liu et al., 2015; Jiang et al., 2019).

Alpha-diversity indices (Shannon and Chao1) were calculated using MOTHUR. Significant differences in α- diversity and in soil properties were analyzed by ANOVA using SPSS (Version 21.0, SPSS, Chicago, IL, United States). Fungal sub-community alpha-diversity indices were calculated using the R package "vegan," and analyzed by one-way ANOVA in SPSS.

Volcano plots were used to show the differential abundance of OTUs. We selected the false discovery rate (FDR) to adjust *p*-values (Love et al., 2014). The $\log_2$ Fold Change ($\log_2$FC) and adjusted *p*-values were calculated using the R package "DESeq2." The OTUs' differential abundance plots were constructed using the R package "ggplot2." PCA was performed using the "Principal Component analysis" application in the OriginPro 2021 (Version 9.80).

Linear discriminant analysis (LDA), combined with effect size (LEfSe) measurements, were performed in order to find statistical biomarkers between treatments (Segata et al., 2011). This analysis was conducted on the Hutlab Galaxy website application[1]. Spearman correlations in R (Version 4.1.1) were used to evaluate the relationship between various environmental factors. The correlation between fungal sub-communities and environmental factors were normalized by using the Mantel test (Diniz-Filho et al., 2013).

The assembly processes of fungal sub-communities were constructed using the "ses.comdistnt" function in the "MicEco" package (Stegen et al., 2012), with the beta mean nearest taxon distance (βMNTD) metric used to determine turnover in the phylogenetic structure of community. Meanwhile, the stochastic and deterministic ecological processes of fungal sub-communities were evaluated through null model analysis (Stegen et al., 2012). This method uses randomizations to estimate the standard deviation of the observed βMNTD compared to a null distribution (999 randomizations) for each βMNTD estimate. The β-nearest taxon index (βNTI) was used to evaluate the deviation between the mean of the null βMNTD and observed βMNTD, expressed in units of standard deviations (Stegen et al., 2013).

In order to analyze the fungal assembly processes quantitatively, we determined the proportion of dispersal limitation and undominated for stochastic processes, and variable selection and homogeneous selection for deterministic

[1]http://huttenhower.sph.harvard.edu/galaxy/

processes. This was done in the "iCAMP" R package (Ning et al., 2020). Variations were investigated by comparing β-diversity metrics (βNTI values and RC_{bray}). Meanwhile, the influences of the variable selection and homogeneous selection fractions were determined according to thresholds of βNTI values > 2 and < -2, respectively. The relative influence of dispersal limitation was quantified as a pairwise comparison between $|\beta NTI| < 2$ and $RC_{bray} > 0.95$, whereas $|\beta NTI| < 2$ and $RC_{bray} < 0.95$ was used to estimate the influence of undominated process (Stegen et al., 2013, 2015).

Microbial networks were created to construct the interaction networks for tillage (CT and CTS) treatments and no-tillage (NT and NTS) treatments, straw mulching (CTS and NTS) and no-straw mulching (CT and NT). In our study, abundant, intermediate and rare OTUs were used for network construction. The Spearman's rank correlation coefficients were calculated using the R package "psych." The correlation coefficients ($P < 0.05$ and $r > 0.6$) were corrected for analysis of network. The network graphs were visualized and the topological properties of the network were calculated using Gephi software (Version 0.9.2[2]). OTUs in the networks were identified as potential plant pathogens using FUNGuild tool (Nguyen et al., 2016).

RESULTS

Richness and Diversity Indices of Fungal Communities

The Shannon estimator for fungal diversity was found to be significantly greater in conventional tillage (CT) than in no-tillage and straw mulching treatments (CTS, NT, and NTS). While the Chao1 richness index was markedly lower in the conventional treatment (CT) than in the conservation tillage groups (CTS, NT, and NTS) (**Figure 1**). We discovered that the richness indices had significantly higher values in the plots subject to straw retention practices (CTS and NTS) than in those subject to other tillage practices (CT and NT).

Depleted and Enriched Operational Taxonomic Units Response to Different Tillage Systems, and Variations of Fungal Community Structure

We conducted groups comparison analysis to identify OTUs where abundance was strongly influenced by straw mulching (CTS vs. CT and NTS vs. NT) and tillage (CT vs. NT and CTS vs. NTS). OTUs where relative abundance significantly increased or decreased were referred to as "Enriched OTUs" and "Depleted OTUs," respectively. These enriched and depleted OTUs were found to occur only in abundant and intermediate taxa (**Figure 2**). There were, respectively, 12 and 46 enriched OTUs in the "CTS vs. CT" group, and 8 and 27 enriched OTUs were in the "NTS vs. NT" group (**Figures 2A,B**). Thus, there were more enriched than depleted OTUs under straw mulching practice. Furthermore, both abundant and intermediate fungal taxa exhibited more enriched than depleted OTUs in the tillage comparison groups (**Figures 2C,D**).

Principle component analysis (PCA) of the fungal sub-communities among the four treatments showed that the structure of abundant and intermediate taxa communities were obviously distinct (**Supplementary Figure 1**). Our results indicated that conservation tillage significantly affected the abundant and intermediate taxa communities. In respect to the abundant taxa, the two principal components account for 54.4% of the total variance. In contrast, in rare taxa, the two principle components account just 33.2% of the total variance.

[2]https://gephi.org/

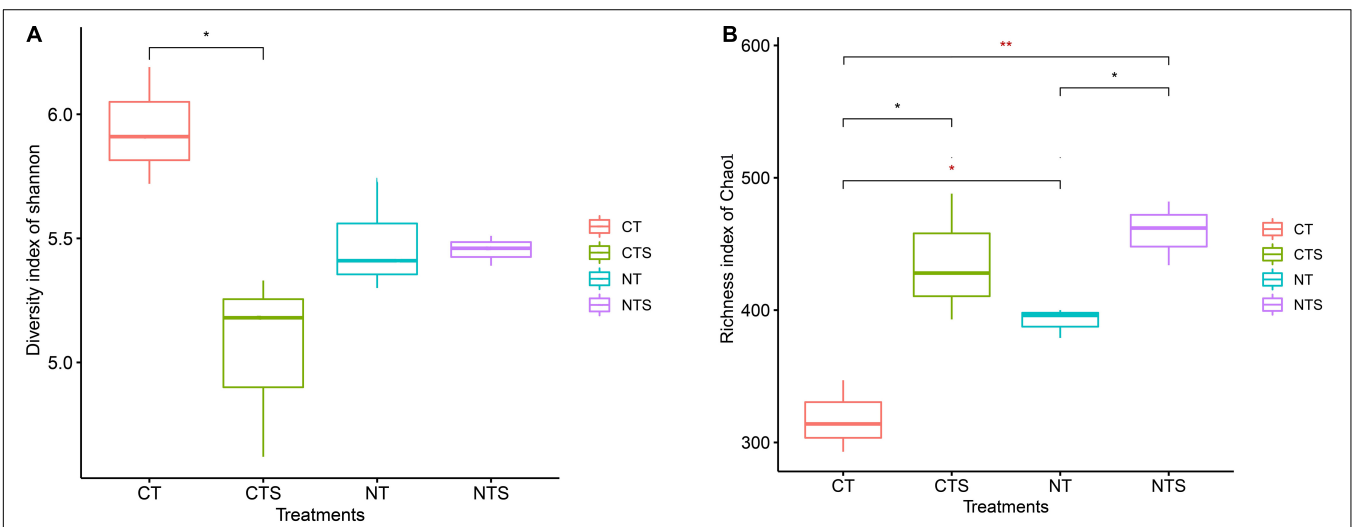

FIGURE 1 | Fungal alpha diversity associated with different soil treatments. The indices of diversity and richness are shown as Shannon **(A)** and Chao1 **(B)**, respectively. Statistically significant differences between different treatments were determined by *T*-test ANOVA ($p < 0.05$). The symbols *, ** are used to show statistical significance at the 0.05, 0.01 level, respectively.

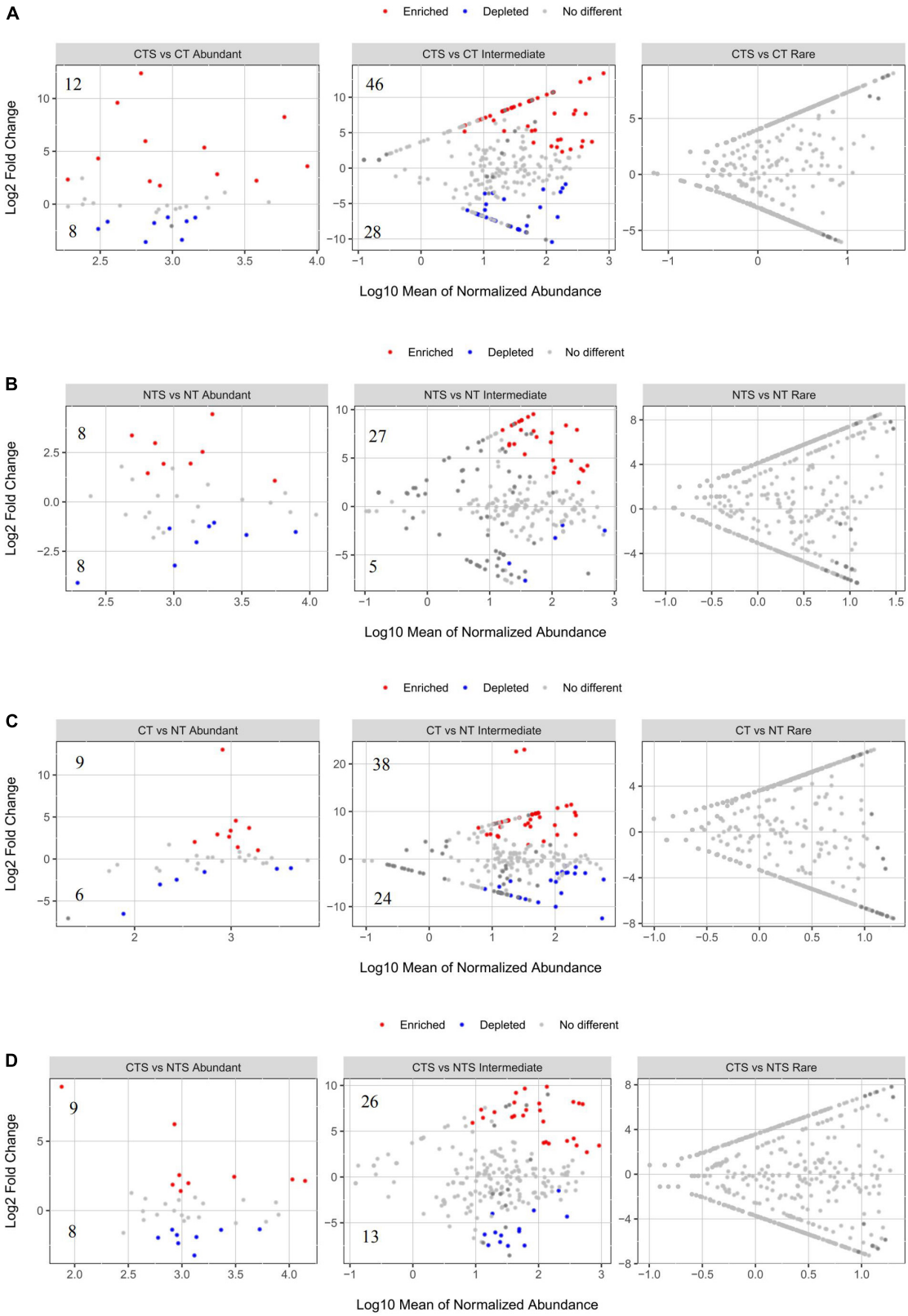

FIGURE 2 | Volcano plots illustrating OTUs significantly enriched (red) and depleted (blue) by straw mulching **(A,B)** and tillage **(C,D)** managements for fungal sub-communities. Each point indicates an individual OTU.

Fungal Communities With Statistically Significant Differences

The Cladogram in **Figure 3** shows the phylogenetic distribution of fungal lineages that were markedly associated (LDA value > 3) with samples from different tillage management fields (**Figures 3A,B**). LEfSe was applied to find statistically different biomarkers among four treatments. In the CTS treatment, three groups were significantly enriched, namely *Eurotiales* (from order to genus), *Cordycipitaceae* (from family to genus) and *Stachybotrys* (genus). In the NTS treatment, the enriched fungi was *Agaricomycetes* (from class to genus). In the CT treatment, the fungal taxa were mostly enriched at the family level, including *Myxotrichaceae*, *Erysiphaceae*, *Chaetomiaceae*, *Incertaesedis*, and *Lasiosphaeriaceae*. In the NT treatment, six groups were found to be significantly enriched, namely *Onygenales* (from order to genus), *Oidiodendron* and *Podospora* (genus), *Microascales* (from order to genus), *Gloeophyllales* (from order to genus), *Chaetomiaceae* (its genus *Humicola* and *Mycothermus*) (**Figure 3A**).

Correlation of Fungal Sub-Communities With Environmental Factors

As illustrated in **Supplementary Table 1**, TOC, TN, AN, AP, and AK were increased in straw mulching treatments (CTS and NTS) compared to no-straw treatments (NT and CT), indicating that straw mulching probably helps to accumulate nutrition in the soil layer. No-tillage (NT and NTS), meanwhile, increased the content of TP compared to tillage (CT and CTS) treatments.

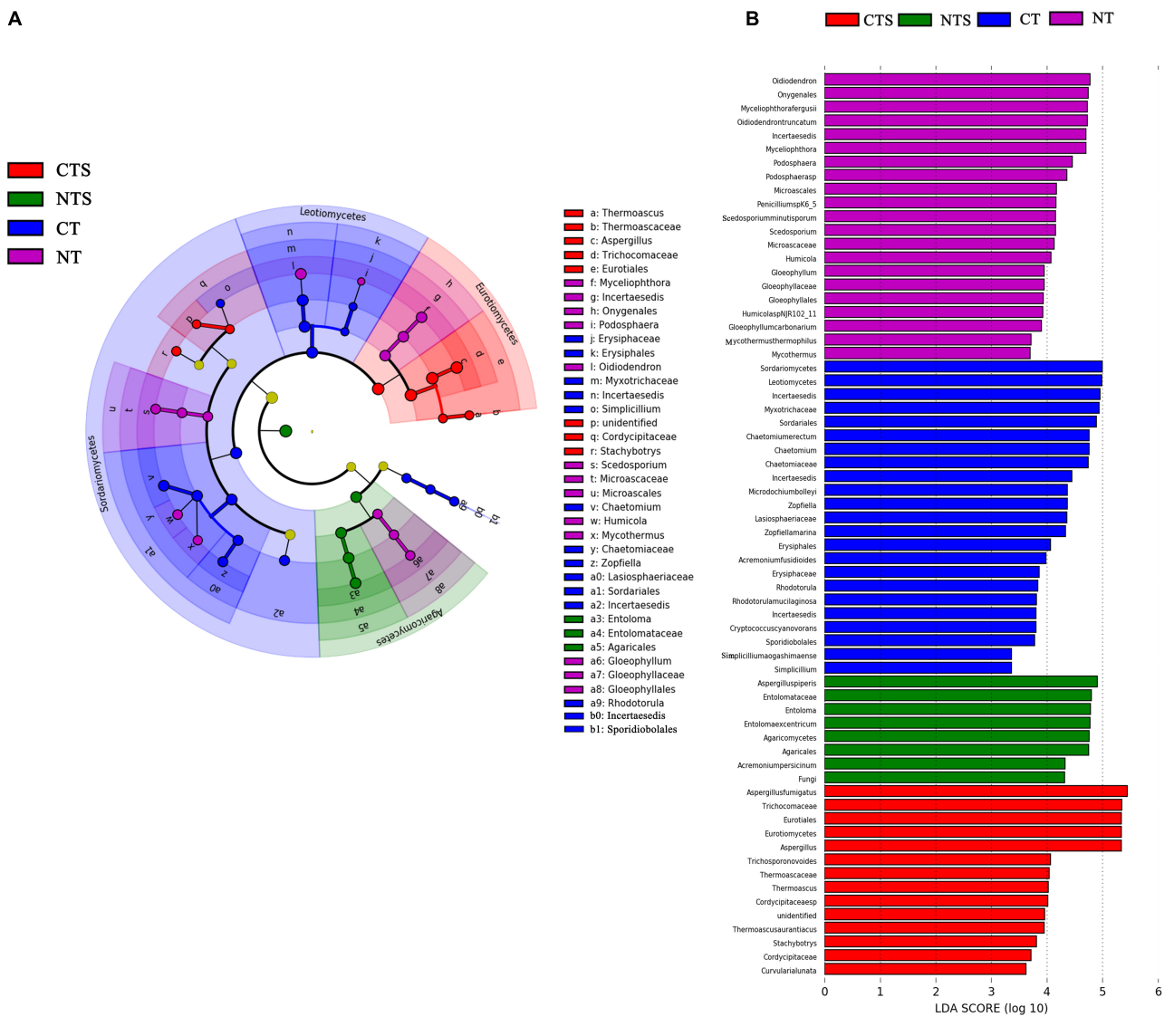

FIGURE 3 | Cladogram showing the phylogenetic distribution of the fungal lineages associated with soil from the four treatments **(A)**. Indicator fungi with LDA scores > 3 **(B)**. Different colors represent different treatments (red, CTS; green, NTS; blue, CT; purple, NT). The circle from inside to outside represents phylogenetic levels from domain to genus, with its diameter reflecting the relative abundance of fungi.

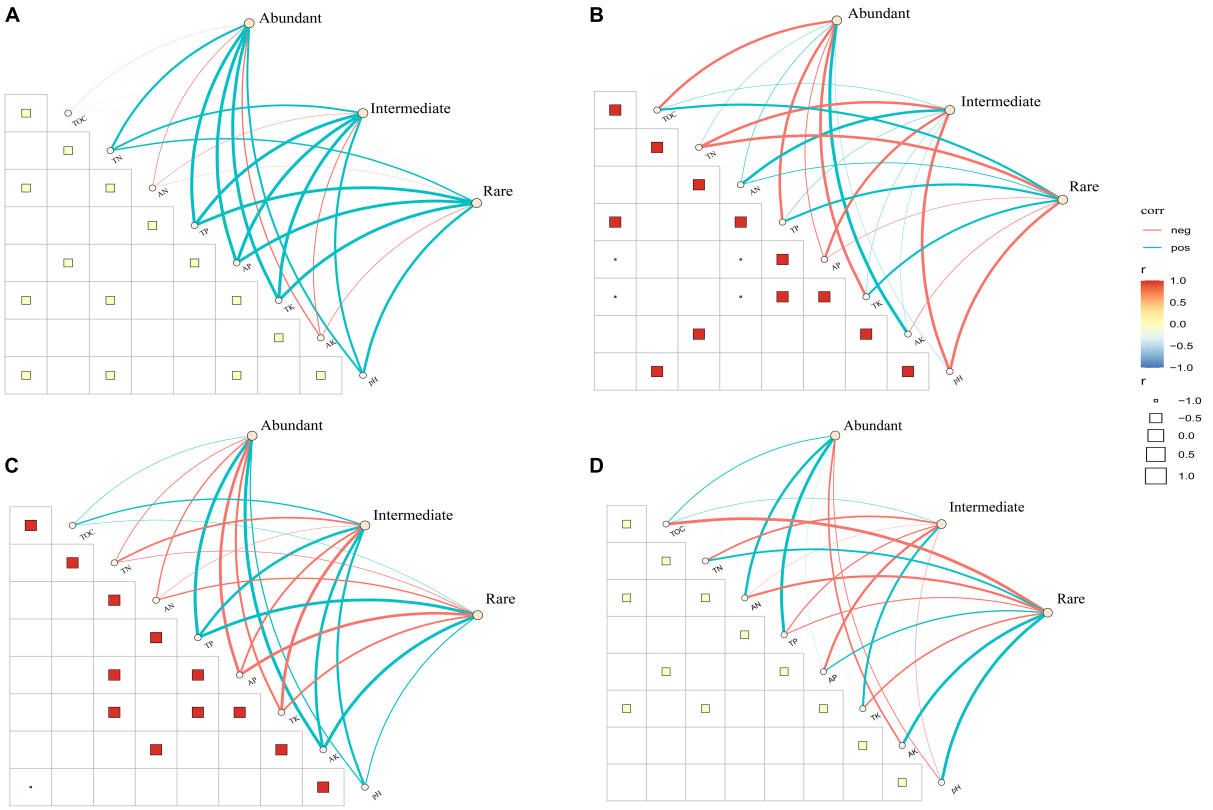

FIGURE 4 | Pairwise comparisons of soil properties and their effects on fungal sub-communities composition in four treatments, with a color gradient denoting Spearman's correlation coefficients. Edges represent Mantel's r for correlations, and the color corresponding to the significance. **(A)** CT. **(B)** CTS. **(C)** NT. **(D)** NTS.

In the CT treatment, most of the soil properties were negatively correlated with fungal sub-communities (**Figure 4A**). The results showed that AP has a positive correlation with abundant, intermediate and rare taxa, TN and pH significantly influence the composition of intermediate and rare taxa in the CTS treatment (**Figure 4B**). In the NT treatment, three fungal sub-communities were simultaneously influenced by multiple factors including TN, AN, AP and TK (**Figure 4C**). As for NTS treatment, environmental factors (TOC, TN, AN, TP, AP and TK) mainly had a positive influence on intermediate and rare taxa (**Figure 4D**). Overall, AP and TN were the strongest correlates of fungal sub-communities in the conservation tillage treatments.

Assembly Processes in Abundant, Intermediate and Rare Fungal Sub-Communities

Briefly, $|\beta NTI| > 2$ and $|\beta NTI| < 2$ represent the deterministic and stochastic processes, respectively. With respect to abundant fungal taxa, deterministic processes comprised more than 62.5% of the assembly processes in CTS and NT treatments, while stochastic processes comprised more than 55.5% of the processes shaping abundant sub-communities in CT and NTS treatments (**Figure 5A**). The distribution of βNTI values across all treatments showed that stochastic processes comprised more than 84.8% of the assembly processes shaping intermediate and rare taxa (**Figures 5B,C**).

We also used quantitative analyses to explain the assembly processes of fungal sub-communities more fully. Dispersal limitation contributed the largest fraction to the assembly of both rare (>84.8%) (**Figure 5C**) and intermediate sub-communities (>93.9%) (**Figure 5B**). Homogeneous selection (> 62%) contributed a large fraction to the assembly of abundant sub-communities in CTS and NT treatments, followed by dispersal limitation. By contrast, in NTS and CT treatments, dispersal limitation (> 55%) contributed a lager fraction to the assembly of abundant sub-communities than homogeneous selection (**Figure 5A**). The undominated process (6.06%), variable selection (< 9%) contributed smaller fractions to the fungal assembly processes (**Figures 5B,C**).

Tillage and Straw Mulching Practices Changed the Fungal Co-occurrence Patterns

We used co-occurrence network analysis to reveal the complexity of fungal sub-community networks (**Figure 6**). The fungal empirical co-occurrence patterns differed significantly with the application of the different tillage practices. As revealed by the network parameters (**Supplementary Table 2**), both abundant

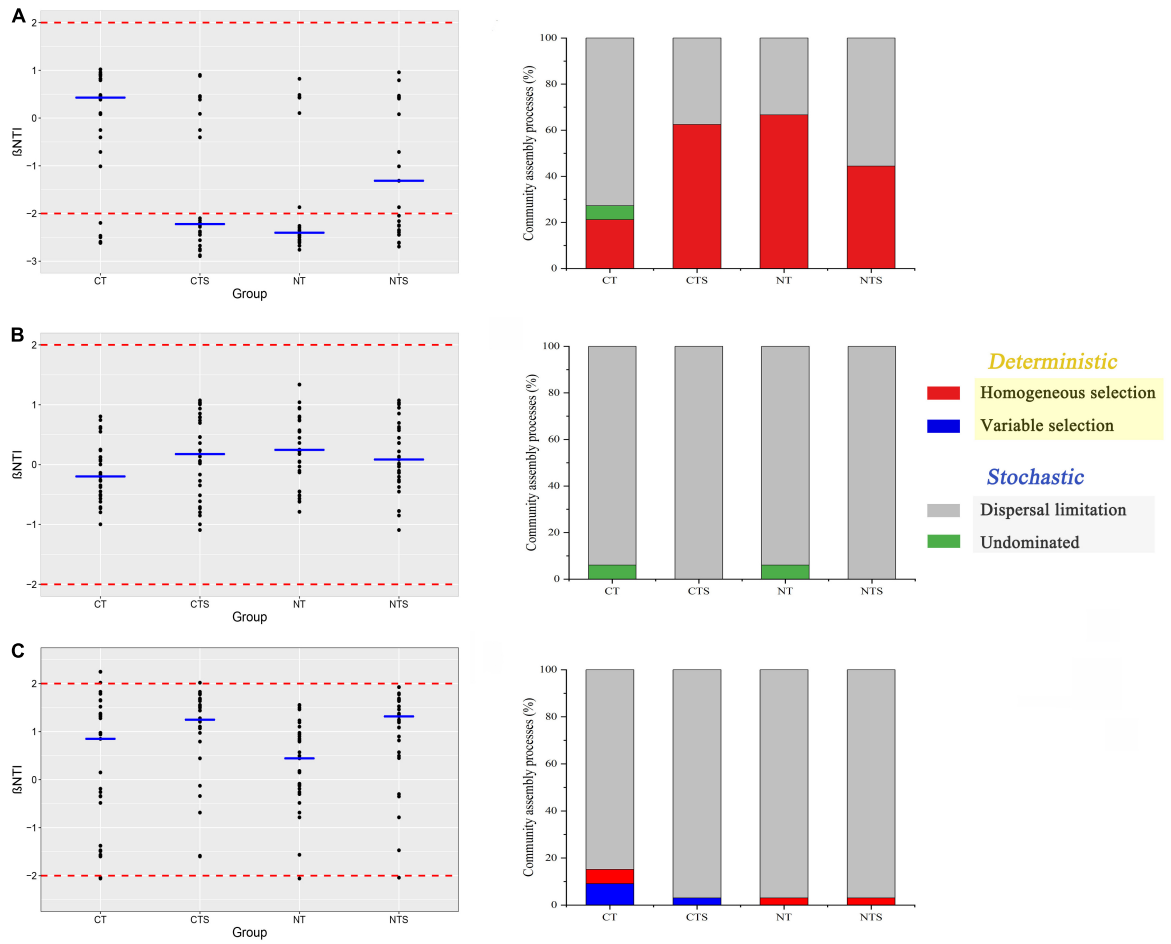

FIGURE 5 | Scatter plot of βNTI values and quantitative analysis of the assembly processes that govern the turnover of abundant **(A)**, intermediate **(B)** and rare **(C)** taxa communities in four treatments. The blue line represents the median values of each treatment.

and intermediate fungal taxa had less complexity in no-tillage treatments than in tillage treatments. The tillage treatments increased the number of nodes and edges compared to no-tillage treatments for abundant and intermediate taxa. The number of edges increased by 2.5-fold (the sum of abundant and intermediate taxa) in tillage treatments, indicating the strong interaction in the abundant and intermediate fungal sub-communities. In respect to rare taxa, however, there were no significant changes in the number of nodes and edges in tillage treatments compared to no-tillage treatments. Straw mulching, meanwhile, reduced the complexity of the abundant and intermediate taxa network, as reflected by the lower number of nodes and edges (**Supplementary Table 2**).

Tillage management was also shown to have different effects on the topological properties of fungal sub-communities. The average clustering coefficient and average degree decreased, but the average path length increased in the no-tillage treatment for abundant and intermediate taxa, while the changes in these topological properties were not significant for the rare taxa. These results showed that the tillage practice notably increased the proportions of positive links in the abundant taxa (61.1% in tillage and 45.3% in no-tillage). The proportions of positive links in abundant taxa were also increased in no-straw mulching (57.7%) compared to straw mulching (43.8%). Our result showed that no-tillage practices had more abundances of potential plant pathogens than tillage practices. Straw mulching had the highest abundance of potential plant pathogens among four practices, and these pathogens mainly have positive correlation with other fungi in the networks (**Supplementary Table 3**).

DISCUSSION

Effects of Conservation Tillage on Fungal Alpha Diversity

Our finding indicated that conservation tillage practices significantly increased fungal richness, but no significant effect on fungal diversity, which is consistent with previous research (Wang et al., 2020a). In this study, we found that conservation tillage (CTS, NTS, and NT) has higher organic carbon content than CT (**Supplementary Table 1**), previous study indicated that fungal richness was significantly correlated to soil organic carbon

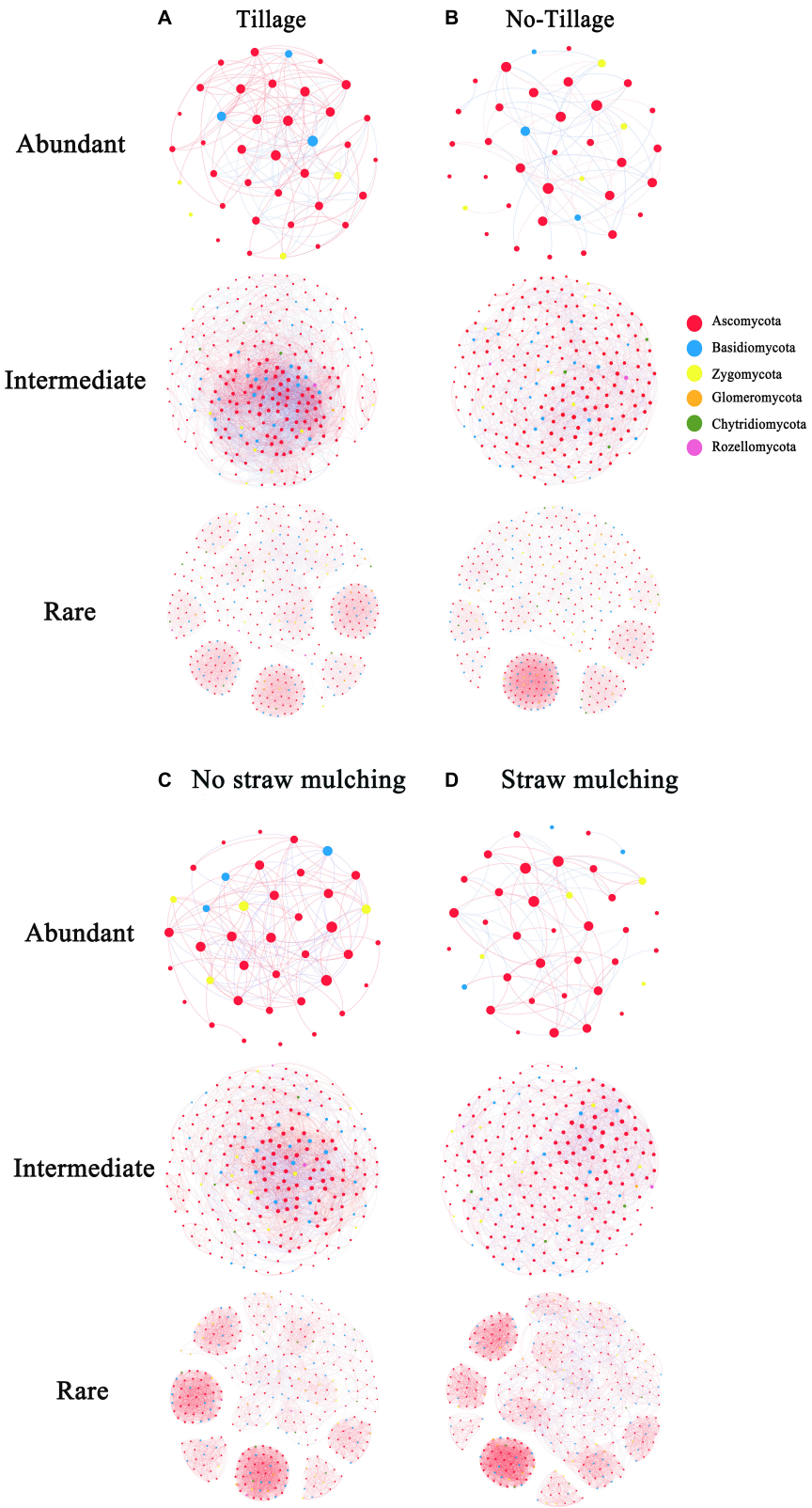

FIGURE 6 | Network co-occurrence analysis (Spearman's $\rho > 0.6$ and significant $p < 0.05$) of fungal sub-communities in the tillage (CT and CTS) and no-tillage (NT and NTS) treatments, no-straw mulching (CT and NT) and straw mulching (CTS and NTS). The size of each node is proportional to the relative abundance; red lines and blue lines represent positive and negative correlations, respectively; nodes of the same color belongs to the same phylum. **(A)** Tillage (CT and CTS). **(B)** No-tillage (NT and NTS). **(C)** No straw mulching (CT and NT). **(D)** Straw mulching (CTS and NTS).

(SOC) content, thereby contributing to soil fungal richness (Yang et al., 2012). Additionally, a previous study demonstrated that conservation tillage practices have no favorable effect on fungal community diversity (Degrune et al., 2016). A study noted that conservation tillage management can lead to greater fungal diversity by changing soil microenvironment (Wang et al., 2017). This disparate understanding of the effect of conservation tillage on the soil fungal community may be due to the complexity of the environmental conditions. For example, soil with high clay and sand fractions can lead to modification in fungal community (Bach et al., 2010). Soil management histories (forest to cultivated land vs. longstanding cultivation) are also key factors (Degrune et al., 2016).

Straw Mulching and Conventional Tillage Affected the Fungal Community Structure

We picked out enriched and depleted OTUs to analyze the differences in the fungal sub-communities using differential abundance analysis. We observed that abundant and intermediate taxa had enriched and depleted OTUs except rare taxa. Compared to CT treatment, CTS increased the proportion of enriched OTUs more than depleted OTUs in abundant and intermediate taxa. NTS, meanwhile, increased the proportion of enriched OTUs more than depleted OTUs. These results indicate that straw mulching helps fungi to grow. Organic farming practice has positive effect on fungi biomass, probably because carbon content is the key factor that governing microbial growth (Birkhofer et al., 2008; Wang et al., 2017). Although NT can provide more favorable conditions (a cooler and moist environment) than CT (Helgason et al., 2009), our results showed that CT enriched more OTUs than NT. Soil disturbance can improve the distribution of plant residues and substrate availability, distributing soil aggregates and releasing particulate organic matter, and thus supporting the growth of micro-biota (Chaer et al., 2009).

Conservation tillage can affect residue decomposition and alter gas and water movement, leading to changes in fungal community patterns (Holland, 2004; Wang et al., 2016, 2017). The PCA results showed that conservation practices significantly influence on abundant and intermediate taxa, but do not have a significant influence on rare taxa community structure. In our study, the total variance explained by PCA was much higher for the abundant sub-community (54.4%) than for the rare sub-community (33.2%). Rare microbes have a large proportion of unexplained variation because rare taxa are more subject to biotic interaction (e.g., competition) and have discrepant ecological niches (Liu et al., 2015; Lopes and Fernandes, 2020).

We used LEfSe analysis to understand the variation in fungal communities in long-term conservation fields more fully. This method can analyze the microbial community at any clade. We retained the taxa with significant differences and filtered out those without significant differences. Statistical analysis was performed from phylum to genus.

According to the LEfSe results, *Eurotiales* were enriched in CTS treatment. A recent study has demonstrated that organic matter can promote the growth of fungal taxa, and *Eurotiales* is important for the SOC decomposition process (Wang et al., 2021). One of the other fungi found to be enriched in the plots subject to NTS treatment was *Agaricomycetes*, which can degrade various substrates, such as cellulose and lignin (Zhang et al., 2021). Decomposition of crop residues on the soil surface could therefore be enhanced by this fungal growth. In addition, a recent study has shown that *Scedosporium*, which is considered to have pathogenic potentials, is enriched in NT treatment (Kitisin et al., 2021). On the other hand, Wang et al. (2020a) found that NT may increase the risk of stubble-borne diseases.

Environmental Drivers of Fungal Sub-Communities Under Conservation Practice

No-tillage and straw mulching have been found to have a significant effect on soil nutrient parameters (Wang et al., 2020a). Our results showed that NT significantly increased the soil TP content, probably on the basis that it increases the residual P concentration in the soil surface layer (Jansa et al., 2003). Furthermore, crop residual acts as a carbon source as well as increasing the organic matter content of soil (Bu et al., 2020). We also observed that TOC, TN, and AN contents increased under straw mulching treatments, which is in line with a previous study reporting that the use of cover crops accounted for most of the N increase associated with crop rotation effects (McDaniel et al., 2014). Overall, conservation tillage treatment improves nutrition conditions for soil microbial communities.

Previous studies have revealed that pH and nutrient levels are key predictors of fungal composition (Lauber et al., 2008; Rousk et al., 2010). Our findings support this by showing that AP, TN and AN simultaneously influenced the fungal sub-communities in NT treatment. In the conservation tillage treatments (CTS, NT and NTS), soil properties were closely related to fungal sub-communities. In contrast, there was a mainly negative relationship between fungal sub-communities and soil properties in CT treatment.

Different Assembly Processes Experienced by Abundant, Intermediate and Rare Fungal Sub-Communities

Uncovering the underlying microbial assembly processes is a key subject for microbial ecology (Nemergut et al., 2013). It is generally recognized that spatial heterogeneity and environmental filtering contribute to the microbial assembly and community structure (Xue et al., 2021). Our results showed that, in CTS and NT treatments, the assembly of abundant fungal taxa was governed by deterministic processes, whereas in CT treatment sub-community assembly was governed by stochastic processes. A significant correlation between environment factors and fungal sub-communities was found in NT and CTS. Specifically, it was revealed that no-tillage and straw mulching positively influenced the soil properties and this shaped the abundant sub-community.

The assembly of abundant taxa in plots subject to NTS treatment, however, was governed by stochastic processes, and

the abundant taxa diversity indices (Shannon) were higher in NTS plots than those subject to CTS or NT treatments (**Supplementary Table 4**). NTS can therefore serve to increase the diversity of the fungal community, probably by increasing the total C and bioavailable C (Navarro-Noya et al., 2013). High-diversity communities are dominated by stochastic processes, while low diversity ones depended on deterministic processes which constrained the community function (Xun et al., 2019), which in line with our findings. Similarly, intermediate and rare fungal taxa were dominated by stochastic processes, while abundant taxa exhibited a remarkably wide response to the ecological preferences in agriculture fields (Jiao and Lu, 2020a).

Previous studies have indicated that the assembly processes of different sub-communities rely on different environmental factors in the agro-ecosystem (Jiao et al., 2017; Jiao and Lu, 2020b). Homogeneous selection significantly affected the abundant sub-communities in NT and CTS treatments, whereas rare and intermediate sub-communities were more subject to dispersal limitations. This contrasts with previous research appearing to show that rare sub-communities are governed mainly by homogeneous selection (Jiao and Lu, 2020a). These discrepancies may be caused by straw retention and no-tillage practices and by geography (Shi et al., 2018). The dominant status of homogeneous selection in CTS and NT plots suggests that abundant taxa are more sensitive to conservation practice, while the fact that dispersal limitation dominated the rare and intermediate taxa implied a weak link with no-tillage and straw retention practices.

To understand the assembly processes of fungal sub-communities more fully, we established a conceptual model (**Figure 7**). This presents that ecological processes can emerge in the following forms: (i) under weak strength selection (conventional tillage conditions), the establishment of fungal sub-communities is dominated by stochastic processes, (ii) under medium strength selection (straw mulching or no-tillage practices), changes in environment conditions enhance the selection leading to deterministic processes dominating in the abundant sub-community, (iii) strong strength selection (straw mulching combined with no-tillage) increase the diversity of the abundant fungal sub-community and thus induced stochastic processes. Notably, intermediate and rare sub-communities are consistently dominated by stochasticity. These microbial taxa are characterized by high levels of organismal dispersal, and influenced by stochastic birth or death rather than environmental filters (Dini-Andreote et al., 2015).

No-Tillage and Straw Mulching Practices Changed the Co-occurrence Patterns of Abundant and Intermediate Taxa

In the present study, we observed that no-tillage and straw mulching reduced the complexity of the abundant and intermediate fungal taxa network, while conservation tillage practices had no significant influence on the rare taxa network. Previous studies showed that agricultural intensification reduced the complexity of the microbial network, and tillage practice was considered to be harmful to the extension of fungal mycelia (Caesar-TonThat et al., 2010; Banerjee et al., 2019). Our results, however, showed that conventional tillage increased the fungal interaction and the proportion of positive link in abundant taxa compared to conservation tillage treatments.

Our results might be explained by recognizing that microbial ecology is affected by nutrient availability, aeration, moisture and pH (Fierer and Jackson, 2006). It may be that the

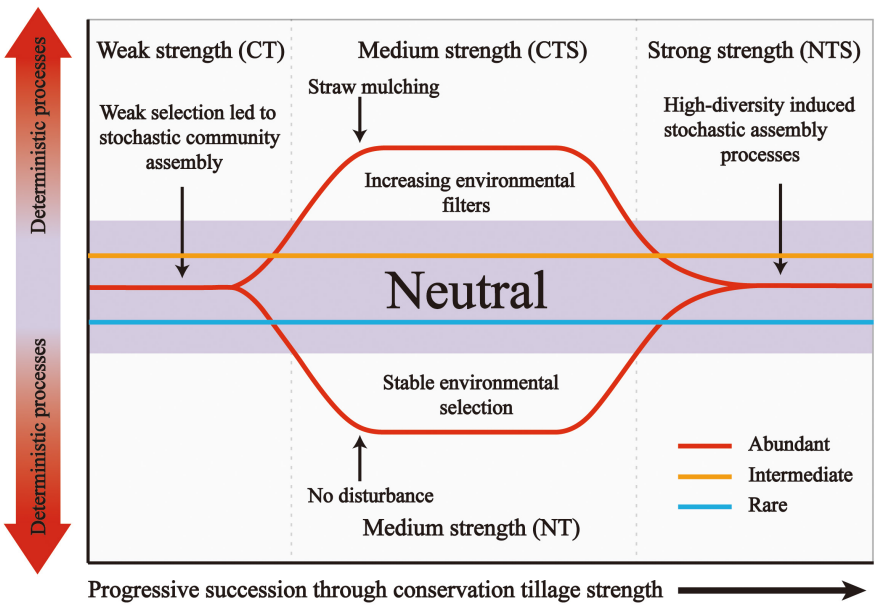

FIGURE 7 | Conceptual model showing how ecological selection dominates the structure of fungal sub-communities through conservation tillage strength. Weak conservation tillage strength, CT; Medium conservation tillage strength, CTS and NT; Strong conservation tillage strength, NTS.

full soil inversion created by tillage practice can promote soil nutrients used by fungi and strengthen the links in the microbial network (Six et al., 2000; Le Guillou et al., 2012). Furthermore, another recent study has found that the abundance of arbuscular mycorrhizal fungi (AMF) was higher in conventional than in conservation tillage, which the authors explained by the dilution of P in the surface soil layer in tillage practice (Lopes and Fernandes, 2020). Wang et al. (2020a) found that straw mulching decreased both fungal and bacterial network complexity, possibly due to straw mulching created favorable nutrient conditions for fungi, decreasing microbial inhibition and competition, and thus weakening interaction and negative relevance (Bronstein, 1994; Cao et al., 2018). These results indicated that conventional tillage practice can deliver a more stable fungal network in corn-wheat rotation systems. In our study, compared to tillage practices, NT practices had more abundances of potential plant pathogens. Straw mulching practices had more abundances of plant pathogens than no-straw mulching practices. Conservation tillage can enhance the growth of plant pathogens by concentrating plant debris, and tillage practices might alleviate the ecological risks of the pathogens (Sturz et al., 1997; Hu et al., 2021). We observed that potential plant pathogens mainly have positive correlation with other fungi in the networks, might due to the cooperation of fungi in the decomposition of straw residues (Hu et al., 2021).

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Genomic sequencing data has been deposited in the NCBI Sequence Read Archive (BioProject ID PRJNA764374, submission ID SUB10399701).

AUTHOR CONTRIBUTIONS

HC and AZ revised the manuscript and developed the experiment idea. CZ and HL conducted most of the experiments, prepared the manuscript, and analyzed the data. SL analyzed the data and revised the manuscript. SH contributed to supervision. LZ, XY, KC, and XX participated in revision of the manuscript. All authors read and agreed the final manuscript.

FUNDING

This work has been jointly supported by the National Natural Science Foundation of China (Grants 41877023 and 42077026).

ACKNOWLEDGMENTS

We would like to thank the Fengqiu National Agro-ecological Experiment Station for assistance in conduct of long-term conservation field trials.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.829152/full#supplementary-material

REFERENCES

Bach, E. M., Baer, S. G., Meyer, C. K., and Six, J. (2010). Soil texture affects soil microbial and structural recovery during grassland restoration. *Soil Biol. Biochem.* 42, 2182–2191. doi: 10.1016/j.soilbio.2010.08.014

Balesdent, J., Chenu, C., and Balabane, M. (2000). Relationship of soil organic matter dynamics to physical protection and tillage. *Soil Tillage Res.* 53, 215–230. doi: 10.1016/S0167-1987(99)00107-5

Banerjee, S., Walder, F., Büchi, L., Meyer, M., Held, A. Y., Gattinger, A., et al. (2019). Agricultural intensification reduces microbial network complexity and the abundance of keystone taxa in roots. *ISME J.* 13, 1722–1736. doi: 10.1038/s41396-019-0383-2

Bao, S. D. (2000). *Soil and Agricultural Chemistry Analysis*. Beijing: China Agriculture Press.

Beare, M., Hu, S., Coleman, D., and Hendrix, P. (1997). Influences of mycelial fungi on soil aggregation and organic matter storage in conventional and no-tillage soils. *Appl. Soil Ecol.* 5, 211–219. doi: 10.1016/S0929-1393(96)00142-4

Birkhofer, K., Bezemer, T. M., Bloem, J., Bonkowski, M., Christensen, S., Dubois, D., et al. (2008). Long-term organic farming fosters below and aboveground biota: implications for soil quality, biological control and productivity. *Soil Biol. Biochem.* 40, 2297–2308. doi: 10.1016/j.soilbio.2008.05.007

Bronstein, J. L. (1994). Conditional outcomes in mutualistic interactions. *Trends Ecol. Evol.* 9, 214–217. doi: 10.1016/0169-5347(94)90246-1

Bu, R., Ren, T., Lei, M., Liu, B., Li, X., Cong, R., et al. (2020). Tillage and straw-returning practices effect on soil dissolved organic matter, aggregate fraction and bacteria community under rice-rice-rapeseed rotation system. *Agric. Ecosyst. Environ.* 287:106681. doi: 10.1016/j.agee.2019.106681

Caesar-TonThat, T., Lenssen, A. W., Caesar, A. J., Sainju, U. M., and Gaskin, J. F. (2010). Effects of tillage on microbial populations associated to soil aggregation in dryland spring wheat system. *Eur. J. Soil Biol.* 46, 119–127. doi: 10.1016/j.ejsobi.2009.12.004

Cao, X., Zhao, D., Xu, H., Huang, R., Zeng, J., and Yu, Z. (2018). Heterogeneity of interactions of microbial communities in regions of Taihu Lake with different nutrient loadings: a network analysis. *Sci. Rep.* 8:8890. doi: 10.1038/s41598-018-27172-z

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* 6, 1621–1624. doi: 10.1038/ismej.2012.8

Chaer, G. M., Fernandes, M. F., Myroldi, D., and Bottomley, P. J. (2009). Shifts in microbial community composition and physiological profiles across a gradient of induced soil degradation. *Embrapa Agrobiologia-Artigo em periódico indexado (ALICE)* 73, 1327–1334.

Chave, J. (2004). Neutral theory and community ecology. *Ecol. Lett.* 7, 241–253. doi: 10.1111/j.1461-0248.2003.00566.x

Chesson, P. (2000). Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* 31, 343–366. doi: 10.1146/annurev.ecolsys.31.1.343

Degrune, F., Theodorakopoulos, N., Dufrêne, M., Colinet, G., Bodson, B., Hiel, M.-P., et al. (2016). No favorable effect of reduced tillage on microbial community diversity in a silty loam soil (Belgium). *Agric. Ecosyst. Environ.* 224, 12–21. doi: 10.1016/j.agee.2016.03.017

Dini-Andreote, F., Stegen, J. C., Van Elsas, J. D., and Salles, J. F. (2015). Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial succession. *Proc. Natl. Acad. Sci. U.S.A.* 112, E1326–E1332. doi: 10.1073/pnas.1414261112

Diniz-Filho, J. A. F., Soares, T. N., Lima, J. S., Dobrovolski, R., Landeiro, V. L., Telles, M. P. D. C., et al. (2013). Mantel test in population genetics. *Genet. Mol. Biol.* 36, 475–485. doi: 10.1590/S1415-47572013000400002

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381

Fierer, N., and Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proc. Natl. Acad. Sci. U.S.A.* 103, 626–631. doi: 10.1073/pnas.0507535103

Guo, L.-J., Zhang, Z.-S., Wang, D.-D., Li, C.-F., and Cao, C.-G. (2015). Effects of short-term conservation management practices on soil organic carbon fractions and microbial community composition under a rice-wheat rotation system. *Biol. Fertil. Soils* 51, 65–75. doi: 10.1007/s00374-014-0951-6

Hartmann, M., Frey, B., Mayer, J., Mäder, P., and Widmer, F. (2015). Distinct soil microbial diversity under long-term organic and conventional farming. *ISME J.* 9, 1177–1194. doi: 10.1038/ismej.2014.210

Helgason, B. L., Walley, F. L., and Germida, J. J. (2009). Fungal and bacterial abundance in long-term no-till and intensive-till soils of the Northern Great Plains. *Soil. Sci. Soc. Am. J.* 73, 120–127. doi: 10.2136/sssaj2007.0392

Hobbs, P. R., Sayre, K., and Gupta, R. (2008). The role of conservation agriculture in sustainable agriculture. *Philos. Trans. R. Soc. B: Biol. Sci.* 363, 543–555. doi: 10.1098/rstb.2007.2169

Holland, J. M. (2004). The environmental consequences of adopting conservation tillage in Europe: reviewing the evidence. *Agric. Ecosyst. Environ.* 103, 1–25. doi: 10.1016/j.agee.2003.12.018

Hu, X., Liu, J., Liang, A., Li, L., Yao, Q., Yu, Z., et al. (2021). Conventional and conservation tillage practices affect soil microbial co-occurrence patterns and are associated with crop yields. *Agric. Ecosyst. Environ.* 319:107534. doi: 10.1016/j.agee.2021.107534

Jansa, J., Mozafar, A., Kuhn, G., Anken, T., Ruh, R., Sanders, I., et al. (2003). Soil tillage affects the community structure of mycorrhizal fungi in maize roots. *Ecol. Appl.* 13, 1164–1176. doi: 10.1890/1051-0761(2003)13[1164:statcs]2.0.co;2

Jiang, Y., Song, H., Lei, Y., Korpelainen, H., and Li, C. (2019). Distinct co-occurrence patterns and driving forces of rare and abundant bacterial subcommunities following a glacial retreat in the eastern Tibetan Plateau. *Biol. Fertil. Soils* 55, 351–364. doi: 10.1007/s00374-019-01355-w

Jiao, S., Chen, W., and Wei, G. (2017). Biogeography and ecological diversity patterns of rare and abundant bacteria in oil-contaminated soils. *Mol. Ecol.* 26, 5305–5317. doi: 10.1111/mec.14218

Jiao, S., and Lu, Y. (2020a). Abundant fungi adapt to broader environmental gradients than rare fungi in agricultural fields. *Glob. Change Biol.* 26, 4506–4520. doi: 10.1111/gcb.15130

Jiao, S., and Lu, Y. (2020b). Soil pH and temperature regulate assembly processes of abundant and rare bacterial communities in agricultural ecosystems. *Environ. Microbiol.* 22, 1052–1065. doi: 10.1111/1462-2920.14815

Jiao, S., Zhang, B., Zhang, G., Chen, W., and Wei, G. (2021). Stochastic community assembly decreases soil fungal richness in arid ecosystems. *Mol. Ecol.* 30, 4338–4348. doi: 10.1111/mec.16047

Joergensen, R. G., and Wichern, F. (2018). Alive and kicking: why dormant soil microorganisms matter. *Soil Biol. Biochem.* 116, 419–430. doi: 10.1016/j.soilbio.2017.10.022

Kassam, A., Friedrich, T., Shaxson, F., Bartz, H., Mello, I., Kienzle, J., et al. (2014). The spread of conservation agriculture: policy and institutional support for adoption and uptake. *Field Actions Sci. Rep. J. Field Actions* 7. Available online at: http://factsreports.revues.org/3720

Kitisin, T., Muangkaew, W., Ampawong, S., Chutoam, P., Thanomsridetchai, N., Tangwattanachuleeporn, M., et al. (2021). Isolation of fungal communities and identification of Scedosporium species complex with pathogenic potentials from a pigsty in Phra Nakhon Si Ayutthaya, Thailand. *New Microbiol.* 44, 33–41.

Lauber, C. L., Strickland, M. S., Bradford, M. A., and Fierer, N. (2008). The influence of soil properties on the structure of bacterial and fungal communities across land-use types. *Soil Biol. Biochem.* 40, 2407–2415.

Le Guillou, C., Angers, D., Leterme, P., and Menasseri-Aubry, S. (2012). Changes during winter in water-stable aggregation due to crop residue quality. *Soil Manag.* 28, 590–595.

Levy-Booth, D. J., Prescott, C. E., and Grayston, S. J. (2014). Microbial functional genes involved in nitrogen fixation, nitrification and denitrification in forest ecosystems. *Soil Biol. Biochem.* 75, 11–25. doi: 10.1016/j.soilbio.2014.03.021

Li, Y., Li, T., Zhao, D., Wang, Z., and Liao, Y. (2021). Different tillage practices change assembly, composition, and co-occurrence patterns of wheat rhizosphere diazotrophs. *Sci. Total Environ.* 767:144252. doi: 10.1016/j.scitotenv.2020.144252

Liu, L., Yang, J., Yu, Z., and Wilkinson, D. M. (2015). The biogeography of abundant and rare bacterioplankton in the lakes and reservoirs of China. *ISME J.* 9, 2068–2077. doi: 10.1038/ismej.2015.29

Lopes, L. D., and Fernandes, M. F. (2020). Changes in microbial community structure and physiological profile in a kaolinitic tropical soil under different conservation agricultural practices. *Appl. Soil Ecol.* 152:103545. doi: 10.1016/j.apsoil.2020.103545

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 1–21. doi: 10.1186/s13059-014-0550-8

McDaniel, M., Tiemann, L., and Grandy, A. (2014). Does agricultural crop diversity enhance soil microbial biomass and organic matter dynamics? A meta-analysis. *Ecol. Appl.* 24, 560–570. doi: 10.1890/13-0616.1

Mueller, R. C., Paula, F. S., Mirza, B. S., Rodrigues, J. L., Nüsslein, K., and Bohannan, B. J. (2014). Links between plant and fungal communities across a deforestation chronosequence in the Amazon rainforest. *ISME J.* 8, 1548–1550. doi: 10.1038/ismej.2013.253

Navarro-Noya, Y. E., Gómez-Acata, S., Montoya-Ciriaco, N., Rojas-Valdez, A., Suárez-Arriaga, M. C., Valenzuela-Encinas, C., et al. (2013). Relative impacts of tillage, residue management and crop-rotation on soil bacterial communities in a semi-arid agroecosystem. *Soil Biol. Biochem.* 65, 86–95. doi: 10.1016/j.soilbio.2013.05.009

Nemergut, D., Schmidt, S., Fukami, T., O'Neill, S., Legg, T., Stanish, L., et al. (2013). Microbial community assembly: patterns and processes. *Microbiol. Mol. Biol. Rev* 77:e356. doi: 10.1128/MMBR.00051-12

Nguyen, N. H., Song, Z., Bates, S. T., Branco, S., Tedersoo, L., Menke, J., et al. (2016). FUNGuild: an open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecol.* 20, 241–248. doi: 10.1016/j.funeco.2015.06.006

Ning, D., Yuan, M., Wu, L., Zhang, Y., Guo, X., Zhou, X., et al. (2020). A quantitative framework reveals ecological drivers of grassland microbial community assembly in response to warming. *Nat. Commun.* 11, 1–12. doi: 10.1038/s41467-020-18560-z

Rousk, J., Bååth, E., Brookes, P. C., Lauber, C. L., Lozupone, C., Caporaso, J. G., et al. (2010). Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J.* 4, 1340–1351. doi: 10.1038/ismej.2010.58

Schmidt, R., Mitchell, J., and Scow, K. (2019). Cover cropping and no-till increase diversity and symbiotroph: saprotroph ratios of soil fungal communities. *Soil Biol. Biochem.* 129, 99–109. doi: 10.1016/j.soilbio.2018.11.010

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12, 1–18. doi: 10.1186/gb-2011-12-6-r60

Shi, Y., Li, Y., Xiang, X., Sun, R., Yang, T., He, D., et al. (2018). Spatial scale affects the relative role of stochasticity versus determinism in soil bacterial communities in wheat fields across the North China Plain. *Microbiome* 6, 1–12. doi: 10.1186/s40168-018-0409-4

Six, J., Paustian, K., Elliott, E. T., and Combrink, C. (2000). Soil structure and organic matter I. Distribution of aggregate-size classes and aggregate-associated carbon. *Soil Sci. Soc. Am. J.* 64, 681–689.

Stegen, J. C., Lin, X., Fredrickson, J. K., Chen, X., Kennedy, D. W., Murray, C. J., et al. (2013). Quantifying community assembly processes and identifying features that impose them. *ISME J.* 7, 2069–2079. doi: 10.1038/ismej.2013.93

Stegen, J. C., Lin, X., Fredrickson, J. K., and Konopka, A. E. (2015). Estimating and mapping ecological processes influencing microbial community assembly. *Front. Microbiol.* 6:370. doi: 10.3389/fmicb.2015.00370

Stegen, J. C., Lin, X., Konopka, A. E., and Fredrickson, J. K. (2012). Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME J.* 6, 1653–1664. doi: 10.1038/ismej.2012.22

Sturz, A., Carter, M., and Johnston, H. (1997). A review of plant disease, pathogen interactions and microbial antagonism under conservation tillage in temperate humid agriculture. *Soil Tillage Res.* 41, 169–189.

Tripathi, B. M., Stegen, J. C., Kim, M., Dong, K., Adams, J. M., and Lee, Y. K. (2018). Soil pH mediates the balance between stochastic and deterministic assembly of bacteria. *ISME J.* 12, 1072–1083. doi: 10.1038/s41396-018-0082-4

Vellend, M. (2010). Conceptual synthesis in community ecology. *Q. Rev. Biol.* 85, 183–206. doi: 10.1086/652373

Wang, H., Guo, Q., Li, X., Li, X., Yu, Z., Li, X., et al. (2020a). Effects of long-term no-tillage with different straw mulching frequencies on soil microbial community and the abundances of two soil-borne pathogens. *Appl. Soil Ecol.* 148: 103488.

Wang, Z., Li, Y., Li, T., Zhao, D., and Liao, Y. (2020b). Conservation tillage decreases selection pressure on community assembly in the rhizosphere of arbuscular mycorrhizal fungi. *Sci. Total Environ.* 710:136326. doi: 10.1016/j.scitotenv.2019.136326

Wang, Z., Li, Y., Li, T., Zhao, D., and Liao, Y. (2020c). Tillage practices with different soil disturbance shape the rhizosphere bacterial community throughout crop growth. *Soil Tillage Res.* 197: 104501.

Wang, L., Yuan, X., Liu, C., Li, Z., Chen, F., Li, S., et al. (2019). Soil C and N dynamics and hydrological processes in a maize-wheat rotation field subjected to different tillage and straw management practices. *Agric. Ecosyst. Environ.* 285:106616. doi: 10.1016/j.agee.2019.106616

Wang, X., Bian, Q., Jiang, Y., Zhu, L., Chen, Y., Liang, Y., et al. (2021). Organic amendments drive shifts in microbial community structure and keystone taxa which increase C mineralization across aggregate size classes. *Soil Biol. Biochem.* 153:108062. doi: 10.1016/j.soilbio.2020.108062

Wang, Y., Li, C., Tu, C., Hoyt, G. D., DeForest, J. L., and Hu, S. (2017). Long-term no-tillage and organic input management enhanced the diversity and stability of soil microbial community. *Sci. Total Environ.* 609, 341–347. doi: 10.1016/j.scitotenv.2017.07.053

Wang, Z., Chen, Q., Liu, L., Wen, X., and Liao, Y. (2016). Responses of soil fungi to 5-year conservation tillage treatments in the drylands of northern China. *Appl. Soil Ecol.* 101, 132–140.

Xue, R., Zhao, K., Yu, X., Stirling, E., Liu, S., Ye, S., et al. (2021). Deciphering sample size effect on microbial biogeographic patterns and community assembly processes at centimeter scale. *Soil Biol. Biochem.* 156:108218. doi: 10.1016/j.soilbio.2021.108218

Xun, W., Li, W., Xiong, W., Ren, Y., Liu, Y., Miao, Y., et al. (2019). Diversity-triggered deterministic bacterial assembly constrains community functions. *Nat. Commun.* 10, 1–10. doi: 10.1038/s41467-019-11787-5

Yang, A., Hu, J., Lin, X., Zhu, A., Wang, J., Dai, J., et al. (2012). Arbuscular mycorrhizal fungal community structure and diversity in response to 3-year conservation tillage management in a sandy loam soil in North China. *J. Soils Sediments* 12, 835–843. doi: 10.1007/s11368-012-0518-9

Zhang, M., Zhao, G., Li, Y., Wang, Q., Dang, P., Qin, X., et al. (2021). Straw incorporation with ridge–furrow plastic film mulch alters soil fungal community and increases maize yield in a semiarid region of China. *Appl. Soil Ecol.* 167:104038. doi: 10.1016/j.apsoil.2021.104038

Zhou, Z., Li, Z., Chen, K., Chen, Z., Zeng, X., Yu, H., et al. (2021). Changes in soil physicochemical properties and bacterial communities at different soil depths after long-term straw mulching under a no-till system. *Soil* 7, 595–609. doi: 10.5194/soil-7-595-2021

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

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

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 06 April 2022
doi: 10.3389/fmicb.2022.880118

Esterases From *Bifidobacteria* Exhibit the Conversion of Albiflorin in Gut Microbiota

*Ran Peng[1], Pei Han[1], Jie Fu[1], Zheng-Wei Zhang[1], Shu-Rong Ma[1], Li-Bin Pan[1], Yuan-Yuan Xia[2], Hang Yu[1], Hui Xu[1], Chang-Xiao Liu[2] and Yan Wang[1]**

[1]State Key Laboratory of Bioactive Substance and Function of Natural Medicines, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [2]Tianjin Institute of Pharmaceutical Research, Research Unit for Drug Metabolism, Chinese Academy of Medical Sciences, Tianjin, China

OPEN ACCESS

Edited by:
Juntao Liu,
Shandong University,
Weihai, China

Reviewed by:
Wei Hu,
Shandong University, China
José María Landete,
Instituto Nacional de Investigación y Tecnología Agroalimentaria (INIA), Spain
Ting Wang,
Qilu University of Technology, China
Bing Xu,
Beijing University of Chinese Medicine, China

***Correspondence:**
Yan Wang
wangyan@imm.ac.cn

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

Received: 21 February 2022
Accepted: 16 March 2022
Published: 06 April 2022

Citation:
Peng R, Han P, Fu J, Zhang Z-W, Ma S-R, Pan L-B, Xia Y-Y, Yu H, Xu H, Liu C-X and Wang Y (2022) Esterases From Bifidobacteria Exhibit the Conversion of Albiflorin in Gut Microbiota. Front. Microbiol. 13:880118. doi: 10.3389/fmicb.2022.880118

Bifidobacteria is an important microbe that inhabits the human gut. It is capable of metabolizing complex compounds in the human diet. Albiflorin, an antidepressant natural product from *Radix Paeoniae Alba* in China, is difficult to absorb after oral administration, and its metabolism has been proven to be closely related to the gut microbiota. In this study, we demonstrated *in vitro* that several *Bifidobacteria* species were able to convert albiflorin to benzoic acid, and four esterases (B2, B3, B4, and BL) from *Bifidobacterium breve* and *Bifidobacterium longum* were found through genome mining and modeled by SWISS-MODEL. B2 and B3 presented the strongest albiflorin metabolism ability. The optimal conditions, including temperature, buffer, and pH, for the conversion of albiflorin by the four esterases were investigated. Furthermore, the effect of esterase on the metabolism of albiflorin *in vivo* was confirmed by transplanting bacteria containing esterase B2. This study demonstrated the vital role of esterases from *Bifidobacteria* in the metabolism of natural compounds containing ester bonds, which could contribute to the development of new enzymes, microbial evolution, and probiotic adjuvant compounds for treatment.

Keywords: gut microbiota, genome mining, esterase structure, metabolism, *Bifidobacteria*, albiflorin, benzoic acid

INTRODUCTION

The human gut microbiota is mainly composed of anaerobic bacteria, and the number of cells per gram of intestinal contents exceeds 10^{11} (Sender et al., 2016). The dense gut microbiota greatly affects the host's metabolic capacity, nutritional status, and immune system development (Nicholson et al., 2012). In adults, intestinal bacteria mainly comprise species from Firmicutes, Bacteroides, Actinobacteria, and Proteobacteria. Actinomycetes are mainly represented by species of the genus *Bifidobacterium*, accounting for 2–10% of adult intestinal bacteria (Turroni et al., 2008; Arumugam et al., 2011). Nowadays, gut microbiota deviations are linked with many diseases including obesity, type 2 diabetes, hepatic steatosis, intestinal bowel diseases (IBDs), and several types of cancer (de Vos et al., 2022). The gut microbiota has both direct and indirect effects on drug and xenobiotic metabolisms, and this can have consequences for both efficacy and toxicity (Wilson and Nicholson, 2016). *Bifidobacteria* are some of the most important probiotics in gut microbiota (Rossi and Amaretti, 2010). They play a beneficial role *via* multiple

mechanisms, such as immune stimulation, anticancer activity, inhibition of pathogen growth, production of vitamins and amino acids, reduction of cholesterol, alleviation of the symptoms of irritable bowel syndrome, treatment of mood disorders, and bioconversion of a variety of natural compounds into bioactive compounds (Whorwell et al., 2006; Fanning et al., 2012; Ventura et al., 2014; Zanotti et al., 2015). Therefore, research on the relationship between humans and *Bifidobacteria*, including its beneficial effects on human health and its symbiosis mechanism, has received extensive attention. As important probiotics, *Bifidobacteria* species possess abundant enzymes that are of great significance. Recent studies have shown that the esterase of *Bifidobacterium longum* can metabolize hydroxycinnamic acid from food (Kelly et al., 2018), and the sialidase of *Bifidobacterium bifidum* can catalyze the removal of terminal sialic acids from various complex carbohydrates (Ashida et al., 2018). The esterases in *Bifidobacteria* can metabolize chlorogenic acid in food (Raimondi et al., 2015). However, little is known about the esterases of *Bifidobacteria* and drug metabolism.

As the main component of Xiaoyao Wan, a common Chinese patent medicine prescribed for the treatment of depression-like disorders, albiflorin has poor bioavailability (5.4%) and can hardly cross the blood–brain barrier. The level of albiflorin in the blood and brain is considerably low after oral administration (Huang et al., 2014). Previous research by our group found that *Bifidobacteria* in intestinal bacteria are closely related to the metabolism of albiflorin. The characteristic metabolite is benzoic acid (BA). The esterase in *Bifidobacteria* may be one of the key metabolic enzymes in the metabolic process (Zhao et al., 2018). Similarly, the esterase of *Bifidobacteria* is involved in the conversion of albiflorin, which is also the main component of Xiaoyao Wan (Yu et al., 2019). Esterases and lipases are two major hydrolases that can cleave ester bonds. Both of them have an α/β-hydrolase fold and contain a consensus sequence (Gly-X-Ser-X-Gly) adjacent to the catalytic triad Ser-Asp-His (Kroon et al., 2000; Bornscheuer, 2002). In contrast to lipases, esterases generally obey classical Michaelis-Menten kinetics and hydrolyze compounds that have less than six carbons (Arpigny and Jaeger, 1999; Bornscheuer, 2002).

This study aimed to explore the ability of probiotic *Bifidobacteria* species to hydrolyze albiflorin into benzoic acid and to identify the enzymes involved in this reaction. Hydrolysis seems to be a characteristic phenomenon of *Bifidobacteria*. Four *Bifidobacteria* strains found in the human intestine were chosen for the conversion of albiflorin. We detected the activity of esterases in *Bifidobacterium breve* and *B. longum*, which can hydrolyze albiflorin. By genome mining and modeling with SWISS-MODEL, four esterases were found, and they all have the core domain of esterase, but the surrounding regions are very different. The functional characterization of the four esterases, which were expressed in *Escherichia coli*, was performed and compared under different conditions. Further, we proved the important role of esterase B2 in affecting the metabolism of albiflorin and improving the concentration of benzoic acid in plasma by pharmacokinetic study *in vivo* with bacteria transplantation. This demonstrated the possible role of *Bifidobacteria* in albiflorin metabolism, which revealed new prospects for the development of novel enzyme preparations and probiotics specifically designed for the enhancement of the bioconversion of traditional Chinese medicinal (TCM) chemicals into biologically active compounds.

MATERIALS AND METHODS

Chemicals, Bacterial Strains, and Culture Conditions

Albiflorin, benzoic acid, propranolol, anaerobic medium, and MRS broth medium were all purchased from Solarbio (Beijing, China). The purity of the compounds was higher than 98% (HPLC). Albiflorin and benzoic acid were dissolved to prepare a 0.5 mmol/L stock solution in water. HPLC-grade ammonia, acetonitrile, and methanol were obtained from Fisher Scientific (Fair Lawn, NJ, United States).

The bacterial strains and plasmids employed in this study are listed in **Supplementary Table S1**. Four *Bifidobacteria* strains (*B. breve* ATCC15700, *B. longum* ATCC15697, *Bifidobacterium animalis* ATCC27673, and *Bifidobacterium adolescentis* ATCC15703) were provided by the ATCC Biological Resource Center. The *Bifidobacteria* strains were cultured in MRS medium at 37°C under anaerobic conditions in N_2 gas. The *E. coli* strains DH5α and BL21 (DE3) were used to clone the pET-28a plasmid and express the esterase protein, respectively. The *E. coli* strain was grown in Luria-Broth (LB) medium (10 g/L trypsin, 5 g/L yeast extract, and 5 g/L NaCl, pH 7.2) at 37°C. If necessary, 10–15 μg/ml kanamycin (km) was added to the medium.

The concentrations of albiflorin and benzoic acid were determined using an HPLC-MS/MS 8050 system from Shimadzu Corporation (Kyoto, Japan). The substance to be tested was separated in liquid phase with an Alltima C_{18} column (100 mm × 2.1 mm × 5 μm, Grace, England). The gradient elution started with 80% mobile phase A (water) and decreased to 50% mobile phase A over 5 min at a flow rate of 0.4 ml/min. Then, it was maintained at 50% for 2 min and returned to 80% mobile phase A over 1 min, where it was maintained for 10 min. Mobile phase B consisted of methanol. The column temperature and autosampler temperature were set at 30°C and 4°C, respectively. The mass spectrometer was run in multiple reaction monitoring (MRM) mode: 503.00 → 341.05 (*m/z*) for albiflorin (+), 121.10 → 77.10 (*m/z*) for benzoic acid (−), and 260.20 → 116.10 (*m/z*) for propranolol (−). The sample processing method was as follows: 100 μl of sample was added to 300 μl of methanol containing 50 ng/ml of internal standard (propranolol). Then, the mixture was vortexed for 30 s and centrifuged at 12,000 rpm for 10 min. The supernatant was prepared for quantitative analysis by LC-MS/MS, and the injection volume was 1 μl.

Animals

Sprague-Dawley (SD) male rats weighing 200 ± 20 g and ICR mice weighting 18 ± 22 g were provided by the Institute of Laboratory Animal Science, Chinese Academy of Medical Sciences (Beijing, China). The animals were placed in a cage

TABLE 1 | The information of esterases in *Bifidobacteria* strains.

Name	NCBI reference sequence	Source	Number of amino acids	Molecular weight	Theoretical pI	Superfamily	Accession	Definition
B2	WP_003829196.1	*Bifidobacterium breve* ATCC 15700	273	30,219.38	5.19	MhpC	cl33968	Alpha/beta hydrolase
B3	WP_003828396.1	*Bifidobacterium breve* ATCC 15700	281	30,261.08	4.50	protocat_pcaD	cl31213	Alpha/beta hydrolase
B4	WP_003828023.1	*Bifidobacterium breve* ATCC 15700	319	35,264.78	5.22	MhpC	cl33968	Alpha/beta hydrolase
BL	KAB6720564.1	*Bifidobacterium longum* ATCC 15697	252	27,695.84	4.79	Abhydrolase	cl21494	Esterase

with free access to chow and water. The temperature of the cage was kept at approximately 22°C with 50% humidity and a 12-h day/night cycle. Rats were fasted for 12 h before the experiment but had free access to water. This study was conducted in accordance with institutional and ethics guidelines approved by the Laboratory Institutional Animal Care and Use Committee of the Chinese Academy of Medical Sciences and Peking Union Medical College.

Bioconversion of Albiflorin With Rat Intestinal Bacteria and *Bifidobacteria* Strains

We obtained the colon contents from six SD rats and transferred 5 g of the collected sample into 100 ml anaerobic medium. Then, the mixture was preincubated under anaerobic conditions (N_2 atmosphere) at 37°C for 1 h. After preincubation, 10 µl of albiflorin was transferred into 990 µl of the culture. The cultures were inactivated by high temperature and high pressure at 121°C as a negative control. The final concentrations of albiflorin in the incubation system were 0.05 mM and 0.5 mM. The cultures were incubated at 37°C for 0, 12, or 24 h. The method was performed in accordance with the previously described procedure (He et al., 2017).

Albiflorin and its metabolite were analyzed by an Alltima C_{18} column (100 mm × 2.1 mm × 5 µm, Grace, England) with an LC/MS^n-IT-TOF system (Shimadzu Corporation, Kyoto, Japan) in both positive and negative modes. Mass spectra were acquired in the range of *m/z* 100–1,000 for MS^1. The MS^n data were collected in automatic mode. An elution gradient was employed at a flow rate of 0.4 ml/min, with water as mobile phase A and methanol as mobile phase B, by using the following program: 0.10 min (90% A and 10% B), 3.00 min (60% A and 40% B), 7.00 min (40% A and 60% B), 9.00 min (10% A and 90% B), 9.01 min (90% A and 10% B), and 12.00 min (controller stop).

Four strains of *Bifidobacteria*, *B. breve* ATCC15700, *B. longum* ATCC15697, *B. animalis* ATCC27673, and *B. adolescentis* ATCC15703 were cultured to convert albiflorin *in vitro*. After resuscitation, the strains were cultured overnight in MRS medium at 37°C under anaerobic conditions (N_2 atmosphere), and the bacterial concentration was uniformly maintained at 3×10^6 cells/ml. The cultures were incubated at 37°C for 0, 12, and 24 h, and the final concentration of albiflorin in the incubation system was 0.05 mM. The contents of albiflorin and its metabolite benzoic acid were determined and analyzed by LC-MS/MS.

Bioinformatics Analyses

The NCBI database[1] was used to search the genomes of *B. breve* and *B. longum* for genes labeled "carboxylesterase" and "esterase." By searching for genes annotated as "carboxylesterase" in the genome of *B. breve* ATCC15700 (NZ_AP012324.1), we found three carboxylesterase genes, NZ_AP012324.1: 1013236–1014057, NZ_AP012324.1: 506173–507018, and NZ_AP012324.1: 506173–507018. The amino acid sequences can be accessed as WP_003829196.1 (B2, 273 aa), WP_003828396.1 (B3, 281 aa), and WP_003828023.1 (B4, 319 aa). By searching for the gene annotated as "esterase" in the genome of *B. longum* ATCC15697 (NC_011593.1), we found an esterase gene, NC_011593.1: c2783438-2782680, and its protein sequence can be found in KAB6720564.1 (BL, 252 aa). Detailed information on each protein was obtained from the NCBI database, as shown in **Table 1**.

3-D structural modeling of the esterase was conducted with SWISS-MODEL,[2] which was the first fully automated protein homology modeling server (Guex and Peitsch, 1997; Schwede et al., 2003; Bordoli et al., 2009). The SWISS-MODEL workspace can be freely accessed at http://swissmodel.expasy.org/workspace/. We used the automatic modeling mode and applied the protein sequences of four esterases, which were available in GenBank. The data obtained from the homology model were visualized using DeepView ver. 4.0.1 (The Swiss Institute of Bioinformatics).

Plasmids Construction for Heterologous Expression of the Esterases From *Bifidobacterium breve* and *Bifidobacterium longum*

To obtain the genomic DNA of *B. breve* and *B. longum*, extraction was carried out according to the instructions of the bacterial genome extraction kit (Tiangen, China). Using the genomic DNA of *B. breve* and *B. longum* as templates, Primer Premier 5.0 was used to design primers based on the four esterase genes (*b2*, *b3*, *b4*, and *bl*) and the vector pET-28a. All of the primers (b2-F, b2-R, b3-F, b3-R, b4-F, b4-R, bl-F, bl-R, pET28a-F, and pET28a-R) are listed in **Supplementary Table S2**. By using the genomic DNA of *B. breve* and *B. longum* as templates, the four esterase genes were amplified by PCR and then purified along with the PCR-amplified vector DNA using the Gel Recovery Purification Kit (Tiangen, China). A Gibson Assembly Kit (NEB, United States) was used

[1]http://www.ncbi.nlm.nih.gov
[2]https://swissmodel.expasy.org

to insert the four esterase DNA fragments into the vector pET-28a to construct the recombinant vectors. *Escherichia coli* DH5α cells were transformed by the heat shock method and used for plasmid propagation to obtain the recombinant plasmids pET-28a-*b2*, pET-28a-*b3*, pET-28a-*b4*, and pET-28a-*bl*. All fragments were validated by Sanger 3730 sequencing.

Expression of *b2, b3, b4,* and *bl* in *Escherichia coli* and Functional Verification

The recombinant plasmids pET-28a-*b2*, pET-28a-*b3*, pET-28a-*b4*, and pET-28a-*bl* were transformed into *E. coli* BL21 (DE3) cells. Cells were grown in liquid LB media supplemented with kanamycin (40 μg/ml). Cultures were maintained at 37°C until the $OD_{600\,nm}$ reached 0.6. Then, the temperature was lowered to 16°C, and protein expression was induced with 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) for 20 h. The bacterial cells were harvested by centrifugation (12,000 × *g*, 30 min, 4°C), concentrated 10-fold, and suspended in approximately 10 ml of lysis buffer (10 mM PBS buffer or 25 mM Tris-HCl buffer). After ultrasonic lysis, the supernatant and precipitate were prepared for SDS-PAGE analysis, with the empty pET-28a plasmid in *E. coli* BL21 cells as a control. The B2 and B3 proteins were purified using the PrepEase His-tag protein purification kit (Takara, Japan).

Escherichia coli cells with B2, B3, B4, and BL and the cell extract supernatant were used for the bioconversion of albiflorin to characterize B2, B3, B4, and BL. *Escherichia coli* BL21 (DE3) transformed with the empty pET-28a vector was used as a negative control. The final concentration of albiflorin in the incubation system was 0.05 mM. The cultures were incubated at 37°C for 0, 12, and 24 h. Albiflorin and its metabolite benzoic acid were detected by LC-MS/MS. The enzymatic reaction conditions were as follows: temperatures, 30°C and 37°C; buffers, PBS buffer and Tris-HCl buffer; and pH, 6.0, 7.0, and 8.0. The LC-MS/MS conditions refer to the above experimental methods.

Microbiota Transplantation and Pharmacokinetic Study *in vivo*

Three groups of ICR mice ($n = 6$/group) were orally administered *E. coli* (1×10^{10} CFU), the *E. coli* with esterase B2 (1×10^{10} CFU), and *B. breve* (1×10^{10} CFU) once a day for 3 days, respectively. Another six ICR mice were orally administered an equivalent volume of saline. Pharmacokinetic evaluations were performed 3 days after the final administration.

Before the oral administration of a single dose of albiflorin (7 mg/kg), the four groups of mice were fasted overnight with free access to water. Blood samples were collected before and at 5, 10, 20, 30, 45, 60, 90, 120, 180, 240, 360, and 480 min after drug treatment.

Statistical Analysis

Statistical analyses were performed with GraphPad Prism Version 5 (GraphPad Software, CA, United States) using two-way ANOVA and Student's *t*-test. Data are expressed as the mean ± standard deviation, and p values less than 0.05 were considered statistically significant.

RESULTS

Conversion of Albiflorin Into Benzoic Acid in Gut Microbiota

Intestinal bacteria from SD rats were collected for anaerobic incubation in the presence of albiflorin, and the concentrations of albiflorin in the culture solution were 0.05 and 0.5 mM, respectively, with the inactivated intestinal bacteria culture functioning as a negative control. The structure of albiflorin (**Figure 1A**) includes a benzoyl group, which might be hydrolyzed by esterases present in gut microbiota. LC/MS^n-IT-TOF was used for identifying albiflorin and BA. The mass spectrum of albiflorin is shown in **Figure 1B**. The *m/z* of the $[M+Na]^+$ peak was 503, the *m/z* of the secondary fragments were 201, 307, and 341, and the tertiary ion fragment had an *m/z* of 175. The mass spectrum of benzoic acid is shown in **Figure 1C**, and the $[M+H]^+$ peak of benzoic acid had an *m/z* of 122. Moreover, different concentrations of albiflorin (0.05 and 0.5 mM) were cultured with the intestinal bacterial culture and inactivated intestinal bacterial culture at 37°C for 12 and 24 h, and the concentrations of albiflorin and benzoic acid were determined after incubation. As shown in **Figures 1D,E**, regardless of the concentration, the albiflorin cultured with the intestinal bacteria solution was completely converted after 12 h, and the metabolite benzoic acid was produced. However, albiflorin cultured with the inactivated intestinal bacteria solution was not converted into benzoic acid. The results indicated that the intestinal bacteria of SD rats were able to metabolize albiflorin and convert it into benzoic acid.

Bioconversion of Albiflorin by Four *Bifidobacteria* Strains

The capability of four common *Bifidobacteria* strains (*B. breve*, *B. longum*, *B. animalis*, and *B. adolescentis*) found among gut microbes to metabolize albiflorin under anaerobic conditions over a period of 24 h was examined. The cells were maintained at a consistent concentration with 3×10^6 cells/ml in the presence of 0.05 mM albiflorin. Samples were removed at 0, 12, and 24 h for LC-MS/MS detection. As shown in **Figures 1F,G**, *B. breve*, *B. longum*, and *B. adolescentis* showed significant metabolism of albiflorin at 12 and 24 h, while *B. longum* had the strongest ability to metabolize albiflorin and resulted in a high conversion rate at 24 h of 18.1%, followed by the rates of *B. breve* (16.8%) and *B. adolescentis* (1.5%). Accordingly, benzoic acid was produced differently by the tested *Bifidobacteria* strains, resulting in 7.14 μg/ml for *B. breve*, 5.29 μg/ml for *B. longum*, and 4.96 μg/ml for *B. adolescentis*. According to the above results, *B. breve* and *B. longum* have a strong albiflorin metabolism ability, and we propose that both play an important role in albiflorin metabolism in the gut microbiota.

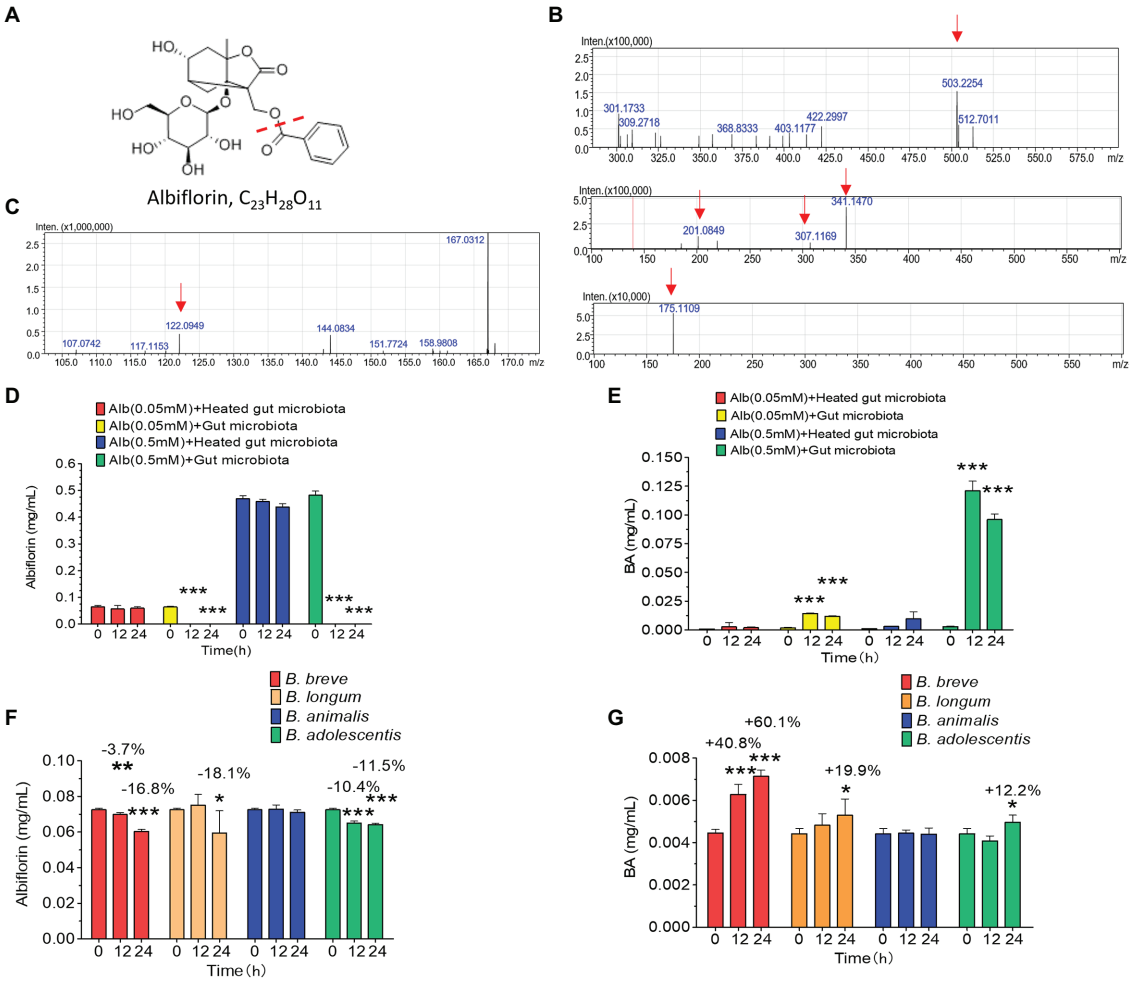

FIGURE 1 | The intestinal microbes of rats converted albiflorin into benzoic acid. **(A)** Schematic diagram of the structure of albiflorin and benzoic acid. **(B)** High-resolution mass spectrum of albiflorin. **(C)** High-resolution mass spectrum of benzoic acid. **(D)** Conversion of albiflorin by the intestinal microbes of rats. **(E)** Benzoic acid (BA) generated by the intestinal microbes of rats. **(F)** Conversion of albiflorin in the four *Bifidobacteria* strains. **(G)** Benzoic acid (BA) generated by the four *Bifidobacteria* strains. (Student's *t*-test, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$, data are expressed as the mean ± SD).

In silico Search of *Bifidobacterium* Esterases

By searching for genes annotated as "carboxylesterase" in the genome of *B. breve* ATCC15700 (NZ_AP012324.1), three carboxylesterase genes were found, including NZ_AP012324.1: 1013236–1014057, NZ_AP012324.1: 506173–507018, and NZ_AP012324.1: 506173–507018. The corresponding amino acid sequences are WP_003829196.1 (B2, 273 aa), WP_003828396.1 (B3, 281 aa), and WP_003828023.1 (B4, 319 aa). By searching for the genes annotated as "esterase" in the genome of *B. longum* ATCC15697 (NC_011593.1), an esterase gene was found, which was named NC_011593.1:c2783438-2,782,680, and its protein identifier is KAB6720564.1 (BL, 252 aa). Detailed information on each protein was obtained from the NCBI database, and the ProtParam tool by ExPASy[3] was used to predict the size and isoelectric point of each protein. As shown in **Table 1**, the predicted molecular weight of B2 was 30,219.38 Da, and the isoelectric point was 5.19. The values of the molecular weight and isoelectric point of the other proteins were as follows: B3: 30,261.08 Da, 4.50; B4: 35,264.78 Da, 5.22.; and BL: 27,695.84 Da, 4.79.

We found that the four enzymes, B2, B3, B4, and BL, belong to three superfamilies. As shown in **Table 1**, B2 and B4 belong to the MhpC superfamily, which is annotated as "Pimeloyl-ACP methyl ester carboxylesterase." B3 belongs to the protocat_pcaD superfamily, which is annotated as "3-oxoadipate enol-lactonase. Note that the substrates are 3-oxoadipate enol-lactone, 2-oxo-2,3-dihydrofuran-5-acetate, 4,5-dihydro-5-oxofuran-2-acetate, and 5-oxo-4,5-dihydrofuran-2-acetate." BL belongs to the abhydrolase superfamily, which is annotated as "alpha/beta hydrolases. A functionally diverse superfamily containing proteases, lipases, peroxidases, esterases, epoxide hydrolases and dehalogenases."

[3]https://web.expasy.org/protparam/

We modeled the four esterases, B2, B3, B4, and BL, by using SWISS-MODEL. As shown in **Figure 2A**, they all have the core domain of an esterase, but the surrounding regions are very different. The modeled structure showed that their core structures are similar and that they may all be esterases. These enzymes are composed of highly similar α/β-hydrolase folds and different helix domains which suggested that they have different cleavage patterns. They are esterases rather than lipases as they elicit a preference for smaller carbon backbone substrates less than six carbons. It has previously been reported that the bifidobacterial esterase from *B. animalis* subsp. *lactis* WC 0432 exhibits hydrolytic activity against chlorogenic acid (Raimondi et al., 2015).

Expression of *Bifidobacterium* Esterases in *Escherichia coli*

To investigate the suitability of *E. coli* BL21 (DE3) for *Bifidobacterium* esterase expression, we first verified the absence of homologous genes in its genome by analyzing its genome, and the *E. coli* cells with the empty vector pET-28a were incapable of metabolizing any albiflorin (data not shown). Several primer pairs designed based on the B2, B3, B4, and BL coding sequences of *B. breve* ATCC15700 and *B. longum* ATCC15697 were used to clone the *four genes* (the primer sequences are shown in **Supplementary Table S1**). The four PCR-amplified esterase DNA fragments (as shown in **Figure 2B**) were recombined into the vector pET-28a by Gibson assembly. DH5α *E. coli* were used to clone the esterase coding sequences of *B. breve* and *B. longum*, and four plasmids, pET-28a-*b2*, pET-28a-*b3*, pET-28a-*b4*, and pET-28a-*bl*, were obtained. *Escherichia coli* BL21 (DE3) was transformed with the above expression vectors bearing four esterase genes. SDS-PAGE showed that the expression of the four esterases occurred at high levels in IPTG-induced recombinant cells (**Figures 2C,D**).

Esterase Activity Characterization *in vitro*

Based on the above results, four *E. coli* strains (EB2, EB3, EB4, and EBL) expressing the four esterases were obtained. After being cultured overnight with IPTG, the four *E. coli* strains were individually used to investigate albiflorin metabolism at 37°C for 24 h. The cells of each strain were controlled at a consistent concentration of $OD_{600\,nm} = 2.0$ in the presence of 0.05 mM albiflorin. Samples were removed at 0, 12, and 24 h for LC-MS/MS analysis. As shown in **Figures 3A,B**, *E. coli* EB2 had a significant effect on albiflorin metabolism, the level of which was reduced by 31.9% at 24 h compared with that at 0 h. Moreover, the production of benzoic acid at 24 h was increased by 5.37-fold compared with that at 0 h. The concentration of benzoic acid reached 23.9 μg/ml. However, EB3, EB4, and EBL had no effect on albiflorin

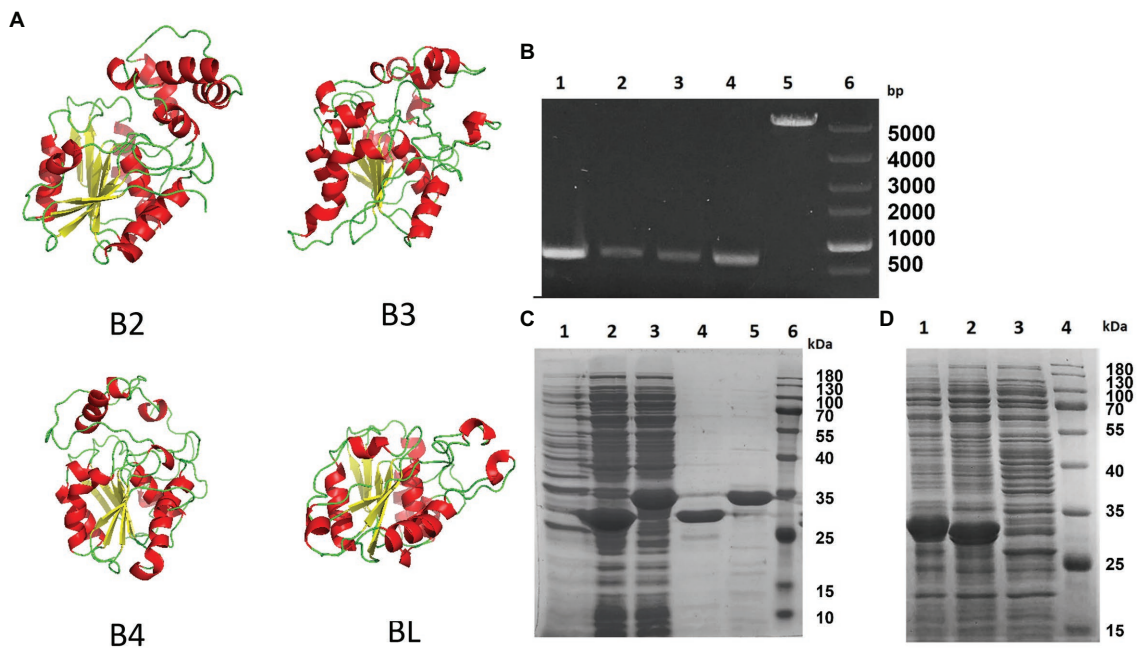

FIGURE 2 | Homology model and expression of four *Bifidobacterium* esterases. **(A)** Homology model of the four esterases, B2, B3, B4, and BL, generated by SWISS-MODEL. **(B)** Agarose gel electrophoresis image showing the PCR-amplified products of four esterase genes from *Bifidobacterium breve* and *Bifidobacterium longum*. Lane 1, *bb2*; lane 2, *bb3*; lane 3, *bb4*; lane 4, *bl*; lane 5, pET-28a; and lane 6, supercoiled DNA marker. **(C)** SDS-PAGE analysis of two proteins, B2 and B3, from *B. breve* after being expressed in *Escherichia coli* BL21. Lane 1, total protein from the lysate of *E. coli* transformed with pET-28a; lane 2, total protein of *E. coli* lysate containing esterase B2; lane 3, total protein of *E. coli* lysate containing esterase B3; lane 4, esterase B2 after purification; lane 5, esterase B3 after purification; and lane 6, prestained protein molecular weight marker. **(D)** SDS-PAGE analysis of two esterases, B4 and BL, from *B. breve* and *B. longum* in *E. coli* BL21. Lane 1, total protein of *E. coli* lysate containing esterase B4; lane 2, total protein of *E. coli* lysate containing esterase BL; lane 3, total protein of *E. coli* lysate transformed with pET-28a; and lane 4, prestained protein molecular weight marker.

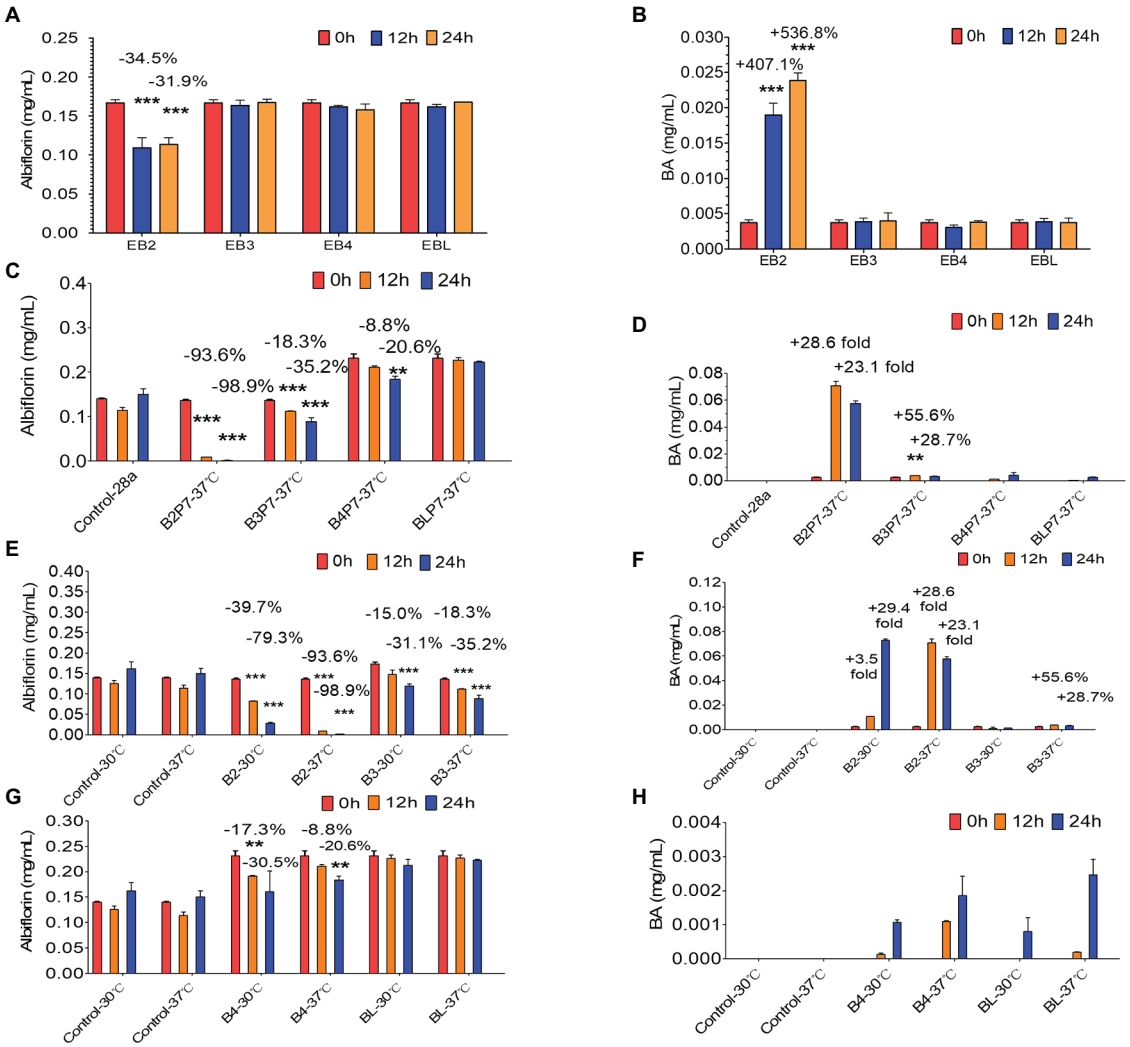

FIGURE 3 | Albiflorin metabolism by four esterases (B2, B3, B4, and BL). **(A)** Conversion of albiflorin in *Escherichia coli* stains (EB2, EB3, EB4, and EBL). **(B)** Benzoic acid (BA) generated by the *E. coli* strains (EB2, EB3, EB4, and EBL). **(C)** Conversion of albiflorin by four esterases. **(D)** Benzoic acid (BA) generated by four esterases. **(E)** Conversion of albiflorin by B2 and B3 at 30°C or 37°C. **(F)** Benzoic acid (BA) generated by B2 and B3 at 30°C or 37°C. **(G)** Conversion of albiflorin by B4 and BL at 30°C or 37°C. **(H)** Benzoic acid (BA) generated by B4 and BL at 30°C or 37°C (Student's *t*-test, $^{**}p < 0.01$ and $^{***}p < 0.001$, data are expressed as the mean ± SD).

metabolism, and no significant increase was detected in the production of benzoic acid. This result showed the metabolic ability of the four esterases expressed by *E. coli* cells in suspension culture, and the *E. coli* strain expressing esterase B2 showed the greatest potential for albiflorin metabolism.

Furthermore, the *E. coli* cells expressing the four esterases were collected and resuspended in PBS buffer (pH = 7) and sonicated to obtain the B2, B3, B4, and BL proteins. The supernatants were mixed with 0.05 mM albiflorin and incubated at 37°C. Samples were taken at 0, 12, and 24 h for the detection of albiflorin and its metabolite benzoic acid. As shown in **Figure 3C**, after 24 h of reaction, 98.9% of albiflorin was converted by B2, 35.2% by B3, and 20.6% by B4, but the conversion by BL was not significant. The production of benzoic acid is shown in **Figure 3D**. The concentrations of benzoic acid produced by B2 and B3 were 70.8 and 3.7 μg/ml, respectively, but the concentrations produced by B4 and BL were not obvious compared with those at 0 h. The results indicated that the B2 and B3 proteins have a strong ability to metabolize albiflorin, and B2 has the strongest conversion effect, followed by B3 and B4. Similarly, the B2 reaction system produced the most benzoic acid, followed by the B3 and B4 reaction systems, which was consistent with the above results.

Effects of Esterases on Albiflorin Metabolism Under Different Conditions

Furthermore, the effects of different temperatures, buffers, and pH on the ability of the four enzymes to convert albiflorin

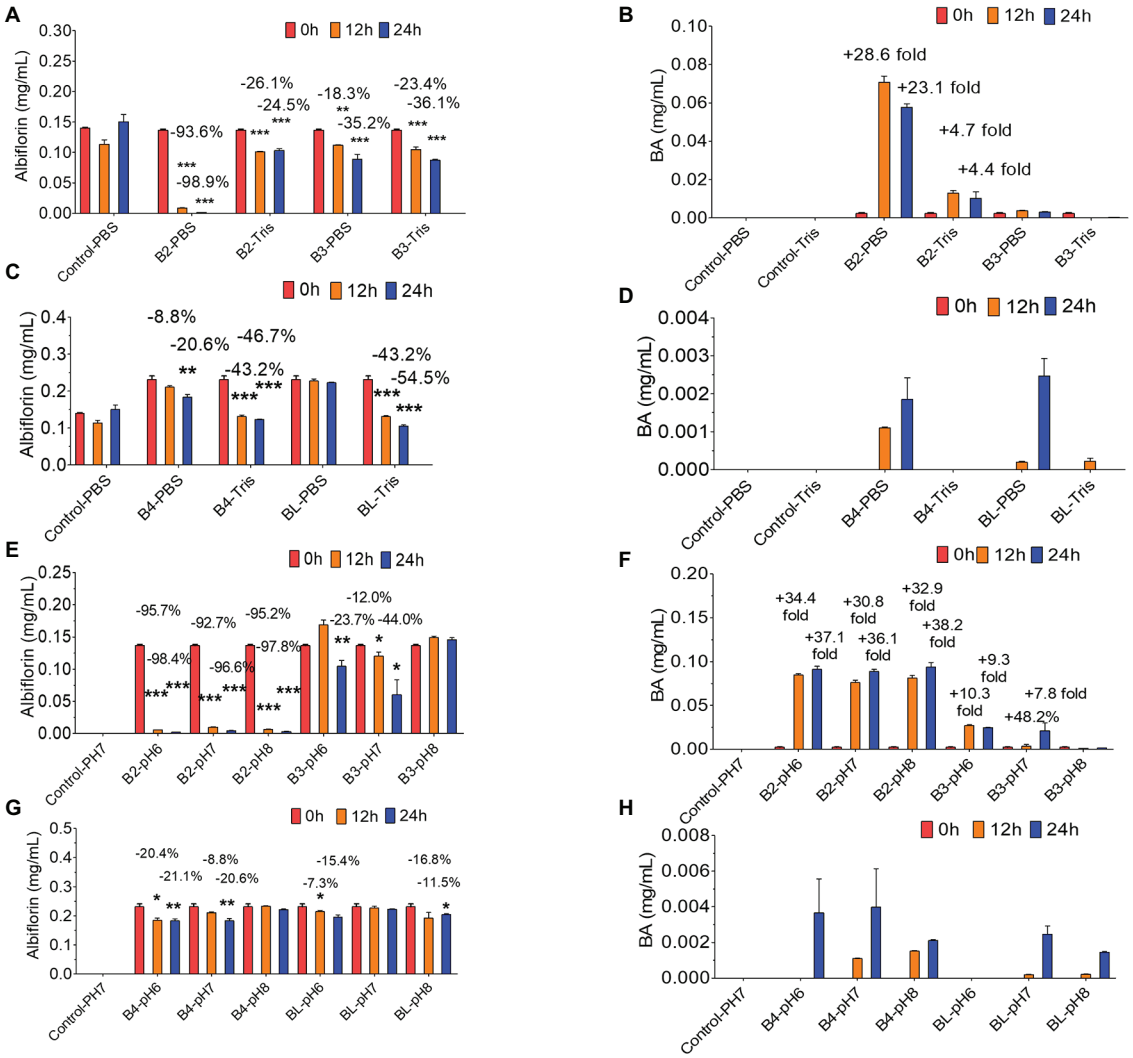

FIGURE 4 | Albiflorin metabolism by four esterases (B2, B3, B4, and BL) under different conditions. **(A)** Conversion of albiflorin by B2 and B3 in PBS and Tris-HCl buffer. **(B)** Benzoic acid (BA) generated by B2 and B3 in PBS and Tris-HCl buffer. **(C)** Conversion of albiflorin by B4 and BL in PBS and Tris-HCl buffer. **(D)** Benzoic acid (BA) generated by B4 and BL in PBS and Tris-HCl buffer. **(E)** Conversion of albiflorin by B2 and B3 at pH 7.0, 8.0, and 9.0. **(F)** Benzoic acid (BA) generated by B2 and B3 at pH 7.0, 8.0, and 9.0. **(G)** Conversion of albiflorin by B4 and BL at pH 7.0, 8.0, and 9.0. **(H)** Benzoic acid (BA) generated by B4 and BL at pH 7.0, 8.0, and 9.0 (Student's *t*-test, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$, data are expressed as the mean ± SD).

were investigated. First, as shown in **Figures 3E,F**, after 24 h of reaction at 37°C, B2 and B3 converted 98.9 and 35.2% of albiflorin, respectively, which was greater than that observed at 30°C (79.3 and 31.1%). The production of benzoic acid at 37°C was also greater than that at 30°C. The conversion capability of B4 and BL are shown in **Figures 3G,H**. B4 converted 30.5% albiflorin at 30°C in 24 h, and the percentage at 37°C was 20.6%. However, there was no significant difference for BL between 30°C and 37°C in terms of albiflorin metabolism. Both temperatures led to the production of very little benzoic acid under the reaction conditions.

Then, the effect of the reaction conditions of the PBS buffer and Tris-HCl buffer on the conversion of albiflorin by the four enzymes was investigated. As shown in **Figures 4A,B**, under the reaction conditions of the PBS buffer solution, after 24 h of reaction, B2 converted 98.9% of albiflorin, which was better than the 24.5% conversion observed under the reaction conditions of the Tris-HCl buffer solution, and produced 57.6 μg/ml of benzoic acid in PBS buffer, which was better than the 13.6 μg/ml produced in Tris-HCl buffer. B3 converted 35.2% of the albiflorin in PBS buffer, which is slightly lower than the 36.1% converted in Tris-HCl buffer, and the production of benzoic acid in PBS buffer was also slightly lower than that in Tris-HCl buffer. As shown in **Figures 4C,D**, the amount of albiflorin converted by B4 was reduced by 46.7% in Tris-HCl buffer, which was better than the 20.6% conversion rate in PBS buffer. The amount of albiflorin converted by BL was reduced by 54.5% in Tris-HCl buffer, which was better than

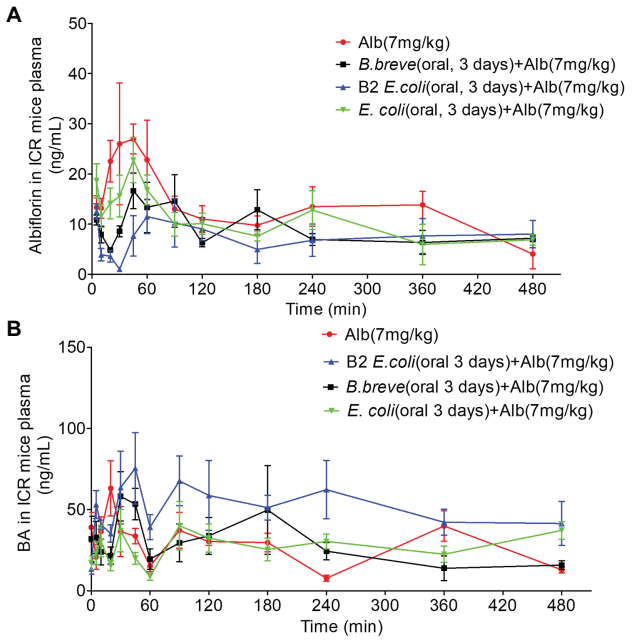

FIGURE 5 | Pharmacokinetic study of albiflorin *in vivo*. **(A)** Plasma concentration-time profiles of albiflorin (Alb) in four groups of mice (oral 3 days, NS, *Bifidobacterium breve* 1×10^{10} CFU/day, *Escherichia coli* with B2 1×10^{10} CFU/day, and *E. coli* 1×10^{10} CFU/day) after oral administration (7 mg/kg). **(B)** Plasma concentration-time profiles of Benzoic acid (BA) in four groups of mice (oral 3 days, NS, *B. breve* 1×10^{10} CFU/day, *E. coli* with B2 1×10^{10} CFU/day, and *E. coli* 1×10^{10} CFU/day) after oral administration of albiflorin (7 mg/kg).

20.6% conversion rate in PBS buffer. However, under the two buffer conditions, B4 and BL produced very little benzoic acid.

Finally, the effect of different pH buffer reaction conditions on the conversion of albiflorin by the four enzymes was investigated. As shown in **Figure 4E**, under the conditions of pH 6.0, pH 7.0, and pH 8.0, after 24 h of reaction, B2 converted 98.4, 96.6, and 97.8% albiflorin, respectively, and B3 converted 23.7, 44.0, and 0% albiflorin, respectively. As shown in **Figure 4F**, under the conditions of pH 6.0, pH 7.0, and pH 8.0, after 24 h of reaction, B2 produced 91.2, 88.8, and 93.8 μg/ml benzoic acid, respectively, and B3 produced 24.6, 21.2, and 1.6 μg/ml benzoic acid, respectively. As shown in **Figure 4G**, at pH 6.0 and pH 7.0, after 24 h of reaction, B4 converted 21.1 and 20.6% albiflorin, respectively, but the conversion at pH 8.0 was not obvious. The conversion rate of albiflorin by BL was 15.4 and 11.5% at pH 6.0 and pH 8.0, respectively, and little conversion occurred at pH 7.0. As shown in **Figure 4H**, after 24 h at pH 6.0, pH 7.0, and pH 8.0, B4 and BL produced very little benzoic acid.

Bacterial Transplantation Affects Drug Metabolism of Albiflorin *in vivo*

After 3 days of oral administration of different bacteria (Normal saline, NS, *E. coli* 1×10^{10} CFU/day, *E. coli* with Esterase B2 1×10^{10} CFU/day, *B. breve* 1×10^{10} CFU/day), ICR mice were given a single dose of albiflorin (7 mg/kg) to investigate the differences in albiflorin and benzoic acid concentrations over time in the four groups (**Figures 5A,B**). Pharmacokinetic parameters were shown in **Supplementary Table S3**. The $AUC_{(0-t)}$ of the four groups were 5,965.590 ± 2,004.089, 5,452.317 ± 1,045.409, 4,254.94 ± 1,591.654, and 4,355.606 ± 934.475 μg/L h, respectively. The C_{max} of the four groups were 44.064 ± 24.12, 45 ± 10.607, 21.61 ± 5.363, and 26.528 ± 11.105 μg/L, respectively. Similar pharmacokinetic parameters were observed in the NS group and the *E. coli* group. The transplantation of *E. coli* had little effect on the drug metabolism of albiflorin. The administration of *E. coli* containing esterase B2 obviously affected the absorption and metabolism of albiflorin *in vivo*, while the administration of *B. brevis* containing esterase B2 also affected the metabolism of albiflorin *in vivo* to a certain extent. At the same time, the plasma concentration-time profiles of albiflorin metabolites benzoic acid were determined, the results showed that the benzoic acid concentration of B2 group obviously greater than the NS group, the *B. breve* group could also increase the benzoic acid concentration in plasma, and the benzoic acid concentration in plasma of *E. coli* group and NS group is similar. *In vivo* experiments showed that supplement of esterase B2 could affect the metabolism of albiflorin in mice and significantly increased the concentration of benzoic acid, the metabolite of albiflorin in plasma. Esterase B2 is an important enzyme for the conversion of albiflorin to benzoic acid.

DISCUSSION

As a result of the in-depth study of gut microbiota, increasing evidence shows that gut microbiota is involved in drug metabolism, which plays a key role in determining the therapeutic effect and host metabolism (Sousa et al., 2009; Spanogiannopoulos et al., 2016; Koppel et al., 2017; Wu and Tan, 2018). Different bacteria produce different metabolic enzymes, such as β-glucuronidase, β-galactosidase, β-glucosidase, nitroreductase, azo reductase, protease, 7-αhydroxylase, and various carbohydrate enzymes, which are responsible for metabolizing different drugs (Schackmann, 1992; Lavrijsen et al., 1995; Ionescu and Caira, 2005; Pollet et al., 2017). Many natural compounds have low bioavailability, and their metabolism depends on the gut microbiota. Among them, natural compounds containing ester bonds are very abundant, including chlorogenic acid, geniposide, asiaticoside, ferulic acid, caffeic acid, and motherwort. Our previous laboratory research showed that albiflorin can be metabolized by intestinal bacteria (Zhao et al., 2018). By comparing the ability of 18 intestinal bacterial strains, we found that four bacteria had a significant ability to metabolize albiflorin, including *B. breve*. *Bifidobacteria* is an important constituent of human intestinal bacteria (accounting for 2–10%), and it contains abundant enzyme resources, such as α-glucosidase, β-glucosidase, hydrolase, and esterase (Kim and Lee, 2010; Kim et al., 2017; Florindo et al., 2018).

Several taxa at the family and genus levels, specifically family Prevotellaceae, genus *Corprococcus*, and *Faecalibacterium*, were decreased in major depressive disorder

(MDD) compared to non-depressed controls in observational studies, and depressive symptoms were improved compared to controls in interventional studies with probiotics (Sanada et al., 2020). Studies have shown that *B. longum* and *B. breve* strains reduce stress-, anxiety-, and depression-related behaviors in male BALB/c mice with congenital anxiety. This beneficial effect is related to its influence on enteric neurons and vagus nerve signals (Savignac et al., 2014). Preclinical and clinical studies have shown that *Bifidobacteria* may have therapeutic effects on mood disorders (Desbonnet et al., 2010). In addition to its beneficial effect on human health, *Bifidobacteria* was also shown to affect TCM efficacy. For example, daidzin, a glycosidic isoflavone that mainly exists in soy products, can be metabolized into equol by *Bifidobacterium via* glycosidic cleavage and reduction of an α, β-unsaturated ketone (Setchell and Clerici, 2010). For *Bif. animalis* subsp. *lactis*, the metabolism of caffeic acid and chlorogenic acid was described (Couteau et al., 2001; Raimondi et al., 2015; Fritsch et al., 2016a).

In this study, we used four common *Bifidobacteria* strains from the intestinal flora to examine the metabolic effect of *Bifidobacteria* on albiflorin. Albiflorin can be converted into benzoic acid, which is the main metabolite, by intestinal bacteria in large amounts. This is similar to the antidepressant mechanism of albiflorin speculated previously. According to the comparison of the conversion abilities of the four *Bifidobacteria* strains and genome mining results, we selected *B. breve* and *B. longum*, which have a strong albiflorin metabolism ability. Previous studies have shown that esterases from *B. animalis* subsp. *lactis* DSM 10140 and *B. animalis* subsp. *lactis* WC 0432 were able to hydrolyze HCA-containing substrates and chlorogenic acid (Raimondi et al., 2015; Fritsch et al., 2016b). In this study, three esterases, B2, B3, and B4, from *B. breve* and one esterase, BL, from *B. longum*, which belong to three superfamilies, were identified and characterized. Esterase (EC 3.1.1.1), which is generally known as carboxylesterase, belongs to the class of hydrolytic enzymes, the function of which is to hydrolyze carboxylic acid esters (Porro et al., 2019). It can catalyze ester bond cleavage and form the corresponding alcohols and acids in the presence of H_2O molecules. By using SWISS-MODEL, it was found that all four esterases have the core domain of esterase, but the surrounding regions are very different. These enzymes are composed of highly similar α/β-hydrolase folds and different helix domains. The catalytic apparatus of alpha/beta hydrolases typically involves three residues (catalytic triad): a serine, a glutamate/aspartate, and a histidine. The catalytic mechanism often involves a nucleophilic attack on a carbonyl carbon atom (Jochens et al., 2011; Bauer et al., 2020).

Then, the albiflorin metabolism function of the four esterases was verified. It was demonstrated that at 37°C, B2 and B3 showed stronger hydrolysis function and produced more benzoic acid. B4 had a weaker metabolic capacity, and BL could not convert albiflorin. To explore the influence of the enzyme reaction conditions, different temperatures, buffers, and pH conditions were investigated to explore the most suitable reaction conditions for the four enzymes. It is revealed that at 37°C, PBS buffer and pH 7.0 are the optimal conditions for B2, which has the strongest ability to convert albiflorin. This result indicated that B2 might be the most important esterase and play a crucial role in hydrolysis by *B. breve* to convert albiflorin. Furthermore, by pharmacokinetic analysis of albiflorin in animals, we verified that transplantation of *E. coli* containing esterase B2 affected the drug metabolism of albiflorin, and the effect was more obvious than that of *B. brevis* containing B2. Recent studies by Zimmermann et al. (2019a,b) have revealed that the contributions of the microbiome to the metabolism of some drugs are much more than 50%, it provided insight into the important roles of the gut microbiota in the metabolism of many pharmaceuticals. For instance, some phenolic compounds with low bioavailability, such as rosmarinic acid and eriodictyol, are fermented into absorbable and bioactive phenolic acids by the gut microbiota, e.g., hydroxyphenylpropionic acids and phenylpropionic acids (Jochens et al., 2011; Zimmermann et al., 2019a). These bioactive microbial metabolites may be absorbed and transported by the circulatory system to tissues and organs or exert their effects in the intestinal lumen (Aura, 2008; Hanske et al., 2009; Mosele et al., 2014). Gut bacteria produce a range of enzymes that might chemically alter drugs as varied as psychotropics and cancer treatments, rendering them less useful or leading to more side effects (Paolo et al., 2018; Spanogiannopoulos and Turnbaugh, 2018).

In summary, this study examined the metabolism of albiflorin, a compound with antidepressant activity, by the enzymes contained in intestinal bacteria and confirmed that benzoic acid is the main metabolite. The conversion of albiflorin to benzoic acid through an *in vitro* reaction by rat intestinal bacteria was verified, and the four esterases in *B. breve* and *B. longum* that play an important role in the metabolism of natural compounds were identified through a bioinformatics search. Furthermore, their albiflorin metabolism functions were compared, and the best enzyme reaction conditions were verified. This study revealed the mechanism by which *Bifidobacteria* convert natural compounds containing ester bonds and characterized their specific functional enzymes. TCM herbs closely interact with gut microbiota and affect their composition (Peng et al., 2020). Reciprocally, the gut microbiota also plays essential roles in the conversion of carbohydrates, proteins, lipids, and non-nutritive small chemical compounds from TCM herbs into chemical metabolites that may show beneficial or adverse effects on human health (Feng et al., 2019; Yue et al., 2019). This discovery presents new prospects for the development of new enzymes and probiotics that can be used to enhance the bioconversion of TCM chemicals into bioactive compounds and reveals a new way for *Bifidobacteria* to assist in the treatment of depression and other mental diseases. However, the discovery, development, and functional characterization of *Bifidobacteria* esterases still need more research. A deeper understanding of the mechanism of the drug-gut microbiota interaction is required to guide the research and development of diet or drug interventions targeting the microbiome, which may have the potential to enhance drug efficacy or reduce adverse drug reactions. Understanding the interplay between microbes and medicine could lead to new therapies, or to changes in how existing drugs are prescribed.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

This study was performed in accordance with the recommendations of guidelines for Animal Experimental Center, Animal Care and Welfare Committee, Institute of Materia Medica, CAMS and PUMC. The protocol was approved by the Animal Care and Welfare Committee, Institute of Materia Medica, CAMS PUMC.

AUTHOR CONTRIBUTIONS

YW conceptualized the experiments and analyses and project administration. RP performed the molecular biology study. RP, PH, Z-WZ, and JF performed the bioconversion study. L-BP, HY, and HX analyzed the data. YW, RP, and S-RM contributed to the writing-review and editing. Y-YX and C-XL contributed to the revision of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The project was supported by CAMS Innovation Fund for Medical Sciences (CIFMS) (nos. 2016-I2M-3-011, 2019-I2M-5-020, and 2021-1-I2M-007) the National Natural Science Foundation of China (nos. 81973290 and 82173888), Beijing Natural Sciences Fund Key Projects (no. 7181007), the Beijing Key Laboratory of Non-Clinical Drug Metabolism and PK/PD study (Z141102004414062), and the Youth Science Foundation Project from National Natural Science Foundation of China (no. 81803613).

ACKNOWLEDGMENTS

We would like to thank Shimadzu (China), Co., Ltd. for technological support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.880118/full#supplementary-material

REFERENCES

Arpigny, J. L., and Jaeger, K. E. (1999). Bacterial lipolytic enzymes: classification and properties. *Biochem. J.* 343, 177–183. doi: 10.1042/bj3430177

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473, 174–180. doi: 10.1038/nature09944

Ashida, H., Tanigawa, K., Kiyohara, M., Katoh, T., Katayama, T., and Yamamoto, K. (2018). Bifunctional properties and characterization of a novel sialidase with esterase activity from Bifidobacterium bifidum. *Biosci. Biotechnol. Biochem.* 82, 2030–2039. doi: 10.1080/09168451.2018.1497944

Aura, A. M. (2008). Microbial metabolism of dietary phenolic compounds in the colon. *Phytochem. Rev.* 7, 407–429. doi: 10.1007/s11101-008-9095-3

Bauer, T. L., Buchholz, P. C. F., and Pleiss, J. (2020). The modular structure of α/β-hydrolases. *FEBS J.* 287, 1035–1053. doi: 10.1111/febs.15071

Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., and Schwede, T. (2009). Protein structure homology modeling using SWISS-MODEL workspace. *Nat. Protoc.* 4, 1–13. doi: 10.1038/nprot.2008.197

Bornscheuer, U. T. (2002). Microbial carboxyl esterases: classification, properties and application in biocatalysis. *FEMS Microbiol. Rev.* 26, 73–81. doi: 10.1111/j.1574-6976.2002.tb00599.x

Couteau, D., McCartney, A. L., Gibson, G. R., Williamson, G., and Faulds, C. B. (2001). Isolation and characterization of human colonic bacteria able to hydrolyse chlorogenic acid. *J. Appl. Microbiol.* 90, 873–881. doi: 10.1046/j.1365-2672.2001.01316.x

de Vos, W. M., Tilg, H., Van Hul, M., and Cani, P. D. (2022). Gut microbiome and health: mechanistic insights. *Gut*. doi: 10.1136/gutjnl-2021-326789

Desbonnet, L., Garrett, L., Clarke, G., Kiely, B., Cryan, J. F., and Dinan, T. G. (2010). Effects of the probiotic *Bifidobacterium infantis* in the maternal separation model of depression. *Neuroscience* 170, 1179–1188. doi: 10.1016/j.neuroscience.2010.08.005

Fanning, S., Hall, L. J., Cronin, M., Zomer, A., MacSharry, J., Goulding, D., et al. (2012). Bifidobacterial surface-exopolysaccharide facilitates commensal-host interaction through immune modulation and pathogen protection. *Proc. Natl. Acad. Sci. U. S. A.* 109, 2108–2113. doi: 10.1073/pnas.1115621109

Feng, W., Ao, H., Peng, C., and Yan, D. (2019). Gut microbiota, a new frontier to understand traditional Chinese medicines. *Pharmacol. Res.* 142, 176–191. doi: 10.1016/j.phrs.2019.02.024

Florindo, R. N., Souza, V. P., Manzine, L. R., Camilo, C. M., Marana, S. R., Polikarpov, I., et al. (2018). Structural and biochemical characterization of a gh3 B-glucosidase from the probiotic bacteria bifidobacterium adolescentis. *Biochimie* 148, 107–115. doi: 10.1016/j.biochi.2018.03.007

Fritsch, C., Heinrich, V., Vogel, R. F., and Toelstede, S. (2016a). Phenolic acid degradation potential and growth behavior of lactic acid bacteria in sunflower substrates. *Food Microbiol.* 57, 178–186. doi: 10.1016/j.fm.2016.03.003

Fritsch, C., Jänsch, A., Ehrmann, M. A., Toelstede, S., and Vogel, R. F. (2016b). Characterization of cinnamoyl esterases from different lactobacilli and bifidobacteria. *Curr. Microbiol.* 74, 1–10. doi: 10.1007/s00284-016-1182-x

Guex, N., and Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-Pdb viewer: an environment for comparative protein modeling. *Electrophoresis* 18, 2714–2723. doi: 10.1002/elps.1150181505

Hanske, L., Loh, G., Sczesny, S., Blaut, M., and Braune, A. (2009). The bioavailability of apigenin-7-glucoside is influenced by human intestinal microbiota in rats. *J. Nutr.* 139, 1095–1102. doi: 10.3945/jn.108.102814

He, C. Y., Fu, J., Shou, J. W., Zhao, Z. X., Ren, L., Wang, Y., et al. (2017). In vitro study of the metabolic characteristics of eight isoquinoline alkaloids from natural plants in rat gut microbiota. *Molecules* 22:932. doi: 10.3390/molecules22060932

Huang, X., Su, S., Cui, W., Liu, P., Duan, J. A., Guo, J., et al. (2014). Simultaneous determination of paeoniflorin, albiflorin, ferulic acid, tetrahydropalmatine, protopine, typhaneoside, senkyunolide I in beagle dogs plasma by UPLC–MS/MS and its application to a pharmacokinetic study after Oral Administration of Shaofu Zhuyu Decoction. *J. Chromatogr. B* 962, 75–81. doi: 10.1016/j.jchromb.2014.05.032

Ionescu, C., and Caira, M. R. (2005). *Drug Metabolism: Current Concepts*.

Jochens, H., Hesseler, M., Stiba, K., Padhi, S. K., Kazlauskas, R. J., and Bornscheuer, U. T. (2011). Protein engineering of α/β-hydrolase fold enzymes. *Chembiochem* 12, 1508–1517. doi: 10.1002/cbic.201000771

Kelly, S. M., O'Callaghan, J., Kinsella, M., and Van Sinderen, D. (2018). Characterisation of a hydroxycinnamic acid esterase from the *Bifidobacterium longum* subsp. *longum* taxon. *Front. Microbiol.* 9:2690. doi: 10.3389/fmicb.2018.02690

Kim, N.-R., Jeong, D.-W., Ko, D.-S., and Shim, J.-H. (2017). Characterization of novel thermophilic alpha-glucosidase from *Bifidobacterium longum*. *Int. J. Biol. Macromol.* 99, 594–599. doi: 10.1016/j.ijbiomac.2017.03.009

Kim, G. B., and Lee, B. H. (2010). Genetic analysis of a bile salt hydrolase in *Bifidobacterium animalis* subsp. *lactis* kl612. *J. Appl. Microbiol.* 105, 778–790. doi: 10.1111/j.1365-2672.2008.03825.x

Koppel, N., Rekdal, V. M., and Balskus, E. P. (2017). Chemical transformation of xenobiotics by the human gut microbiota. *Science* 356:eaag2770. doi: 10.1126/science.aag2770

Kroon, P. A., Williamson, G., Fish, N. M., Archer, D. B., and Belshaw, N. J. (2000). A modular esterase from *Penicillium funiculosum* which releases ferulic acid from plant cell walls and binds crystalline cellulose contains a carbohydrate binding module. *Eur. J. Biochem.* 267, 6740–6752. doi: 10.1046/j.1432-1033.2000.01742.x

Lavrijsen, K., Dyck, D. V., Houdt, J. V., Hendrickx, J., and Heykants, J. (1995). Reduction of the prodrug loperamide oxide to its active drug loperamide in the gut of rats, dogs, and humans. *Drug Metab. Dispos.* 23, 354–362. PMID: 7628301

Mosele, J. I., Martín-Peláez, S., Macià, A., Farràs, M., Valls, R. M., Catalán, Ú., et al. (2014). Study of the catabolism of thyme phenols combining in vitro fermentation and human intervention. *J. Agric. Food Chem.* 62, 10954–10961. doi: 10.1021/jf503748y

Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., et al. (2012). Host-gut microbiota metabolic interactions. *Science* 336, 1262–1267. doi: 10.1126/science.1223813

Paolo, B., Divekar, D., and Watson, A. J. M. (2018). The intestinal microbiome influences the response of cancers to pd-1-based immunotherapy: might fecal transplantation become part of cancer therapy? *Gastroenterology* 154, 1845–1847. doi: 10.1053/j.gastro.2018.03.060

Peng, Y., Li, X., Zhang, S. Y., Liu, Z., and Yang, J. (2020). Gut microbiota and chinese medicine syndrome: altered fecal microbiotas in spleen (pi)-deficient patients. *J. Tradit. Chin. Med.* 40, 137–143.

Pollet, R. M., D'Agostino, E. H., Walton, W. G., Xu, Y., Little, M. S., Biernat, K. A., et al. (2017). An atlas of beta-glucuronidases in the human intestinal microbiome. *Structure* 25, 967.e5–977.e5. doi: 10.1016/j.str.2017.05.003

Porro, B., Di Minno, A., Rocca, B., Fiorelli, S., Eligini, S., Turnu, L., et al. (2019). Characterization of aspirin esterase activity in health and disease: in vitro and ex vivo studies. *Biochem. Pharmacol.* 163, 119–127. doi: 10.1016/j.bcp.2019.02.014

Raimondi, S., Anighoro, A., Quartieri, A., Amaretti, A., Tomás-Barberán, F. A., Rastelli, G., et al. (2015). Role of bifidobacteria in the hydrolysis of chlorogenic acid. *Microbiologyopen* 4, 41–52. doi: 10.1002/mbo3.219

Rossi, M., and Amaretti, A. (2010). Probiotic properties of bifidobacteria.

Sanada, K., Nakajima, S., Kurokawa, S., Barceló-Soler, A., Ikuse, D., Hirata, A., et al. (2020). Gut microbiota and major depressive disorder: a systematic review and meta-analysis. *J. Affect. Disord.* 266, 1–13. doi: 10.1016/j.jad.2020.01.102

Savignac, H. M., Kiely, B., Dinan, T. G., and Cryan, J. F. (2014). Bifidobacteria exert strain-specific effects on stress-related behavior and physiology in BALB/c mice. *Neurogastroenterol. Motil.* 26, 1615–1627. doi: 10.1111/nmo.12427

Schackmann, A. V. (1992). Degradation of nitroaromatic compounds by microorganisms. *Bioforum.* 42, 499–507. doi: 10.1007/BF00173912

Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* 31, 3381–3385. doi: 10.1093/nar/gkg520

Sender, R., Fuchs, S., and Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.* 14:e1002533. doi: 10.1371/journal.pbio.1002533

Setchell, K. D., and Clerici, C. (2010). Equol: history, chemistry, and formation. *J. Nutr.* 140, 1355S–1362S. doi: 10.3945/jn.109.119776

Sousa, T., Paterson, R., Moore, V., Carlsson, A., Abrahamsson, B., and Basit, A. W. (2009). The gastrointestinal microbiota as a site for the biotransformation of drugs. *Int. J. Pharm.* 363, 1–25. doi: 10.1016/j.ijpharm.2008.07.009

Spanogiannopoulos, P., Bess, E. N., Carmody, R. N., and Turnbaugh, P. J. (2016). The microbial pharmacists within us: a metagenomic view of xenobiotic metabolism. *Nat. Rev. Microbiol.* 14, 273–287. doi: 10.1038/nrmicro.2016.17

Spanogiannopoulos, P., and Turnbaugh, P. J. (2018). Broad collateral damage of drugs against the gut microbiome. *Nat. Rev. Gastroenterol. Hepatol.* 15, 457–458. doi: 10.1038/s41575-018-0028-3

Turroni, F., Ribbera, A., Foroni, E., van Sinderen, D., and Ventura, M. (2008). Human gut microbiota and bifidobacteria: from composition to functionality. *Antonie Van Leeuwenhoek* 94, 35–50. doi: 10.1007/s10482-008-9232-4

Ventura, M., Turroni, F., Lugli, G. A., and van Sinderen, D. (2014). Bifidobacteria and humans: our special friends, from ecological to genomics perspectives. *J. Sci. Food Agric.* 94, 163–168. doi: 10.1002/jsfa.6356

Whorwell, P. J., Altringer, L., Morel, J., Bond, Y., Charbonneau, D., O'Mahony, L., et al. (2006). Efficacy of an encapsulated probiotic *Bifidobacterium infantis* 35624 in women with irritable bowel syndrome. *Am. J. Gastroenterol.* 101, 1581–1590. doi: 10.1111/j.1572-0241.2006.00734.x

Wilson, I. D., and Nicholson, J. K. (2016). Gut microbiome interactions with drug metabolism, efficacy, and toxicity. *Transl. Res.* 179:204. doi: 10.1016/j.trsl.2016.08.002

Wu, X. M., and Tan, R. X. (2018). Interaction between gut microbiota and ethnomedicine constituents. *Nat. Prod. Rep.* 36, 788–809. doi: 10.1039/c8np00041g

Yu, J. B., Zhao, Z. X., Peng, R., Pan, L. B., Fu, J., Ma, S. R., et al. (2019). Gut microbiota-based pharmacokinetics and the antidepressant mechanism of paeoniflorin. *Front. Pharmacol.* 10:268. doi: 10.3389/fphar.2019.00268

Yue, S. J., Wang, W. X., Chen, Y. Y., Shi, X. Q., and Tang, Y. P. (2019). Gut microbiota modulation with traditional Chinese medicine: a system biology-driven approach. *Pharmacol. Res.* 148:104453. doi: 10.1016/j.phrs.2019.104453

Zanotti, I., Turroni, F., Piemontese, A., Mancabelli, L., Milani, C., Viappiani, A., et al. (2015). Evidence for cholesterol-lowering activity by *Bifidobacterium bifidum* PRL2010 through gut microbiota modulation. *Appl. Microbiol. Biotechnol.* 99, 6813–6829. doi: 10.1007/s00253-015-6564-7

Zhao, Z. X., Fu, J., Ma, S. R., Peng, R., Yu, J. B., Cong, L., et al. (2018). Gut-brain axis metabolic pathway regulates antidepressant efficacy of albiflorin. *Theranostics* 8, 5945–5959. doi: 10.7150/thno.28068

Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R., and Goodman, A. L. (2019a). Separating host and microbiome contributions to drug pharmacokinetics and toxicity. *Science* 363:600. doi: 10.1126/science.aat9931

Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R., and Goodman, A. L. (2019b). Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature* 570, 462–467. doi: 10.1038/s41586-019-1291-3

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

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

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 11 April 2022
doi: 10.3389/fmicb.2022.834335

Succession of the Gut Microbiome in the Tibetan Population of Minjiang River Basin

Jun Li[1†], Lin Sun[1†], Xianlu He[2†], Jing Liu[1], Dan Wang[1], Yuanping Han[3], Baijun Chen[1], Xuemei Li[1], Lingmeng Song[1], Wen Yang[1], Luo Zuo[1], Jingping Sun[1], Ling Qin[1], Feng He[1], Yuanqin Tang[4], Lin Yang[5], Lesiji Kang[5], Yonghua He[6], Xiaofeng Qin[7,8*] and Xiaoan Li[9*]*

[1] Department of Gastroenterology, Clinical Medical College, The First Affiliated Hospital of Chengdu Medical College, Chengdu, China, [2] Department of General Surgery, The Second Affiliated Hospital of Chengdu Medical College, Chengdu, China, [3] College of Life Sciences, Sichuan University, Chengdu, China, [4] Barkam County People's Hospital, Barkam, China, [5] Ngawa Tibetan and Qiang Autonomous Prefecture People's Hospital, Ngawa, China, [6] Hongyuan County People's Hospital, Hongyuan, China, [7] Center of Systems Medicine, Institute of Basic Medical Sciences, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China, [8] Suzhou Institute of Systems Medicine, Suzhou, China, [9] Department of Gastroenterology, Mianyang Central Hospital, Mianyang, China

OPEN ACCESS

Edited by:
Zheng Zhang,
Shandong University, China

Reviewed by:
Bärbel Ulrike Fösel,
Helmholtz Center Munich, Helmholtz Association of German Research Centres (HZ), Germany
Yunus Emre Tunçil,
Ordu University, Turkey

***Correspondence:**
Jun Li
lijun84183967@126.com
Xiaofeng Qin
fqin1@foxmail.com
Xiaoan Li
435445611@qq.com

† These authors have contributed equally to this work

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

Received: *13 December 2021*
Accepted: *15 March 2022*
Published: *11 April 2022*

Citation:
Li J, Sun L, He X, Liu J, Wang D, Han Y, Chen B, Li X, Song L, Yang W, Zuo L, Sun J, Qin L, He F, Tang Y, Yang L, Kang L, He Y, Qin X and Li X (2022) Succession of the Gut Microbiome in the Tibetan Population of Minjiang River Basin. Front. Microbiol. 13:834335. doi: 10.3389/fmicb.2022.834335

Tibetans are one of the oldest ethnic groups in China and South Asia. Based on the analysis of 1,059 Tibetans in the Minjiang River basin at an altitude of 500–4,001 m, we found that the dominant phyla of the Tibetan population were *Bacteroidota* and *Firmicutes,* and the main genera were *Prevotella* and *Bacteroides*, which were mostly in consistent with other nationalities. We further evaluated in total 115 parameters of seven categories, and results showed that altitude was the most important factor affecting the variation in the microbial community. In the process of emigration from high altitudes to the plain, the gut microbial composition of late emigrants was similar to that of plateau aborigines. In addition, regarding immigration from low altitude to high altitude, the microbial community became more similar to that of high altitude population with the increase of immigration time. Changes in these microbes are related to the metabolism, disease incidence and cell functions of the Tibetan population. The results of other two cohorts (AGP and Z208) also showed the impact of altitude on the microbial community. Our study demonstrated that altitude of habitation is an important factor affecting the enterotype of the microflora in the Tibetan population and the study also provided a basis to explore the interaction of impact parameters with gut microbiome for host health and diseases.

Keywords: gut microbiome, Tibet, Minjiang River basin, altitude, migration

INTRODUCTION

Gut microflora is the largest and most complex micro-ecosystem in the human body (Qin et al., 2010; Charbonneau et al., 2016). The microorganisms play important roles in digestion, vitamin synthesis, and immune system functioning. The functions of the intestinal microbiota depend largely on their composition in addition to intestinal factors (Flint et al., 2012; Wang et al., 2014).

Abbreviations: NMDS, non-metric multi-dimensional scaling; AGP, American Gut Project Database; LLD, LifeLines-DEEP Database; PCoA, principal coordinate analysis; CH index, Calinski-Harabasz index; PDW, platelet distribution width; ALT, alanine aminotransferase; KEGG, Kyoto Encyclopedia of Genes and Genomes; CTLA-4, cytotoxic T lymphocyte antigen-4; PD-1, death receptor 1.

Microbial taxa and abundances are in a dynamic balance and are influenced by environmental conditions, host diet, and genetic and epigenetic factors (De Filippo et al., 2010; Gopalakrishnan et al., 2018). Features of ethnicity, including particular ethnic groups, cultural heritage, ancestry, origin myth, history, homeland, language, religion, ritual, cuisine, dressing style, art, and physical appearance, impact the nature of gut microbiome. For instance, in a comparative study involving 173 Caucasian infants and 182 South Asian infants, ethnicity was identified as an independent predictor for intestinal microbial composition of infants, and for example, *Bacillus* and *Lactobacillus* were abundant in South Asian infants, while *Fusobacterium* was abundant in Caucasian infants (Stearns et al., 2017).

Although Human Microbiome Project has produced massive information about intestinal microbes, little is known about the intestinal microbes of the Tibetan population with unique ethnicity. The Tibetan population, one of 56 ethnic minorities in China, mainly resides in the Tibetan autonomous region in Qinghai Province and the western Sichuan Province. The Minjiang River basin, with the largest concentration of Tibetan people in Sichuan, has an average altitude of more than 3,000 m. The genetic background, together with its high altitude, low oxygen concentration, low atmospheric pressure, and high radiation of habitation plus the unique cultural, lifestyle and dietary habits, all contribute to the diverse microbiome of the Tibetan ethnicity. In a similar way, different gut microbiome was examined for the populations in Italy (De Filippo et al., 2017), Africa (Smits et al., 2017), China (Zhang et al., 2015) in association with their unique ethnicities. Likewise, migration from non-Western countries to the United States is associated with a reduction in gut microbiome diversity and function and an increased predisposition to metabolic diseases (Vangay et al., 2018).

With a history of over 25,000 years in the area, indigenous Tibetans have adapted to the plateau inhabitation (Aldenderfer, 2011), which serve as a good model for exploring the effects of the environment on the gut microbiota in the ethnicity conditions. The Tibetan population in the area of Minjiang River and tributaries had emigrated to the Chengdu Plain in recent decades, and this unique feature of changing environmental condition, together with other factors, may impact the gut microbiome, which is the subject of this study. To this end, we collected fecal samples from native Tibetan individuals living at altitudes of 500–4,001 m and examined the ethnicity factors such as environmental conditions, dietary habits, disease statuses, drug use, biochemical indexes, exercise and basic information on metabolic tests. We further assessed the effects of migration from the plateau to the plain and vice versa on the intestinal flora.

MATERIALS AND METHODS

Ethics Committee Approval

This experiment was approved by the Ethics Committee of Chengdu Medical College (No. 2017009). Informed consent was obtained from all participants.

Subject Selection and Sampling

Human subjects of Tibetan ethnicity in Sichuan Province were recruited from the high altitude of 3,300–5,100 m in Ngawa Tibetan Autonomous Region and the low altitude at about 500 m in Chengdu Basin. The residual locations are illustrated in **Figure 1A**, showing Minjiang River and tributaries, including Hongyuan, Barkam, Jinchuan, Heishui, Songpan, Wenchuan, Dujiangyan, and Chengdu, ranging from altitudes of 500–4,001 m. A total of 1251 participants were enrolled (anthropometric information is listed in **Supplementary Table 1**), and fecal samples from 1,059 native Tibetan individuals were collected for analysis. To study the impact of migration, a total of 776 Tibetan individuals were divided into three groups (**Figure 2A**): plateau-born (born and living on the plateau of high altitude, n = 586), basin-born (born and living in the basin at low altitude, n = 20), and plateau-*Trans* (born in the basin or the ancestors were born in the basin and moved to the plateau, n = 170). Among the 170 immigrant participants, some of them returned to and from high-altitude and low-altitude areas for various reasons such as work and study. Therefore, 36 immigrants were included finally for our study to analysis the microbiome succession of reverting immigration from the basin to the plateau. They were further classified as (1) born in the basin but immigrated to the plateau (labeled as Immigrant 1, n = 9), (2) born on the plateau while their parents were from the basin (Immigrant 2, n = 20), and (3) born on the plateau while their grandparents or ancestors were from the basin (Immigrant 3, n = 7). A questionnaire survey was completed by the human subjects regarding basic demographic information (age, sex, birthplace, place of residence, ethnicity, etc.), health status (digestive tract diseases, type 2 diabetes, mental health, genetic diseases, etc.), diet (staple food, dietary intake, drinking habit, consumption of coffee, tea, yogurt, etc.), and exercise (daily physical activity, exercise frequency, etc.). The height, weight, and blood pressure of all participants were recorded. Fecal samples were freshly collected into a stool collection cup without any reagent and transferred within 6 h to a −80°C freezer until further use. Peripheral fasting venous blood was collected for routine blood tests (hemoglobin, erythrocyte, white blood cell, and blood platelet counts) and biochemical tests (liver and kidney functions, blood glucose, and lipid levels). Standardized procedures were applied at all collection sites by the trained personnel. Staff and procedures were regularly checked for quality control throughout the data collection period. This experiment was approved by the Ethics Committee of Chengdu Medical College (No. 2017009). Informed consent was obtained from all participants.

Metadata Collection

Based on questionnaire surveys and blood tests, metadata were obtained, including sociodemographic characteristics, anthropometric characteristics, and information about lifestyle, diet, drug use, diseases, and biochemical parameters. In total, 115 factors were screened and further divided into the following seven categories: basic, environment, drug use, disease, diet, exercise, and biochemical parameters (**Supplementary Table 2**).

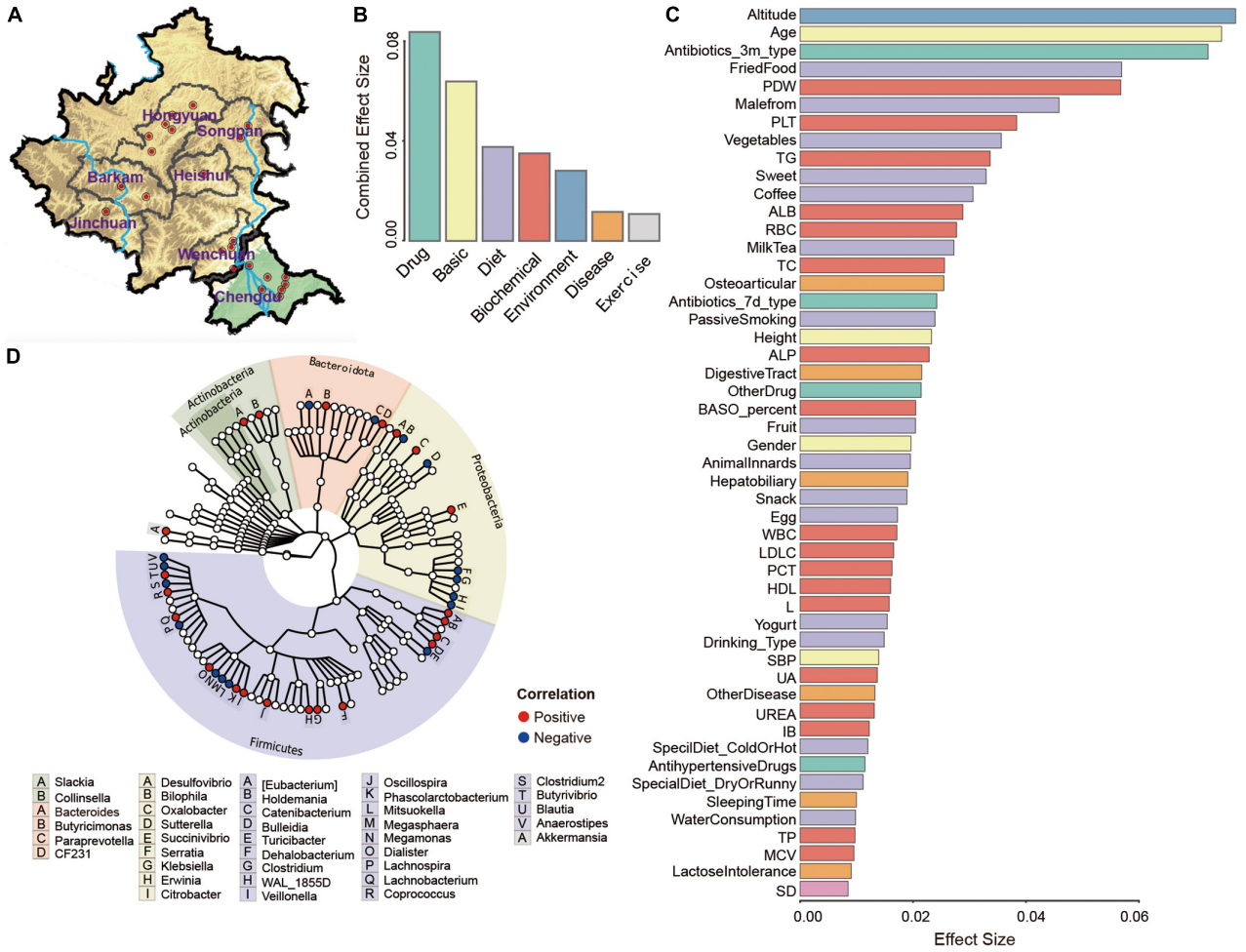

FIGURE 1 | Altitude affects the microbiome of Tibetan ethnicity. **(A)** Overview of sampling regions in Sichuan province of China. Red dots indicated sampling sites at different regions, blue lines represented Dadu River (left) and Min River (right). Effect size of seven categories **(B)** and top 50 factors **(C)** found to significantly influence Tibetan gut microbiota were displayed. Factors were sorted by effect size and colored by seven categories (Basic, Environment, Disease, Diet, Biochemical, Exercise, and Drug). **(D)** Taxonomic tree of 38 genera with significant correlations (adjusted $P < 0.1$) with altitude, as determined using MaAsLin.

Sequencing and Taxonomic Profiling

Microbiota DNA was extracted using the genomic DNA extraction kit (TIANGEN, China). DNA samples were quantified using a Qubit 2.0 Fluorometer. The specific primers 341F (5′-CCTACACGACGCTCTTCCGATCTN-3′) with Barcode and 805R (5′-GACTGGAGTTCCTTGGCACCCGAGAATTCCA-3′) were used to perform PCR amplification of the 16S V3-V4 region (Li et al., 2022). Sequencing was performed with a 2 × 250 paired-end (PE) configuration using the Illumina HiSeq platform. The raw PE reads were merged using FLASH (Version 1.2.7) (Magoč and Salzberg, 2011), and low-quality and polyclonal sequences were filtered using QIIME (Version 1.9.1) (Caporaso et al., 2010). By further comparison with the gold database, chimeric reads were removed using Usearch (Version 8.1.1861) (Edgar et al., 2011). The resulting reads for each sample were clustered into operational taxonomic units (OTUs) at the level of 97% similarity using QIIME (Version 1.9.1). A representative sequence for each OTU was selected, and annotation was performed using QIIME (Version 1.9.1) based on the Greengenes database (Version 13.8) (DeSantis et al., 2006). After random rarefaction of sequences to the minimal number of reads in all samples, microbial composition at each taxonomic level was evaluated using QIIME (Version 1.9.1). The microbial metagenomes were imputed from the 16S rRNA data with Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) based on KEGG Orthology genes and pathways (Langille et al., 2013). The dataset supporting the results of this article has been deposited in the EMBL European Nucleotide Archive (ENA) under BioProject accession code: PRJEB13870.

Multivariate Association

Missing values from the metadata were imputed using the mice package in R (Version 3.3.3) (Zhang, 2015), and collinear variables were detected by a Pearson correlation analysis (Pearson $|r| > 0.8$). Correlations between clinical parameters (categorical

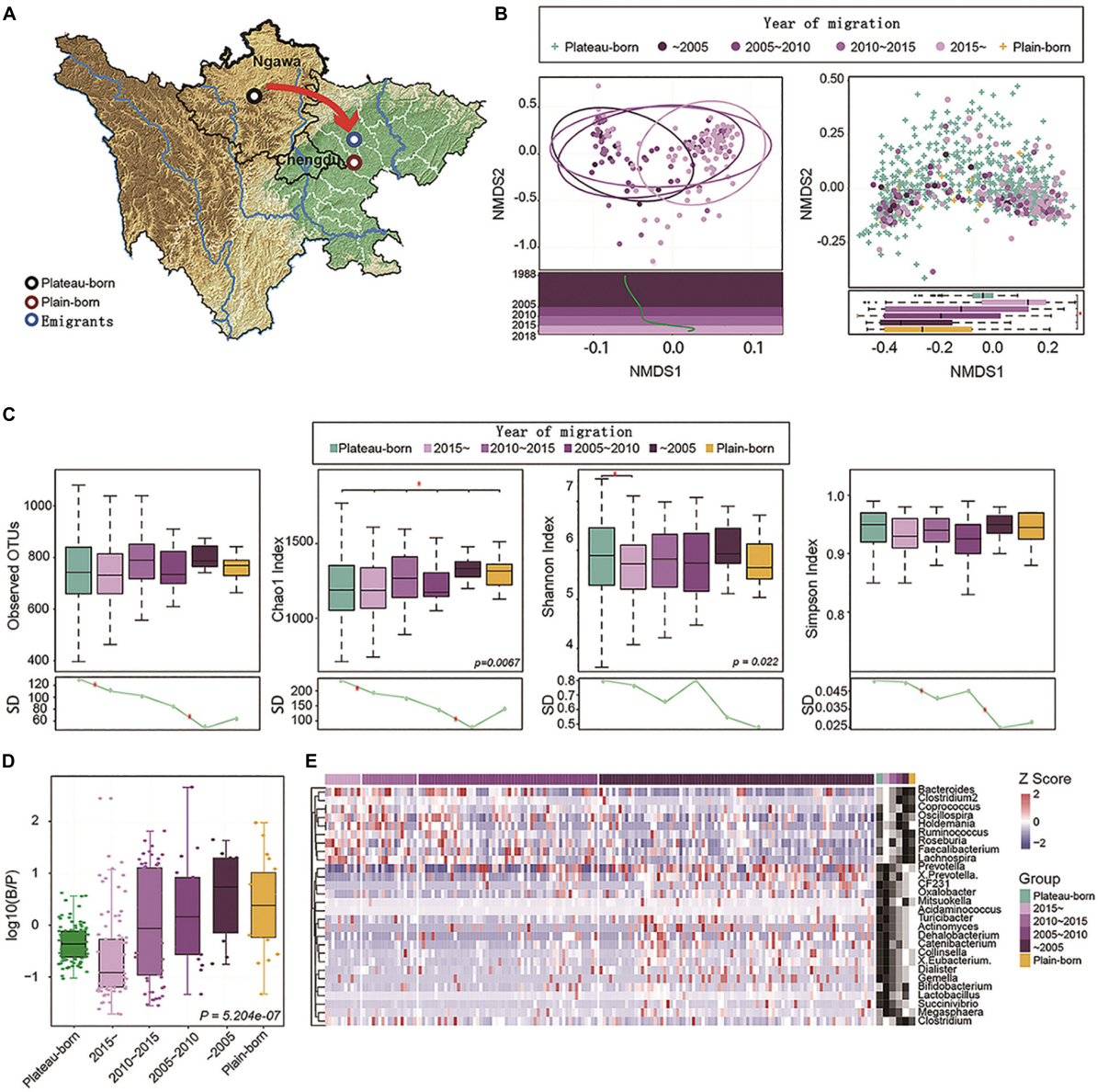

FIGURE 2 | Altitude migration leads to gut microbiome succession in the Tibetan population. **(A)** Plateau-born (n = 586), plain-born (n = 20), and emigrants (Plateau to Plain, n = 170) are marked on the graphics. **(B)** Left: NMDS plot based on the genus profile from 170 plateau to Plain individuals; each dot represents one individual and the color associated with the year of migration. Linear fitting is based on the first NMDS dimension (t = 4.94, P = 1.85e-06). Right: Plateau (plateau-born) and Plains (plain-born) individuals were analyzed by NMDS, and the difference between groups in the first dimension is also visualized by a boxplot (P = 1.589e-11), the NMDS1 values for Plateau-born individuals were the averages of 50 samples by random sampling (100 times). **(C)** Alpha diversity (observed OTUs, Chao1 index, Shannon index, and Simpson index) between Plateau-born, plain-born, and Plateau to Plain individuals, P-values are from the Kruskal–Wallis test. The standard deviation (SD) in different groups revealed convergence in variance from the plateau to the plain, $^*P < 0.05$. **(D)** Distribution of the log-normalized *Bacteroides*-to-*Prevotella* (B/P) ratio between different groups, P-values are based on Kruskal–Wallis test. **(E)** Relative abundance of genera correlated with the year of migration (Spearman's correlation, adjusted $P < 0.05$); the abundance profiles are transformed into Z-scores, where negative a Z-score means a lower abundance than the mean and a positive Z-score represents a higher abundance than the mean. The red arrows indicates the direction of migration.

or numerical) and microbiota community ordination generated by non-metric multidimensional scaling (NMDS) based on Bray–Curtis distances were calculated, as previously described (Falony et al., 2016). For collinear pairs, variables that were weakly correlated with the microbial community were filtered. Envfit was used in the vegan R package to conduct the MANOVA and to estimate linear correlations of categorical and numerical variables of the microbiota. Fifty factors were selected as significant determinants (10,000 permutations; $P < 0.05$; adjusted $P < 0.05$) of the microbial community, and the effect size (r-value) for each factor was determined. The combined effect sizes for the seven categories (basic, environment, disease, diet, biochemical,

exercise, and drug use) were also generated. After Bray–Curtis distance matrixes of sub-metadata and microbiota community data were generated, the correlation between the two distance matrixes was calculated (|*r*|, combined effect size) by the Mantel test in the vegan R package. In addition, clinical variables with significant contributions to core- and unique genus-level community ordination were analyzed. Genera observed in more than 90% of samples were defined as the core microbiota, and genera detected in less than 10% of subjects were defined as unique. The taxonomic tree was visualized using GraPhlAn (Version 1.1.3) (Asnicar et al., 2015). The Kyoto Encyclopedia of Genes and Genomes (KEGG)[1] database was also used. Based on the KEGG pathway analyses, the differential gut microbiomes were annotated and their functions determined.

[1]http://www.genome.jp/kegg

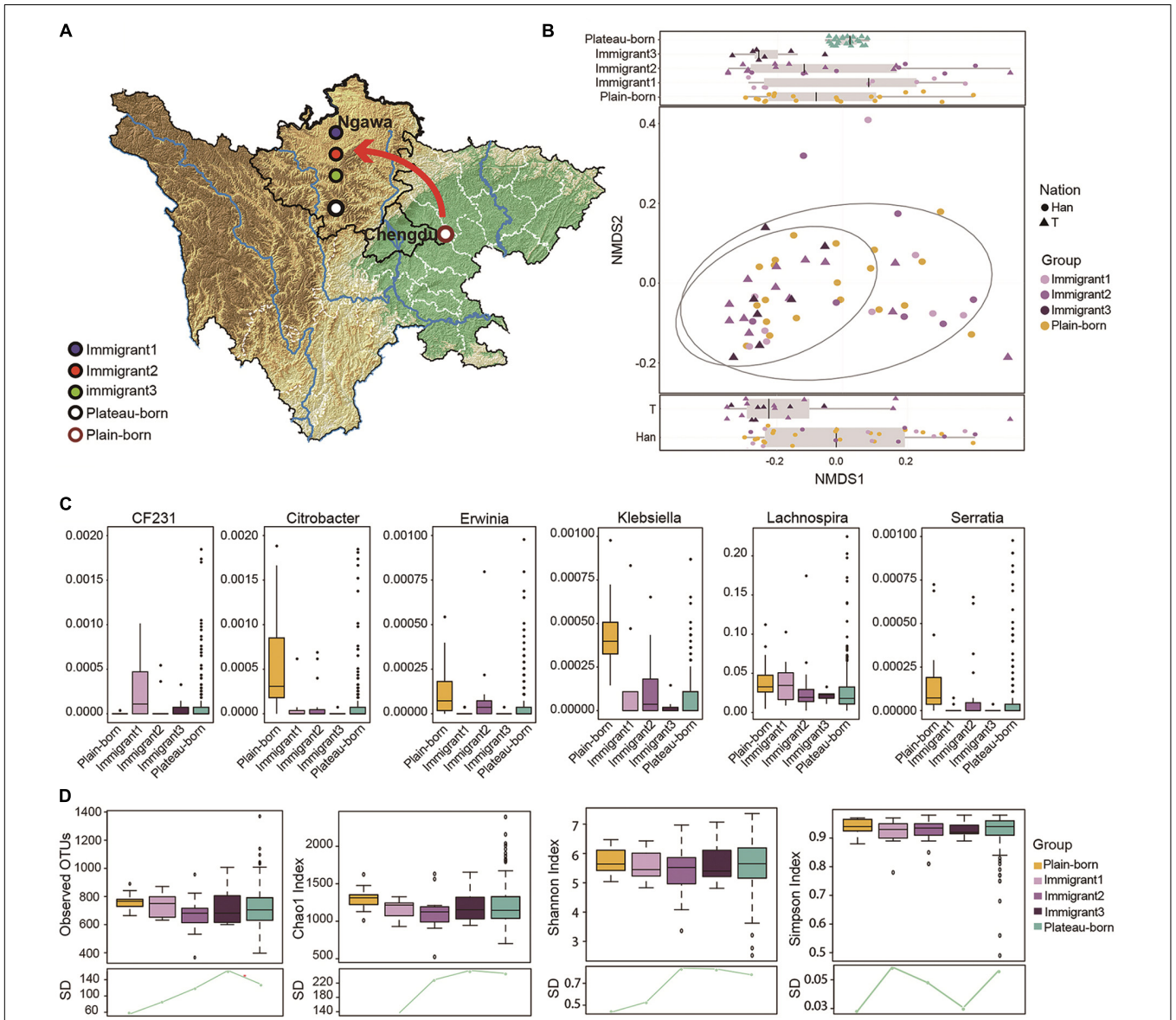

FIGURE 3 | The reverting immigration from the basin to the plateau also promotes microbiome succession. **(A)** Immigrant 1 (born in the plains and moved to the plateau), Immigrant 2 (born in plateau and whose parents born in plain), Immigrant 3 (born in plateau and whose grandparents born in plain), plateau-born (born and living in plateau), plains-born (born and living in plain) and plateau-*Trans* (born in the basin or the ancestors were born in the basin and moved to the plateau) are marked on the graphics. **(B)** NMDS analysis based on the genera profile from plateau-born (n = 586), plain-born (n = 20), and plain to plateau (n = 36) individuals, difference between nations was shown at the bottom (P = 0.009822) and difference between immigrate generations was displayed at the top (P = 0.04825), P-values are from Kruskal–Wallis test, and the NMDS1 values for Plateau-born individuals were the averages of 50 samples at random sampling (20 times). **(C)** Relative abundance of genera significantly correlated with immigrate generation (Spearman's correlation, adjusted $P < 0.05$). **(D)** Alpha diversity between plateau-born, plain-born, and 36 plain to plateau individuals, the standard deviation (SD) in different groups revealed convergence variant from plateau to plain, *$P < 0.05$. The red arrows indicates the direction of migration.

Datasets

Gut microbiota data available in public databases and the literature, including the American Gut Project Database (>), the LifeLines-DEEP Database (LLD), and the flora resources published by Lan et al. (2017) (referred to as S208) and Zang (10) (referred to as S314), were compared with our data set (referred to as Zang). Data were divided according to ethnic groups or regions to analyze the differences in the composition of the intestinal flora and specific characteristics of populations. To validate the impact of altitude on the gut microbial community, two databases including altitude information (AGP and S208) were used to obtain information for 1,244 subjects and 208 Tibetans from six locations in China: Gannan, Gangcha, Tianzhu, Hongyuan, Lhasa, and Nagqu.

Statistics and Reproducibility

Alpha diversity indices (i.e., the observed OTUs, Chao1 index, Shannon index, and Simpson index) were measured using QIIME (Version 1.9.1). To quantify differences (beta diversity) between samples, the phylogeny-based weighted and unweighted UniFrac distances between all pairs of samples were calculated using QIIME. Principal coordinate analysis (PCoA) and NMDS were used to visualize the differences between samples with the ade4 R package. Enterotyping was performed as described previously (Arumugam et al., 2011). Briefly, all samples were analyzed by the partitioning around medoids clustering method based on the Jensen–Shannon distances for genera abundances. The optimal number of clusters was estimated using the Calinski-Harabasz (CH) index (where higher values are better). Only genera detected in at least 10% of samples were included in the analysis. To determine significant associations between clinical variables (categorical or numerical) and genera, a multivariate association analysis was performed using MaAsLin (Morgan et al., 2012). Spearman's correlation coefficients for relationships between continuous variables and microbiota were determined. The differences in alpha diversity indices, genera, and variables between groups were tested by the Wilcoxon rank-sum test or the Kruskal–Wallis test, and *P*-values were calibrated by the Benjamin method. Significance was defined as an adjusted *P*-value of <0.05.

RESULTS

Altitude Affects the Microbiome of Tibetan Ethnicity

Ethnic Tibetan of the main Minjiang River and tributaries at altitudes of 500–4,001 m were recruited and their locations were illustrated in **Figure 1A**. PCoA analysis for the similarity of the groups indicated that *Bacteroidota* and *Firmicutes* were the two most abundant phyla (**Supplementary Figure 1A**). Five core genera were present in Tibetan individuals, *Prevotella* (22.06%), *Bacteroides* (9.08%), *Faecalibacterium* (3.54%), *Lachnospira* (1.43%), and *Ruminococcus* (1.13%), accounting for 32.75% of the total sequences (**Supplementary Figure 1B**). Within the phylum of *Bacteroidota*, the core species mainly belonged to the order *Bacteroidales* and class *Bacteroidia*. The community richness and community diversity of the microbiome in the Tibetan population in the region were mostly consistent with previous reports (**Supplementary Figure 1C**). Based on the CH index, the Tibetan samples were assigned to enterotype 1 (richen in *Prevotella*) and enterotype 2 (richen in *Bacteroides*) (**Supplementary Figure 1D**). To assess whether the flora of the Tibetan population is unique, we compared our dataset Zang with the datasets of LLD (1010 samples), AGP (1,313 samples), S314 (314 samples) and S208 (208 samples). The 3D map of the flora distribution (**Supplementary Figure 2**) indicated that our dataset and S208 of the Tibetan population showed high similarity and were distinguished from the other three datasets, reflecting the specificity of the microflora of the Tibetan population.

To further study the factors affecting the composition of the gut microbes in the Tibetan population, 115 total parameters in seven broad categories were evaluated. Among these categories, drug use (antibiotics, painkillers, etc.) was the main factors affecting the overall flora composition, and basic population parameters took the second place (**Figure 1B**). With respect to the overall flora and the core flora, altitude exerted the strongest effect, followed by age, antibiotic use (within 3 months), fried food, and platelet distribution width (PDW) (**Figure 1C** and **Supplementary Figure 3A**). The unique microbiota was greatly related to liver function determined as plasma levels of alanine aminotransferase (ALT) (**Supplementary Figure 3C**). In terms of the seven categories, drug use was the most important determinant for the core flora, being consistent with previous findings (**Supplementary Figure 3B**; Maier et al., 2018). For unique microbiota, the environment category had the greatest impact (**Supplementary Figure 3D**). Furthermore, 38 genera significantly correlated with altitude were screened using MaAsLin ($P < 0.01$) and we found that *Firmicutes* and *Proteobacteria* were the dominant phyla related to altitude (**Figure 1D**). Taken together, these results indicated that altitude was the most important factor affecting the gut microbiome in Tibetan populations, and our result also uncovered uniqueness of the microflora in individuals living in Tibetan areas.

Altitude Migration Leads to gut Microbiome Succession in the Tibetan Population

As shown, the overall composition of the Tibetan microflora varied across altitudes based on NMDS2 analysis, as indicated by changing abundances of *Megamonas*, *Bacteroides*, *Prevotella*, *Fusobacterium*, and *Lachnospira* with increased altitude (**Supplementary Figure 4A**). In contrast, the abundances of *Coprococcus, Dialister, Succinivibrio, Megasphaera,* and *Prevotella* were enhanced together with the decreased scale of altitude (**Supplementary Figure 4A**). Eight genera were further analyzed for their association with altitude adaptability (**Supplementary Figure 4B**). *Klebsiella* was decreased along with the increased altitude, while *Lachnospira* and *Megamonas* showed good adaptability to high altitudes and maintained relatively high abundances (**Supplementary Figure 4B**). It

should be pointed out that the abundance of *Lachnospira* was increased significantly ($p < 0.05$) at altitudes of 1,000–2,000 m and then decreased with the increased altitude. *Megamonas* showed a higher abundance at 1,000–3,000 m, with a significant ($p < 0.05$) decrease in abundance at altitudes above 3,000 m. A steady increase in the abundance of *Oscillospira* was detected with increased altitude, indicating good adaptability to high altitudes. *Clostridium, Lachnobacterium*, and *Akkermansia* all showed relatively stable abundances, except at altitudes exceeding 3,000 m (**Supplementary Figure 4B**). Taken together, these results show that alpha diversity was positively associated with altitude above 1,000 m (**Supplementary Figure 4C**). Spearman's correlation analyses were used to evaluate relationships between genera and altitude (**Supplementary Figure 5A**). *Clostridium, Oscillospira, WAL_1855D, Succinivibrio*, and *CF231* were positively correlated with altitude, while *Bacteroides*, *Trabulsiella, Serratia, Erwinia*, and *Citrobacter* were negatively correlated with altitude. To eliminate bias due to the uneven distribution of samples at different altitudes, the relative abundances of these genera were acquired from random sampling and transformed into *Z*-scores (**Supplementary Figure 5B**).

Next, a total of 776 Tibetan individuals were divided into three groups (**Figure 2A**): plateau-born, basin-born, and plateau-Trans. An NMDS2 plot was generated based on the genera profile and the time of emigration to the basin, which was then used for a linear fitting analysis. As shown in **Figure 2B**, the gut microbiome composition differed with respect to emigration time. In particular, late emigrants had gut microbial communities that were similar to those of the indigenous population on the plateau, while early emigrants had microbial communities that were similar to those of the native population in the basin. In terms of alpha diversity (**Figure 2C**), earlier emigrants exhibited lower levels of microbial diversity, in line with that of samples from the plain. Standard deviations in prevalence in different groups revealed convergent losses in diversity from the emigration from plateau to the basin, indicating that the time of emigration was correlated with the loss of alpha diversity. The ratio of *Bacteroides* to *Prevotella* (B/P) is an important indicator of the status of bacteria related to weight control and metabolic status (Hjorth et al., 2018). Accordingly, we analyzed the distribution of the log-normalized B/P in different groups (**Figure 2D**). The longer the time since emigration, the higher the B/P ratio and the closer the ratio was to that of the indigenous population on the plains, indicating that the gut microbiome metabolism of the migrating population gradually converges to that of the plain population. Furthermore, Spearman's correlation analysis of the overall distribution of genera showed that the genus type was associated with the time of migration.

Based on abundance profiles transformed into *Z*-scores (**Figure 2E**), species with significantly higher abundances in the earlier emigrants were also more abundant in the population in the basin. In brief, these results showed that changing altitude drives gut microbiome succession in the Tibetan population, and Tibetan migration is further associated with the loss of diversity in the gut microbiome.

The Reverting Immigration From the Basin to the Plateau Also Promotes Microbiome Succession

Based on the subpopulations described in **Figure 3A** we examined the genera profile used to determine the overall microflora structure in each generation of immigrants that reverted migration from the basin to plateau with respect to ethnicity (**Figure 3B**). We focused on six genera in which their abundances were significantly correlated with immigration by the MaAsLin method. As shown, Lachnospira had a high abundance in samples from the basin but decreased across generations. Conversely, Klebsiella had a high abundance only in the basin group (**Figure 3C**). The alpha diversity among the plateau-born, basin-born, and basin-to-plateau immigrant individuals was determined (**Figure 3D**), and the SD in the different groups revealed convergence in variation from the basin to the plateau immigration. In general, the shorter the time since immigration, the smaller the difference in sample diversity, and conversely, the longer the immigration time, the greater the difference in sample diversity (**Figure 3D**).

Altitude Migration Affects the Diversity of the Tibetan Gut Microbiome

To further explore the influence of altitude migration on the gut microbiome of the Tibetan population, we analyzed the change in core flora in different periods from high altitude to low altitude (years of migration), and the core flora changed in different generations from low altitude to high altitude (inverting immigration). Through correlation analysis, the correlation network showed that 15 bacterial communities changed in the process of altitude migration. Among them, Lachnospira, *Bacteroides,* and *Clostridium 2* were negatively correlated with altitude migration, while the changes in *Lactobacillus,* [*Prevotella*], *Dialister, Prevotella, Succinivibrio, Catenibacterium, Collinsella, [Eubacterium], CF231, Slackia, Oxalobacter,* and *Dehalobacterium* were positively correlated with altitude migration (**Figure 4A**).

We further analyzed the influence of 115 parameters of ethnicity on the evolution of the gut microbiome in the process of altitude migration. For people migrating from the plateau to the basin, altitude and fried food significantly contributed to the diversification of the Chao1 index (**Figure 4B**), and alcohol drinking significantly contributed to the taxonomic composition of the gut microbiome (**Figure 4C**). For people migrating from the basin to the plateau, altitude had the greatest impact on the Chao1 index (**Figure 4B**), and altitude significantly contributed to taxonomic composition (**Figure 4C**).

Altitude Migration Affects the Physiological Function and Disease Incidence in the Tibetan Population

To study the influence of altitude migration on the host, we conducted Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis. Predicted pathways based on structural data revealed that metabolism, signal transduction, and transcription

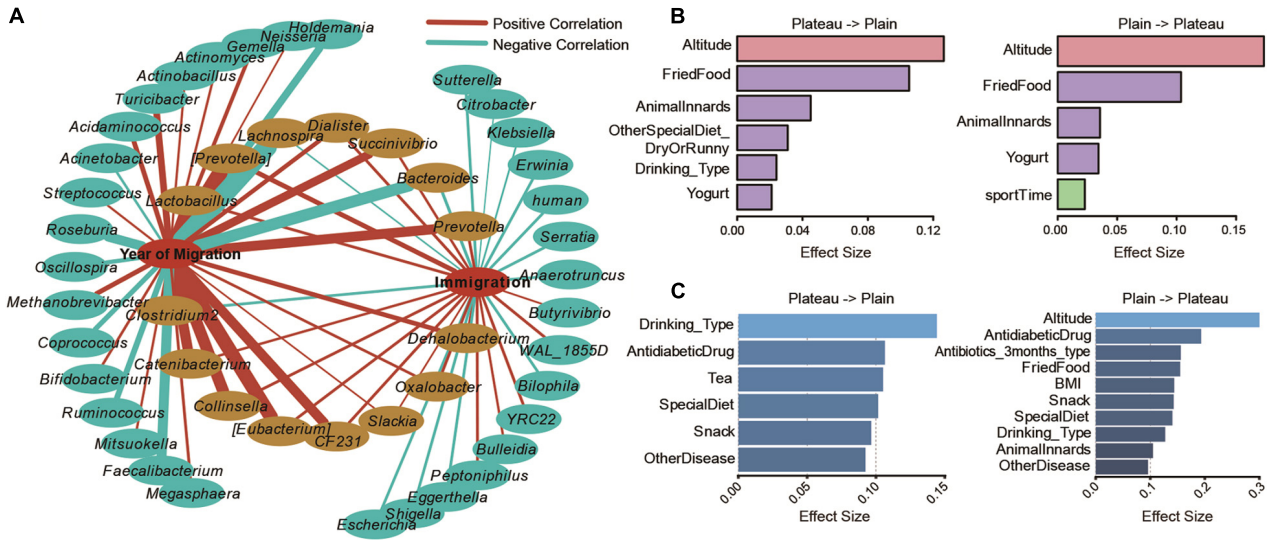

FIGURE 4 | Altitude migration affects the diversity of the Tibetan gut microbiome. **(A)** Correlation network of genera significantly correlated with the year of migration or the generation of immigration (adjusted $P < 0.05$). Factors significantly contributed to diversification of Chao1 index **(B)** and taxonomic composition **(C)** were also displayed.

were positively correlated with the year of migration from the plateau to the plain (**Figure 5A**). Cancer, immune, and digestive systems were negatively correlated with the year of migration from the plateau to the plain (**Figure 5A**). We further analyzed the influence of gut microbial changes on the Tibetan population from the plain to the plateau. KEGG analysis showed that metabolism, signal transduction, transcription, and signaling pathways were negatively correlated with the generations that immigrated from the plain to the plateau (**Figure 5B** and **Supplementary Figure 6**), while cell growth and death, replication and repair and digestive system pathways were positively correlated with the generations that immigrated from the plain to the plateau (**Figure 5B** and **Supplementary Figure 6**). The above analysis showed that with the migration between altitudes, the changes in microorganisms mainly affected the metabolism, signal transduction, transcription, cancer, digestive system, cell growth and death and replication and repair pathways of the Tibetan population.

Construction of the Network of Impact Parameters, Gut Microbiome, and Functions

Gram staining of bacteria is one of the important methods used to distinguish bacterial species, which could guide the diagnosis and treatment of diseases (Boyanova, 2018). Therefore, we analyzed the changes in the abundance of the predicted gram-positive and gram-negative bacteria in the process of altitude migration. Among the people who emigrated from the plateau to the basin, with a shorter emigration time, there were more gram-negative bacteria and less gram-positive bacteria, the pathogenic potential of the bacteria was stronger, and the content of mobile elements was lower (**Figure 6A**). Because the change in altitude mainly leads to a change in environmental oxygen content, we further analyzed the changes in the abundance of aerobic and anaerobic bacteria in the process of altitude migration. A shorter emigration time led to less anaerobic bacteria, more facultative anaerobic bacteria, and better oxygen tolerance (**Supplementary Figure 7**). In addition, for people who immigrated from the plain to the plateau, a longer immigration time led to a lower content of mobile elements, an increased abundance of gram-negative bacteria, and a decreased abundance of gram-positive bacteria (**Supplementary Figure 8**).

Finally, we constructed a network of impact parameter-gut microbiome functions based on altitude and dietary factors, 11 bacterial communities that changed in abundance in the process of altitude migration and changes in the physiological functions associated with the gut microbiome. *Prevotella* was the most closely related to the relevant KEGG pathways, and the factors most related to the genus and metabolic pathways were altitude, along with the consumption of milk, tea, vegetables, fried food, and sweets (**Figure 6B**).

Altitude Determines the Composition of the Gut Microbiome in Other Ethnic Groups

To analyze the effect of altitude on other populations, we compared our data with AGP and S208. We used PCoA plots based on Bray–Curtis dissimilarity at the genus level to depict the overall distribution of the intestinal flora at different altitudes. As shown, altitude clearly distinguished the samples (**Figure 7A**). Similarly, when we analyzed AGP and S208 separately, the high-altitude samples in AGP and S208 could be differentiated in the PCoA plots (**Supplementary Figures 9A,B**). Based on Spearman's correlation coefficients, we identified correlations

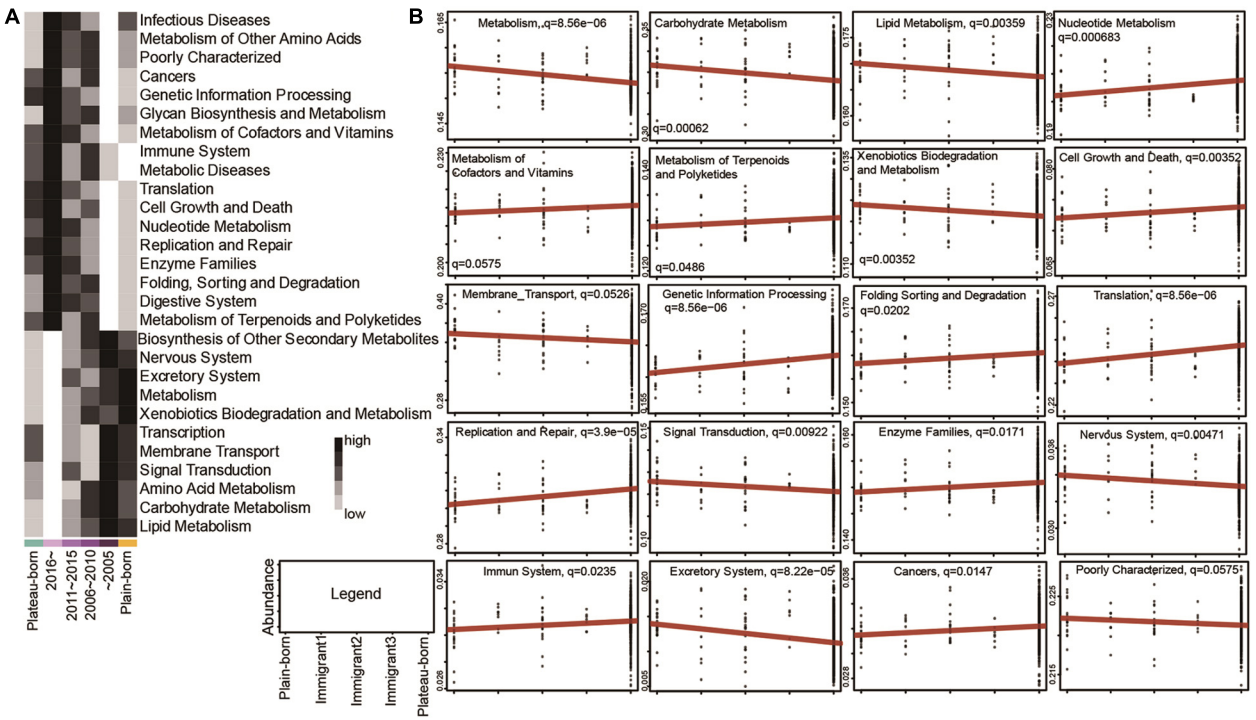

FIGURE 5 | Altitude migration affects the physiological function and disease incidence in the Tibetan population. **(A)** Abundance of predicted KEGG pathways significantly correlated with the year of emigration from plateau to plain (Spearman's correlation, adjusted $P < 0.05$); colors of heatmap mean lower abundance or higher abundance in different groups. **(B)** Relative abundance of 20 predicted KEGG pathways which were significantly correlated with immigration generations from plain to plateau (MaAsLin q-value < 0.1).

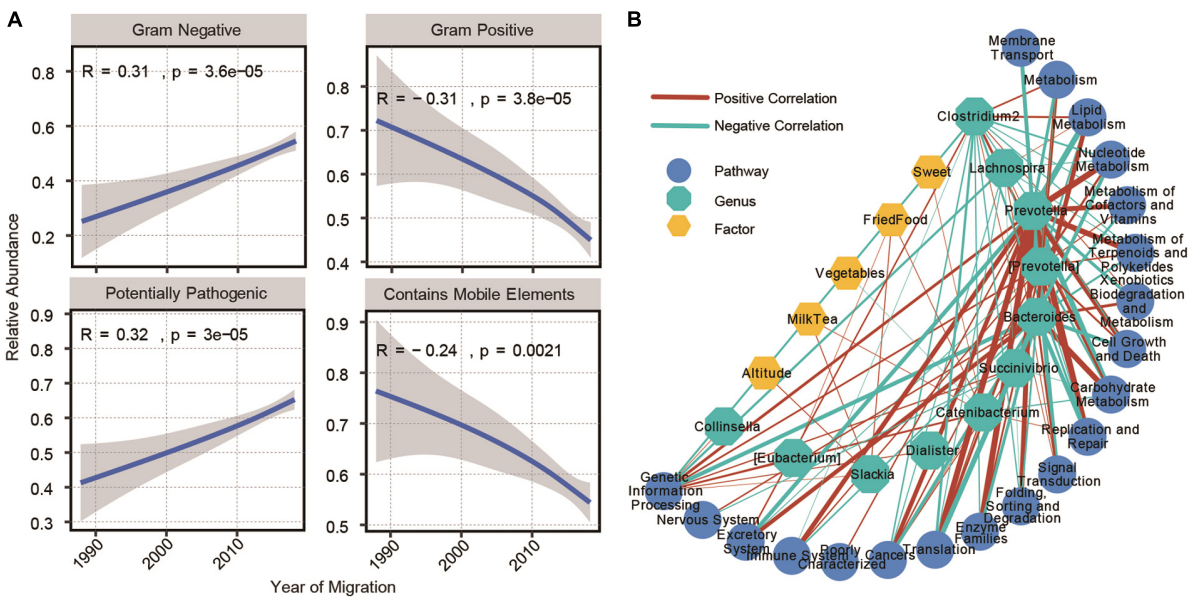

FIGURE 6 | Construction of the network of impact parameters, gut microbiome, and functions. **(A)** Microbiome phenotypes significantly correlated with the year of emigration from plateau to plain (Spearman's correlation, adjusted $P < 0.05$) and immigration generations from plain to plateau (Kendall's correlation, adjusted $P < 0.05$). **(B)** Correlation network between 15 common significantly genera and 20 common predicted KEGG pathways. Their correlations with factors were also displayed (Spearman's correlation, adjusted $P < 0.05$, $|r| > 0.3$).

between the relative abundances of genera and altitude in the S208 (top) and Zang (bottom) datasets (**Figure 7B**). In our dataset (Zang), *Bacteroides* was more abundant in low-altitude samples; however, in the S208 dataset, this taxon was more abundant at high altitudes. In addition, other taxa that were positively correlated with altitude in our dataset were also found in S208. All samples from the S208 dataset were clustered into enterotype 1 (*Prevotella*) and enterotype 2 (*Bacteroides*) by PCoA based on Jensen–Shannon distances (**Supplementary Figure 9C**), consistent with the results for our dataset. In S208, the core genera belonging to *Bacteroidota* were *Prevotella* and *Bacteroides*, and in *Firmicutes*, the core genus was *Faecalibacterium* (**Supplementary Figure 9D**). The core genera identified in S208 were also identified in our dataset, illustrating the similarity in the intestinal flora of Tibetan populations from different regions.

DISCUSSION

This is the first large-scale study of the gut microflora of the Tibetan population in Ngawa. Our results showed that the gut microbiota of the Tibetan population was dominated by phyla of *Bacteroidota* and *Firmicutes*. The flora distribution of the Tibetan population was distinguished from LLD, AGP, and S314 datasets. Bacterial diversity increased with altitude, in agreement with previous results (Kumar et al., 2019). Zhang et al. (2015) studied the gut microbiome of 314 healthy young people from seven ethnic groups in nine provinces of China and identified nine core genera with *Phascolarctobacterium, Roseburia, Bacteroides, Blautia, Faecalibacterium, Clostridium, Subdoligranulum, Ruminococcus,* and *Coprococcus*. In this study, we found unique core genera in the Tibetan population located in Ngawa area as *Prevotella* and *Lachnospira*. It is known that *Prevotella* mainly participates in the metabolism of carbohydrates and plant proteins as well as short-chain fatty acid production (De Filippis et al., 2016). Many *Lachnospira* strains produce butyrate, which plays a crucial role in the maintenance of human gut health (Paggi et al., 2005). These bacteria produce short-chain fatty acids, which can act as anti-inflammatory agents in addition to the source of calories of the body (Paggi et al., 2005). This may explain how Tibetans with a low dietary fiber intake maintain gut health. These bacteria may be related to adaptation to the low-oxygen environment and may be beneficial for Tibetans in the high-altitude areas. The Tibetan microflora can be divided into two enterotypes, as observed in both our dataset (Zang) and dataset S208. A previous study reported that the human gut microbiota can be divided into three enterotypes, *Prevotella*, *Bacteroides,* and *Ruminococcus* (Roager et al., 2014). The gut microbiota of the Tibetan population mainly consisted of *Prevotella* and *Bacteroides*, while *Ruminococcus* accounted for only a small portion of the sequence reads (˜1%). We speculate that this can be explained through the unique diet of Tibetans. *Ruminococcus* is mainly related to the digestion of carbohydrates, such as sugar, starch, and potato (Crost et al., 2018). Tibetans prefer high-protein, high-fat, and low-fiber foods, while their carbohydrate consumption is relatively low.

One study of individuals from six different plateau regions found that location and altitude affected gut microbial composition (Wang et al., 2020). However, the main determinant of the flora was not established in particular small sample size. Based on more than 2,000 individuals from different ethnicities in the same city, one study showed that ethnicity contributes significantly to individual differences in the gut microbiome, independent of metabolic health (Deschasaux et al., 2020). Likewise, based on 7,000 individuals in 14 districts of Guangdong Province, one found that regional factors have a significant impact on the flora (He et al., 2018). Our study also confirmed that altitude among other factors had the greatest effect on the microbial composition. This is also reflected in the analysis of reverse immigrants from the Qinghai-Tibetan Plateau and Han immigrants from the basin to the plateau (Hauenschild et al., 2017). The important influence of altitude on flora was further indicated by two other datasets, AGP and S208. We confirmed that altitude is an important factor affecting the succession of the gut microbiome. We aimed to determine the effects of these changes on the metabolic functions in the Tibetan population. Based on the KEGG pathway analyses, our study found that in the migration process of the Tibetan population, changes in the gut microbiome were most related to metabolism, cancer, the immune system, and the digestive system. Alessia Visconti et al. showed that the gut microbiome was closely related to host systemic metabolism. The metabolic pathway was significantly correlated with 95% of the fecal metabolites, while the microbial species were related to 82% of the fecal metabolites. The carcinogenesis of colorectal cancer was significantly correlated with the gut microbiome. There is some evidence to indicate the association between that pathogenesis of *Fusobacterium nucleatum* and colorectal cancer (Shang and Liu, 2018; Hashemi Goradel et al., 2019). At the same time, the intestinal microflora plays a very important role in the treatment of tumors. In 2015, a landmark paper was published in Science, which showed that the composition of intestinal microbiota can affect the immune checkpoint for cytotoxic T lymphocyte antigen-4 (CTLA-4) inhibitor response and death receptor 1 (PD-1) (Sivan et al., 2015; Vétizou et al., 2015). Recent studies have shown that the gut microbiota is associated with immune system diseases, for example, *P. gingivalis* impacts the development of autoimmune diseases (Ohtsu et al., 2019). Kim et al. (2019) found that the ratio of intestinal P/B decreased in patients with systemic lupus erythematosus, indicating that mucosal immune dysfunction in patients with systemic lupus erythematosus affects the intestinal microbiota. *Bacteroides* found in fragile substances in the human gut play an active regulatory role in the human immune system (Thaiss et al., 2014). These studies highlight the complex interaction between the gut microbiome and host function.

Gram staining of bacteria is one of the important methods used to distinguish bacterial species and can guide the diagnosis and treatment of diseases (Boyanova, 2018). Our research found that a shorter migration time led to an increased abundance of gram-negative bacteria and a decreased abundance of gram-positive bacteria among people migrating from the plateau to the plain. This finding has good clinical guiding significance for the infection of Tibetan people and the choice of antibiotic.

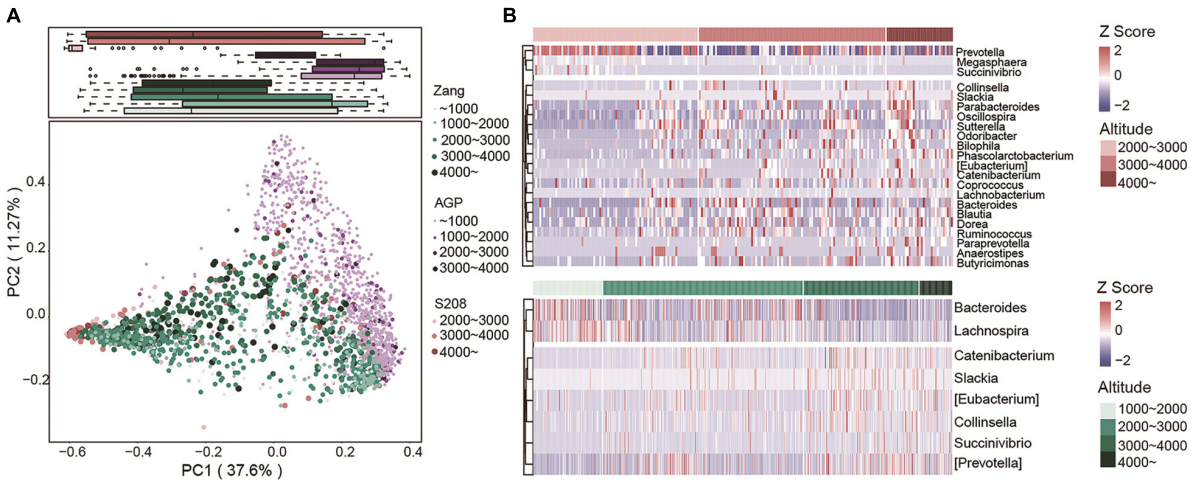

FIGURE 7 | Altitude determines the composition of the gut microbiome in other ethnic groups. **(A)** PCoA plot based on Bray-Curtis dissimilarity at the genus level from three datasets (Zang, $n = 1059$; AGP, $n = 1244$; S208, $n = 208$), the color and size of dots represent different datasets and altitudes. **(B)** Relative abundance of genera significantly correlated with altitude at the dataset of S208 (top) and Zang (bottom) by Spearman's correlation (adjusted $P < 0.05$). The abundance profiles are transformed into Z-score, negative Z-score represents lower abundance than mean and a positive Z-score represents higher abundance than mean.

CONCLUSION

Tibetan ethnicity with its unique lifestyle and customs and high altitude of living environment creates a particular niche for the gut microbiome. Understanding the composition of the gut microbiota of the Tibetan population can provide insight into differences in microbial colonization among regions and ethnic groups as well as the contributions of the unique adaptive lifestyle, customs, and dietary habits to intestinal microecology. We found that altitude was the most important factor affecting the gut microbiome in Tibetan populations and further supported the uniqueness of the microflora in Tibetan areas. The change in altitude promoted the succession of the gut microbial community. AGP and Z208 also showed the impact of altitude on the microbial community. Furthermore, our study provided abundant and unique data to explore the interaction of impact parameter-gut microbiome-host function and disease incidence.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI BioProject – PRJEB13870.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Chengdu Medical College (No. 2017009). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

JuL mainly contributed to the design of this study and wrote the original manuscript. LS and XLH conceived and designed the study. JiL, DW, YPH, BJC, XML, LMS, and WY were responsible for the collection and processing of samples. LZ, JPS, LQ, and FH supervised and administered the project. YQT, LY, LK, and YHH were involved in sample storage and processing. XFQ and XAL were mainly involved in the revision of the manuscript and were responsible for project implementation and quality control. All co-authors contributed substantially to manuscript revisions.

FUNDING

This study was supported by the National Natural Science Foundation of China (Grant Nos. 81702446 and 81802832), Science and Technology Plan Project of Sichuan Province (Grant No. 2018JY0654), and State Key Clinical Specialty Construction Project (Grant Nos. CYFY2017GLPXH001-005 and CYFY201801-06).

ACKNOWLEDGMENTS

We would like to thank Suzhou Geneworks Technology Co., Ltd., for their support on the sequencing and computation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.834335/full#supplementary-material

REFERENCES

Aldenderfer, M. (2011). Peopling the Tibetan plateau: insights from archaeology. *High. Alt. Med. Biol.* 12, 141–147. doi: 10.1089/ham.2010.1094

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., et al. (2011). Enterotypes of the human gut microbiome. *Nature.* 473, 174–180. doi: 10.1038/nature09944

Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ.* 3:e1029. doi: 10.7717/peerj.1029

Boyanova, L. (2018). Direct Gram staining and its various benefits in the diagnosis of bacterial infections. *Postgrad Med.* 130, 105–110. doi: 10.1080/00325481.2018.1398049

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

Charbonneau, M. R., Blanton, L. V., DiGiulio, D. B., Relman, D. A., Lebrilla, C. B., Mills, D. A., et al. (2016). A microbial perspective of human developmental biology. *Nature.* 535, 48–55. doi: 10.1038/nature18845

Crost, E. H., Le Gall, G., Laverde-Gomez, J. A., Mukhopadhya, I., Flint, H. J., and Juge, N. (2018). Mechanistic insights into the cross-feeding of *Ruminococcus* gnavus and *Ruminococcus* bromii on host and dietary carbohydrates. *Front. Microbiol.* 9:2558. doi: 10.3389/fmicb.2018.02558

De Filippis, F., Pellegrini, N., Laghi, L., Gobbetti, M., and Ercolini, D. (2016). Unusual sub-genus associations of faecal Prevotella and *Bacteroides* with specific dietary patterns. *Microbiome.* 4:57. doi: 10.1186/s40168-016-0202-1

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., et al. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14691–14696. doi: 10.1073/pnas.1005963107

De Filippo, C., Di Paola, M., Ramazzotti, M., Albanese, D., Pieraccini, G., Banci, E., et al. (2017). A preliminary investigation in children living in rural and Urban Burkina Faso and Italy. *Front. Microbiol.* 8:1979. doi: 10.3389/fmicb.2017.01979

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

Deschasaux, M., Huybrechts, I., Julia, C., Hercberg, S., Egnell, M., Srour, B., et al. (2020). Association between nutritional profiles of foods underlying Nutri-Score front-of-pack labels and mortality: EPIC cohort study in 10 European countries. *BMJ.* 370:m3173. doi: 10.1136/bmj.m3173

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics.* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381

Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., et al. (2016). Population-level analysis of gut microbiome variation. *Science.* 352, 560–564. doi: 10.1126/science.aad3503

Flint, H. J., Scott, K. P., Louis, P., and Duncan, S. H. (2012). The role of the gut microbiota in nutrition and health. *Nat. Rev. Gastroenterol. Hepatol.* 9, 577–589. doi: 10.1038/nrgastro.2012.156

Gopalakrishnan, V., Helmink, B. A., Spencer, C. N., Reuben, A., and Wargo, J. A. (2018). The influence of the gut microbiome on cancer. *Immun. Cancer Immunother.. Cancer Cell.* 33, 570–580. doi: 10.1016/j.ccell.2018.03.015

Hashemi Goradel, N., Heidarzadeh, S., Jahangiri, S., Farhood, B., Mortezaee, K., Khanlarkhani, N., et al. (2019). *Fusobacterium nucleatum* and colorectal cancer: a mechanistic overview. *J. Cell Physiol.* 234, 2337–2344. doi: 10.1002/jcp.27250

Hauenschild, F., Favre, A., Schnitzler, J., Michalak, I., Freiberg, M., and Muellner-Riehl, A. N. (2017). Spatio-temporal evolution of *Allium* L. in the Qinghai-Tibet-Plateau region: immigration and *in situ* radiation. *Plant Divers.* 39, 167–179. doi: 10.1016/j.pld.2017.05.010

He, Y., Wu, W., Zheng, H. M., Li, P., McDonald, D., Sheng, H. F., et al. (2018). Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* 24, 1532–1535. doi: 10.1038/s41591-018-0164-x

Hjorth, M. F., Roager, H. M., Larsen, T. M., Poulsen, S. K., Licht, T. R., Bahl, M. I., et al. (2018). Pre-treatment microbial *Prevotella*-to-*Bacteroides* ratio, determines body fat loss success during a 6-month randomized controlled diet intervention. *Int. J. Obes. (Lond).* 42, 580–583. doi: 10.1038/ijo.2017.220

Kim, J. W., Kwok, S. K., Choe, J. Y., and Park, S. H. (2019). Recent advances in our understanding of the link between the intestinal microbiota and systemic lupus erythematosus. *Int. J. Mol. Sci.* 20:4871. doi: 10.3390/ijms20194871

Kumar, S., Suyal, D. C., Yadav, A., Shouche, Y., and Goel, R. (2019). Microbial diversity and soil physiochemical characteristic of higher altitude. *PLoS One.* 14:e0213844. doi: 10.1371/journal.pone.0213844

Lan, D., Ji, W., Lin, B., Chen, Y., Huang, C., Xiong, X., et al. (2017). Correlations between gut microbiota community structures of Tibetans and geography. *Sci. Rep.* 7:16982. doi: 10.1038/s41598-017-17194-4

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Li, J., Wang, D., and Sun, J. (2022). Application of fecal microbial transplantation in hepatic encephalopathy after transjugular intrahepatic portosystemic shunt. *Medicine (Baltimore)* 101:e28584. doi: 10.1097/MD.0000000000028584

Magoč, 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

Maier, L., Pruteanu, M., Kuhn, M., Zeller, G., Telzerow, A., Anderson, E. E., et al. (2018). Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature.* 555, 623–628. doi: 10.1038/nature25979

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* 13:R79. doi: 10.1186/gb-2012-13-9-r79

Ohtsu, A., Takeuchi, Y., Katagiri, S., Suda, W., Maekawa, S., Shiba, T., et al. (2019). Influence of *Porphyromonas gingivalis* in gut microbiota of streptozotocin-induced diabetic mice. *Oral Dis.* 25, 868–880. doi: 10.1111/odi.13044

Paggi, R. A., Rodríguez, C., and Fay, J. P. (2005). Growth and pectate-lyase activity of the ruminal bacterium lachnospira multiparus in the presence of short-chain organic acids. *Lett Appl. Microbiol.* 41, 434–439. doi: 10.1111/j.1472-765X.2005.01717.x

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 464, 59–65. doi: 10.1038/nature08821

Roager, H. M., Licht, T. R., Poulsen, S. K., Larsen, T. M., and Bahl, M. I. (2014). Microbial enterotypes, inferred by the prevotella-to-*bacteroides* ratio, remained stable during a 6-month randomized controlled diet intervention with the new nordic diet. *Appl. Environ. Microbiol.* 80, 1142–1149. doi: 10.1128/AEM.03549-13

Shang, F. M., and Liu, H. L. (2018). *Fusobacterium nucleatum* and colorectal cancer: a review. *World J. Gastrointest Oncol.* 10, 71–81. doi: 10.4251/wjgo.v10.i3.71

Sivan, A., Corrales, L., Hubert, N., Williams, J. B., Aquino-Michaels, K., Earley, Z. M., et al. (2015). Commensal *Bifidobacterium* promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science.* 350, 1084–1089. doi: 10.1126/science.aac4255

Smits, S. A., Leach, J., Sonnenburg, E. D., Gonzalez, C. G., Lichtman, J. S., Reid, G., et al. (2017). Seasonal cycling in the gut microbiome of the hadza hunter-gatherers of Tanzania. *Science.* 357, 802–806. doi: 10.1126/science.aan4834

Stearns, J. C., Zulyniak, M. A., de Souza, R. J., Campbell, N. C., Fontes, M., Shaikh, M., et al. (2017). NutriGen alliance. ethnic and diet-related differences in the healthy infant microbiome. *Genome Med.* 9:32. doi: 10.1186/s13073-017-0421-5

Thaiss, C. A., Levy, M., Suez, J., and Elinav, E. (2014). The interplay between the innate immune system and the microbiota. *Curr. Opin. Immunol.* 26, 41–48. doi: 10.1016/j.coi.2013.10.016

Vangay, P., Johnson, A. J., Ward, T. L., Al-Ghalith, G. A., Shields-Cutler, R. R., Hillmann, B. M., et al. (2018). US immigration westernizes the human gut microbiome. *Cell* 175, 962–972. doi: 10.1016/j.cell.2018.10.029

Vétizou, M., Pitt, J. M., Daillère, R., Lepage, P., Waldschmitt, N., Flament, C., et al. (2015). Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science.* 350, 1079–1084. doi: 10.1126/science.aad1329

Wang, H., Zhang, W., Zuo, L., Dong, J., Zhu, W., Li, Y., et al. (2014). Intestinal dysbacteriosis contributes to decreased intestinal mucosal barrier function and increased bacterial translocation. *Lett. Appl. Microbiol.* 58, 384–392. doi: 10.1111/lam.12201

Wang, W., Wang, F., Li, L., Wang, A., Sharshov, K., Druzyaka, A., et al. (2020). Characterization of the gut microbiome of black-necked cranes (*Grus nigricollis*) in six wintering areas in China. *Arch. Microbiol.* 202, 983–993. doi: 10.1007/s00203-019-01802-0

Zhang, J., Guo, Z., Xue, Z., Sun, Z., Zhang, M., Wang, L., et al. (2015). A phylo-functional core of gut microbiota in healthy young Chinese cohorts across lifestyles, geography and ethnicities. *ISME J.* 9, 1979–1990. doi: 10.1038/ismej.2015.11

Zhang, Z. (2015). Missing data exploration: highlighting graphical presentation of missing pattern. *Ann. Transl. Med.* 3:356. doi: 10.3978/j.issn.2305-5839.2015.12.28

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

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

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 22 April 2022
doi: 10.3389/fmicb.2022.836899

Bacterial Community and Fermentation Quality of Ensiling Alfalfa With Commercial Lactic Acid Bacterial Additives

Na Na[1], Moge Qili[1], Nier Wu[1], Lin Sun[1], Haiwen Xu[2], Yi Zhao[1], Xiaobin Wei[3], Yanlin Xue[1] and Ya Tao[3,4]**

[1] Inner Mongolia Key Laboratory of Microbial Ecology of Silage, Inner Mongolia Engineering Research Center of Development and Utilization of Microbial Resources in Silage, Inner Mongolia Academy of Agriculture and Animal Husbandry Science, Hohhot, China, [2] College of Foreign Languages, Inner Mongolia University of Finance and Economics, Hohhot, China, [3] Inner Mongolia Youran Animal Husbandry Co., Ltd., Hohhot, China, [4] Institute of Grassland Research, Chinese Academy of Agricultural Sciences, Hohhot, China

OPEN ACCESS

Edited by:
Zheng Zhang,
Shandong University, China

Reviewed by:
Xianjun Yuan,
Nanjing Agricultural University, China
Kimon Andreas Karatzas,
University of Reading, United Kingdom
Tugay Ayasan,
Osmaniye Korkut Ata University, Turkey

***Correspondence:**
Yanlin Xue
xueyanlin_1979@163.com
Ya Tao
taoya2001@126.com

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

Received: *16 December 2021*
Accepted: *22 February 2022*
Published: *22 April 2022*

Citation:
Na N, Qili M, Wu N, Sun L, Xu H, Zhao Y, Wei X, Xue Y and Tao Y (2022) Bacterial Community and Fermentation Quality of Ensiling Alfalfa With Commercial Lactic Acid Bacterial Additives. Front. Microbiol. 13:836899. doi: 10.3389/fmicb.2022.836899

The aim of this study was to determine the effects of six common commercial lactic acid bacteria (LAB) additives [A1, *Lactobacillus plantarum*, *L. buchneri*, and *Enterococcus faecalis*; A2, *L. plantarum* and *L. casei*; A3, *L. plantarum* and *L. buchneri*; A4, *L. plantarum*, *L. buchneri*, *L. casei*, and *Pediococcus acidilactici*; A5, *L. plantarum* (producing feruloyl esterase); and A6, *L. buchneri*, *P. acidilactici*, β-glucanase, and xylanase] on the bacterial community and fermentation quality of alfalfa silage. Alfalfa was harvested at the squaring stage, wilted in the field for 24 h, and ensiled without any additives (Control) or with A1, A2, A3, A4, A5, or A6. Microbial counts, bacterial community, fermentation parameters, and nutritional composition were determined after ensiling for 90 days. The total abundance of LAB genera on alfalfa pre-ensiling was 0.38% in bacterial community. The abundances of *Lactobacillus*, *Enterococcus*, and *Pediococcus* in the Control silage were 42.18, 40.18, and 8.09% of abundance, respectively. The abundances of *Lactobacillus* in A1-, A2-, A3-, A4-, and A5-treatments were 89.32, 92.93, 92.87, 81.12, and 80.44%, respectively. The abundances of *Pediococcus* and *Lactobacillus* in A6-treatment were 70.14 and 24.86%, respectively. Compared with Control silage, LAB-treated silage had lower pH and less ammonia nitrogen and water-soluble carbohydrates concentrations ($p < 0.05$). Further, the A5- and A6-treatments contained lower neutral detergent fiber, acid detergent fiber, and hemicellulose than other treatments ($p < 0.05$). Overall, LAB genera were presented as minor taxa in alfalfa pre-ensiling and as dominant taxa in alfalfa silage. Adding LAB additives improved the fermentation quality and altered the bacterial community of alfalfa silage. The main bacterial genera in Control silage were *Lactobacillus*, *Enterococcus*, and *Pediococcus*. *Lactobacillus* dominated the bacterial communities of A1-, A2-, A3-, A4-, and A5-treatments, while *Pediococcus* and *Lactobacillus* were dominant bacterial genera in A6-treatment. Inoculating A5 and A6 degraded the fiber in alfalfa silage. It is necessary to ensile alfalfa with LAB inoculants.

Keywords: alfalfa silage, bacterial community, fermentation quality, lactic acid bacterial additives, microbial counts, nutrition composition

INTRODUCTION

Ensiling has become a common and effective method for the long-term preservation of forage for livestock (Sun et al., 2021a). Silage enables anaerobic microbial fermentation to be dominated by lactic acid bacteria (LAB), which utilize water-soluble carbohydrates (WSC) to produce lactic acid (LA), reduce pH, and inhibit harmful microorganisms during ensilage process (Yang et al., 2020). Alfalfa (*Medicago sativa* L.) is a preferred perennial legume forage for livestock producers owing to its high nutritional value, especially its high crude protein (CP) concentration (Hartinger et al., 2019; Besharati et al., 2021; Netthisinghe et al., 2021). However, the second and third cuts of alfalfa in northern China are generally harvested during July and August, a period with an unreliable weather for alfalfa hay processing as it is in the rainy season. As a result, ensiling is the preferable method for conserving alfalfa during this period. Nevertheless, ensiling alfalfa with satisfactory fermentation quality is difficult because of the low dry matter (DM) and WSC concentrations and high buffering capacity (BC) (Sun et al., 2021b). Thus, wilting and applying additives to ensiled alfalfa are necessary to improve the fermentation quality and optimize microbial communities (Gao et al., 2021; Zhang et al., 2021).

Microbial composition, particularly LAB populations, plays a crucial role in the ensiling fermentation quality of silage (Bai et al., 2021). The development of next-generation sequencing technologies has helped to understand the differences in microbial communities and fermentation quality among silages (Wang C. et al., 2021). Previous studies revealed that inoculating LAB additives at ensiling alfalfa promotes bacterial community dynamics (especially *Lactobacillus* dynamics) during the fermentation process (Guo et al., 2018, 2020; Hu et al., 2020; Zhao S. et al., 2021). Other studies have also reported that *Lactobacillus* dominates the bacterial community in terminal alfalfa silage and in the mixing silage of alfalfa and whole-plant corn (Wang et al., 2020; Luo et al., 2021; Wang M. et al., 2021; Yang et al., 2021).

Inoculating LAB at ensiling optimizes the bacterial community and improves the fermentation quality of the terminal silage (Schmidt et al., 2009; Silva et al., 2016; Zhang et al., 2021; Zhao S. et al., 2021). Previous studies reported that the inoculation of ensiling alfalfa with self-screened LAB can promote the succession of *Lactobacillus* during the fermentation process and increase the abundance of *Lactobacillus* in terminal silage with good fermentation quality (Guo et al., 2018; Hu et al., 2020; Yang et al., 2021). Fermentation quality is improved in alfalfa silage treated with functional LAB screened to produce 3-phenyllactic acid (Wu et al., 2020), ferulic acid esterase (Su et al., 2019; Xie et al., 2021), and class IIa bacteriocin (Li et al., 2020). Sun et al. (2021b) revealed that alfalfa silage inoculated with LAB from ensiling material had greater fermentation quality than that inoculated with LAB from other forage sources. However, LAB screening has a low degree of commercialization, and the effect of common commercial LAB additives on fermentation quality and microbial communities of alfalfa silage has rarely been reported.

In the present study, six commercial LAB additives commonly used for ensiling alfalfa silage in northern China were collected. We hypothesized that the application of these additives at ensiling would improve the fermentation quality and optimize the bacterial community of alfalfa silage. The objective of this study was to determine the fermentation quality and bacterial community in alfalfa silage treated with commercial LAB additives.

MATERIALS AND METHODS

Silage Preparation

Alfalfa was grown for 3 years on an experimental farm (40°46.265 N, 111°39.851E) at the Inner Mongolia Academy of Agricultural and Animal Husbandry Science, Hohhot, China, and harvested from four fields as replicates. The second-cut alfalfa was harvested at the squaring stage at 1 p.m. on June 1, 2019, and wilted in the fields for 24 h. The wilted forages from the four fields were separately chopped to 10–20 mm lengths using a chaffcutter (Hongguang Industry and Trade Co., Ltd., Zhejiang, China), thoroughly mixed, and then randomly divided into seven batches for seven treatments. After each additive (5 g) was mixed with distilled water (2,000 ml), the resulting mixture was allowed to rest for 2 h. The seven treatments were as follows: CK (control): 2 ml/kg fresh weight (FW) of distilled water; A1: 2 ml/kg FW of distilled water and 2 g/t FW (recommended amount, RA) of the first additive [*L. plantarum* LP28 ($\geq$1.0 $\times$ 10^{11} CFU/g), *L. buchneri* LBC136 ($\geq$1.0 $\times$ 10^{9} CFU/g), and *Enterococcus faecalis* EF08 ($\geq$1.0 $\times$ 10^{9} CFU/g); Xinlaiwang I-HL for ensiling straw; Xinlaiwang Biotechnology Co., Ltd., Yangzhou, China]; A2, 2 ml/kg FW of distilled water and 2 g/t FW (RA) of the second additive [*L. plantarum* ($\geq$6.0 $\times$ 10^{10} CFU/g) and *Lactobacillus casei* ($\geq$4.0 $\times$ 10^{10} CFU/g); Xinlaiwang I for ensiling alfalfa. Xinlaiwang Biotechnology Co., Ltd., Yangzhou, China]; A3, 2 ml/kg FW of distilled water and 5 g/t FW (RA) of the third additive [*L. plantarum* 550 and 360 ($\geq$1.3 $\times$ 10^{10} CFU/g) and *L. buchneri* 225 ($\geq$7.0 $\times$ 10^{9} CFU/g); Zhuanglemei; Sichuan Gaofuji Biotechnology Co., Ltd., Chengdu, China]; A4, 2 ml/kg FW of distilled water and 1 g/t FW (RA) of the fourth additive [*L. plantarum*, *L. buchneri*, *L. casei*, and *Pediococcus acidilactici* ($\geq$1.0 $\times$ 10^{11} CFU/g); BONSILAGE; Schaumann Agricultural Trading Co., Ltd., Shanghai, China]; A5, 2 ml/kg FW of distilled water and 1 g/t FW (RA) of the fifth additive [*L. plantarum* MF0932189 (producing feruloyl esterase, $\geq$ 1.0 $\times$ 10^{11} CFU/g); QXMG; Gansu Aopujintai Biological Engineering Co., Ltd., Lanzhou, China]; and A6, 2 ml/kg FW of distilled water and 1 g/t FW (RA) of the sixth additive [*L. buchneri* NCIMB 40788 ($\geq$7.5 $\times$ 10^{10} CFU/g), *P. acidilactici* CNCM MA 18/5 M ($\geq$5.0 $\times$ 10^{10} CFU/g), β-glucanase de *Aspergillus niger* MUCL 39199 ($\geq$5,750 IU/g), and xylanase de *Trichoderma longibrachiatum* MUCL 39203 ($\geq$30,000 IU/g), $\geq$ 1.25 $\times$ 10^{11} CFU/g; LaLSiL Dry; Lallemand Biotechnology Co., Ltd., Beijing, China]. After spraying distilled water with or without additives on chopped alfalfa and performing uniform mixing, approximately 500 g of forage was packed into a plastic bag (food grade, 300 mm $\times$ 400 mm;

Qingye, Beijing, China) and sealed using a vacuum sealer (DZ-300; Qingye, Beijing, China). The bags were stored in a dark room for 90 days and then sampled for analysis. After sampling, the alfalfa pre-ensiling and silages were dried in a forced-air oven (BPG-9240A, Shanghai Yiheng Scientific Instrument Co., Ltd., Shanghai, China) at 65°C for 48 h, ground using a mill (FS-6D; Fichi Machinery Equipment Co., Ltd., Shandong, China) with a 1-mm screen, and dried in the same forced-air oven at 105°C until a constant mass was achieved. The dry matter (DM) content of the silages was corrected for the loss of volatiles during drying according to Weissbach and Strubelt (2008).

Microbial Counts and Bacterial Community

The counts of LAB, coliforms, aerobic bacteria, and yeast were determined *via* culture on Man, Rogosa, Sharpe agar, violet red bile agar, nutrient agar, and potato dextrose agar, respectively, in an incubator (LRH-70, Shanghai Yiheng Science Instruments Co., Ltd., Shanghai, China) at 30°C for 72 h (Cai, 1999).

The bacterial communities of alfalfa pre-ensiling and silages were analyzed by Hangzhou Lianchuan Biotechnology Co., Ltd., Hangzhou, China, according to the method described by Sun et al. (2021a). The E.Z.N.A.® Stool DNA Kit (D4015, Omega Inc., Norcross, GA, United States) was used to extract DNA from the bacteria according to the manufacturer's instructions. Polymerase chain reaction (PCR) was carried out to amplify the V3–V4 region of the bacterial rRNA gene with primers 341F (5′-CCTACGGGNGGCWGCAG-3′) and 805R (5′-GACTACHVGGGTATCTAATCC-3′) (Logue et al., 2016), and the following cycling conditions: 98°C for 30 s, followed by 32 cycles of denaturation at 98°C for 10 s, annealing at 54°C for 30 s, and extension at 72°C for 45 s, and a final extension at 72°C for 10 min. The PCR products were purified using AMPure XT beads (Beckman Coulter Genomics, Danvers, MA, United States), quantified using Qubit (Invitrogen, Carlsbad, CA, United States), and then sequenced on an Illumina NovaSeq PE250 platform according to the manufacturer's recommendations. High-quality clean tags were obtained from raw reads *via* quality filtering according to fqtrim (v0.94), and then filtered using Vsearch software (v2.3.4). Bacterial community diversity was calculated using QIIME2. Further, the sequence alignment of species annotation was performed using BLAST; the alignment databases were SILVA and NT-16S. Principal component analysis (PCA) of the bacterial community (at the genus level) of silages and bubble plot of the bacterial community (genus level) of silages were derived using R (version 3.2.1).[1] Sequencing data were submitted to the NCBI Sequence Read Archive database (accession number: PRJNA744283).

Fermentation Quality and Nutrition Composition

Fresh silage (25 g) was homogenized with sterile water (225 ml) using a flap-type sterile homogenizer (JX-05, Shanghai Jingxin Industrial Development Co., Ltd., Shanghai, China) for 100 s and filtered through four layers of cheesecloth to obtain the silage extract (Sun et al., 2021a). The pH of the silage extract was measured using a pH meter (PB-10; Sartorius, Gottingen, Germany). The organic acids [lactic acid (LA), acetic acid (AA), propionic acid, and butyric acid] were assessed using high-performance liquid chromatography (DAD, 210 nm, SPD-20A, Shimadzu Co., Ltd., Kyoto, Japan) and the following conditions: detector, SPD-20A diode array detector (210 nm); column, Shodex RS Pak KC-811 (50°C, Showa Denko K.K., Kawasaki, Japan); and mobile phase, 3 mM $HClO_4$ (1.0 ml/min) (Bai et al., 2021). The concentrations of ammonia nitrogen (NH_3-N) and total nitrogen (TN) were determined using the Kjeldahl method with a Kjeltec autoanalyzer (8400; Foss Co., Ltd., Hillerød, Denmark) (AOAC, 2005). Water-soluble carbohydrates (WSC) were assessed using anthrone-sulfuric acid colorimetry with a spectrophotometer (UV1102II, Shanghai Tianmei Scientific Instrument Co., Ltd., Shanghai, China), according to the method described by McDonald and Henderson (1964). The buffering capacity (BC) was assessed using acid-base titration, as described by Playne and McDonald (1966).

Crude protein (CP) concentration was calculated by multiplying the TN concentration by 6.25. Neutral detergent fiber (NDF) and acid detergent fiber (ADF) were assessed using an Ankom 2000 fiber analyzer (Ankom, Macedon, NY, United States) according to the method described by Van Soest et al. (1991). Hemicellulose concentration was calculated by the NDF concentration minus the ADF concentration. Crude ash was assessed using a muffle roaster (SX-4-10N, Shanghai Jingqi Instrument Co., Ltd., Shanghai, China) at 550°C for 5 h after carbonization.

Statistical Analysis

The differences in microbial counts, sequencing data, alpha diversity, fermentation quality, and nutrition composition among treatments were analyzed with seven treatments and four repetitions using one-factor analysis of variance *via* the general linear model (GLM) procedure of SAS (version 9.1.3; SAS Institute Inc., Cary, NC, United States). The differences were compared using the least significant difference test, and significance was determined at $p \leq 0.05$.

RESULTS

Fermentation Quality and Nutrition Composition

The silage had lower pH and WSC concentration and higher BC content than fresh materials ($p < 0.05$; **Table 1**). The LAB-treatments had lower pH, NH_3-N, and WSC than Control silage. Further, A2- and A6-treatments contained higher NH_3-N than other LAB-treatments ($p < 0.05$). Compared with the Control silage and A1-treatment, the A4-, A5-, and A6-treatments contained lower AA ($p < 0.05$); A4- and A6-treatments had higher LA/AA ($p < 0.05$). The BC was the lowest in A4-treatment and the highest in A5-treatment among LAB-treatments ($p < 0.05$). No propionic and butyric acids were detected in alfalfa silages.

[1] https://www.omicstudio.cn/tool

The fresh material had a lower DM content than silages ($p < 0.05$; **Table 2**). The A4-treatment contained higher DM content than other treatments, with Control silage displaying a lower content than A3-treatment ($p < 0.05$). The NDF, ADF, and hemicellulose concentrations in A5- and A6-treatments were lower than those in other treatments and fresh materials ($p < 0.05$). The A3-tretment had higher NDF than Control silage and A1-treatment, with A1-treatment having a lower NDF than Control silage ($p < 0.05$). The A4-treatment contained the highest ADF, and A3-treatment had the greatest hemicellulose ($p < 0.05$). The crude ash concentration in A1-treatment was lower than that in A2- and A3-treatmetns ($p < 0.05$).

Microbial Counts and Bacterial Community

The Control silage and A2-treatment had greater LAB count than other treatments and fresh materials ($p < 0.05$). Further, A1- and A3-treatments contained higher LAB count than A4-, A5-, and A6-treatments ($p < 0.05$), and the A4-treatment displayed the lowest LAB count ($p < 0.05$; **Table 3**). The aerobic bacterial count in fresh materials was higher than that in the silages ($p < 0.05$), and the aerobic bacterial count in Control silage was higher than that in LAB-treatments ($p < 0.05$). Moreover, the A4- and A6-treatments had lower yeast count than other treatments and fresh materials ($p < 0.05$). Coliforms were detected in fresh materials but not in silages.

A total of 2,672,280 raw reads and 2,453,088 clean reads of the 16S rRNA gene were obtained from the 32 samples (**Table 3**). There were no differences in the number of raw reads among all silages and fresh materials ($p > 0.05$), and the fresh materials had lower clean reads than A1-, A3-, and A5-treatments ($p < 0.05$). More than 83,000 raw reads and 76,000 clean reads were derived for each sample.

The A5-treatment had higher observed species and Chao1 index than A6-treatment ($p < 0.05$; **Table 3**). Fresh materials had a higher Shannon index than silages ($p < 0.05$), and Shannon index in Control silage and A4-treatment were higher than those in the other treatments ($p < 0.05$). The Simpson index for the A3-, A5-, and A6-treatments was higher than that for the A1- and A2-treatments and lower than that for the fresh materials, Control silage, and A4-treatment ($p < 0.05$).

According to PCA, the bacterial communities of Control silage and A6-treatment were cleanly separated from each other and from other treatments (**Figure 1**). However, the A1-, A2-, A3-, A4-, and A5-treatments had aggregated bacterial community. In addition, the fresh materials contained a separate bacterial community from that of the silages.

The most predominant bacterial genus in the A1-, A2-, A3-, A4-, and A5-treatments was *Lactobacillus*, with abundances of 89.32, 92.93, 92.87, 81.12, and 80.44%, respectively (**Figure 2**). The other main genera (>1%) were *Enterococcus*, *Cedecea*, and *Devosia* in A1-treatment; *Enterococcus* in A2- and A3-treatments; *Pediococcus*, *Paracoccus*, *Devosia*, and *Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium* in A4-treatment; and *Pediococcus*, *Enterococcus*, *Paracoccus*, *Devosia*, *Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium*, and *Falsirhodobacter* in A5-treatment. The main bacterial genera in Control silage and A6-treatment were *Lactobacillus* (42.18 and 24.86%, respectively), *Pediococcus* (8.09 and 70.14%, respectively), and *Enterococcus* (40.18 and 1.40%, respectively), followed by *Pantoea*, *Paracoccus*, and *Weissell*a in Control silage (>1%). Further, *Pantoea*, *Enterobacter*, and *Pseudomonas* were the dominant bacterial genera in fresh materials, with abundances of 37.59, 21.20, and 15.74%, respectively (**Figure 2**).

Difference in Bacterial Community Among Silages and Fresh Materials

Compared with fresh materials, silages had higher *Lactobacillus*, *Pediococcus*, and *Enterococcus* ($p < 0.05$) and lower *Pantoea*, *Enterobacter*, *Pseudomonas*, *Cedecea*, and *Rahnella* ($p < 0.05$) (**Figure 3**). Control silage and A6-treatment contained less *Lactobacillus* than other treatments, with Control silage having higher than the A6-treatments ($p < 0.05$). The A6-treatment had higher *Pediococcus* than other treatments, with A1-, A2-, and A3-treatments displaying a lower than A4- and A5-treatments and

TABLE 1 | Fermentation quality, water-soluble carbohydrates (WSC), and buffering capacity (BC) of alfalfa silages ($n = 4$).

Items	FM	CK	A1	A2	A3	A4	A5	A6	SEM	*p*-value
Ph	6.09a	4.70b	4.32c	4.36c	4.39c	4.36c	4.33c	4.38c	0.020	<0.001
LA (g/kg DM)	–	72.2	94.8	90.5	84.8	67.7	63.3	63.3	7.06	0.055
AA (g/kg DM)	–	51.1a	50.4a	40.0ab	30.6ab	20.2b	23.8b	19.8b	6.04	0.002
PA (g/kg DM)	–	ND	ND	ND	ND	ND	ND	ND	–	–
BA (g/kg DM)	–	ND	ND	ND	ND	ND	ND	ND	–	–
LA/AA	–	1.42c	2.07bc	2.52ab	2.78ab	3.38a	2.75ab	3.21a	0.222	<0.001
NH_3-N (g/kg TN)	–	41.3a	23.2c	29.0b	19.2c	21.5c	23.5c	27.4b	1.1	<0.001
WSC (g/kg DM)	46.5a	15.1b	6.85c	6.54c	7.11c	5.79c	2.68c	4.50c	1.18	<0.001
BC (mEq/kg DM)	575d	858ab	842b	835b	838b	813c	867a	838b	5.87	<0.001

Values with different letters indicate significant differences among fresh materials and silages. SEM, standard error of the mean; LA, lactic acid; AA, acetic acid; PA, propionic acid; BA, butyric acid; NH_3-N, ammonia nitrogen; TN, total nitrogen; ND, not detected. FM, fresh materials; CK (Control), 2.00 ml/kg fresh weight (FW) of distilled water; A1, 2 g/t FW of the first additive and 2.00 ml/kg FW of distilled water; A2, 2 g/t FW of the second additive and 2.00 ml/kg FW of distilled water; A3, 5 g/t FW of the third additive and 2.00 ml/kg FW of distilled water; A4, 1 g/t FW of the fourth additive and 2.00 ml/kg FW of distilled water; A5, 1 g/t FW of the fifth additive and 2.00 ml/kg FW of distilled water; A6, 1 g/t FW of the sixth additive and 2.00 ml/kg FW of distilled water.

TABLE 2 | Dry matter (DM, g/kg) and nutrition composition (g/kg DM) of alfalfa silages (n = 4).

Items	FM	CK	A1	A2	A3	A4	A5	A6	SEM	p-value
DM	485d	496c	503bc	506bc	512b	525a	507bc	504bc	2.69	<0.001
Crude protein	196	194	195	193	194	192	194	193	2.33	0.931
Neutral detergent fiber	375ab	359b	340c	373ab	384a	366ab	293d	294d	5.65	<0.001
Acid detergent fiber	230ab	236ab	220b	237ab	220b	245a	198c	201c	5.44	<0.001
Hemicellulose	145b	122c	119c	136bc	164a	121c	94.6d	93.3d	5.01	<0.001
Crude ash	98.2ab	97.0ab	94.5b	99.7a	99.5a	98.3ab	98.3ab	98.3ab	1.06	0.048

Values with different letters indicate significant differences among fresh materials and silages. SEM, standard error of the mean. FM, fresh materials; CK (Control), 2.00 ml/kg fresh weight (FW) of distilled water; A1, 2 g/t FW of the first additive and 2.00 ml/kg FW of distilled water; A2, 2 g/t FW of the second additive and 2.00 ml/kg FW of distilled water; A3, 5 g/t FW of the third additive and 2.00 ml/kg FW of distilled water; A4, 1 g/t FW of the fourth additive and 2.00 ml/kg FW of distilled water; A5, 1 g/t FW of the fifth additive and 2.00 ml/kg FW of distilled water; A6, 1 g/t FW of the sixth additive and 2.00 ml/kg FW of distilled water.

TABLE 3 | Microbial counts, sequencing data, and alpha diversity of bacteria in alfalfa silages (n = 4).

Items		FM	CK	A1	A2	A3	A4	A5	A6	SEM	p-value
Microbial counts (log CFU/g FW)	Lactic acid bacteria	5.59cd	7.29a	6.45b	7.09a	6.00bc	4.45f	5.36de	4.96e	0.157	<0.001
	Coliforms	5.31a	–	–	–	–	–	–	–	0.074	<0.001
	Aerobic bacteria	6.47a	5.55b	4.70c	4.55c	4.82c	4.24c	4.43c	4.77c	0.143	<0.001
	Yeasts	7.40a	7.25a	6.90a	7.49a	7.06a	4.91b	7.00a	5.19b	0.179	<0.001
Sequencing data	Raw reads	85,257	82,306	83,868	82,358	84,706	83,348	83,852	82,378	884	0.181
	Clean reads	71,561b	739,95ab	79,610a	766,55ab	78,702a	76,947ab	78,683a	77,120ab	1,384	0.007
Alpha diversity	Observed species	250ab	179ab	190ab	159ab	160ab	245ab	257a	147b	23.2	0.007
	Chao1	250ab	179ab	191ab	159ab	161ab	246ab	257a	147b	23.3	0.007
	Shannon	4.97a	3.34b	1.32ef	1.13f	1.86de	3.51b	2.56c	2.31cd	0.196	<0.001
	Simpson	0.904a	0.791a	0.291c	0.249c	0.534b	0.846a	0.559b	0.629b	0.035	<0.001
	Goods coverage	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–	–

Values with different letters indicate significant differences among fresh materials and silages. SEM, standard error of the mean. FM, fresh materials; CK (Control), 2.00 ml/kg fresh weight (FW) of distilled water; A1, 2 g/t FW of the first additive and 2.00 ml/kg FW of distilled water; A2, 2 g/t FW of the second additive and 2.00 ml/kg FW of distilled water; A3, 5 g/t FW of the third additive and 2.00 ml/kg FW of distilled water; A4, 1 g/t FW of the fourth additive and 2.00 ml/kg FW of distilled water; A5, 1 g/t FW of the fifth additive and 2.00 ml/kg FW of distilled water; A6, 1 g/t FW of the sixth additive and 2.00 ml/kg FW of distilled water.

Control silage ($p < 0.05$). Control silage had higher *Enterococcus* than other treatments ($p < 0.05$).

DISCUSSION

Characteristics of Alfalfa Pre-ensiling

A previous study reported that ensiling alfalfa with satisfactory fermentation quality is difficult because of low LAB count, less WSC concentration, and higher BC in fresh forage (Sun et al., 2021b). In the present study, the epiphytic LAB count (5.59 log CFU/g FW, **Table 3**) in alfalfa pre-ensiling met the criteria (10^5 log CFU/g FW) required for adequate fermentation (McDonald et al., 1991). However, the alfalfa pre-ensiling contained less WSC and moisture contents (46.5 g/kg DM and 515 g/kg) and higher BC (575 mEq/kg DM) (**Tables 1**, **2**), resulting in higher pH (4.70), more NH_3-N (41.3 g/kg), and lower LA/AA (1.42) in Control silage than those in LAB-treatments (**Table 1**). These suggest that it is necessary to ensile alfalfa with LAB additive for good fermentation quality. The LAB genera (*Weissella*, *Lactobacillus*, *Enterococcus*, and *Pediococcus*) were detected in alfalfa pre-ensiling with total abundance of 0.38% (**Figure 2**). Other studies found that the LAB genera have total abundance of less than 1.0% in alfalfa pre-ensiling (Hu et al., 2020; Zhao S. et al., 2021) and fresh whole-plant corn (Xu et al., 2019; Guan et al., 2020). The LAB genera were demonstrated to be generally presented as minor taxa in forage pre-ensiling. The main bacterial genera in alfalfa pre-ensiling were *Pantoea* (37.59%), *Enterobacter* (21.20%), and *Pseudomonas* (15.74%) (**Figure 2**). The findings agreed with those reported by Zhao S. et al. (2021) for high-moisture alfalfa [*Enterobacter* (33.93%), *Pseudomonas* (16.67%), and *Pantoea* (7.09%)]. However, other studies reported that the main bacterial genera (>10% of abundance) in alfalfa pre-ensiling were *Pseudomonas*, *Exiguobacterium*, and *Massilla* (Yang et al., 2020), Sphingobium (Hu et al., 2020), *Xanthomonas* and *Cyanobacteria* (Guo et al., 2020), and *Exiguobacterium* (Wang et al., 2019). The different bacterial communities in alfalfa pre-ensiling among those studies might be due to the differences in geographical locations (Wang C. et al., 2021).

Fermentation Quality and Nutrition Composition of Silage

Ensiling alfalfa with LAB inoculants improves fermentation quality, as demonstrated by the increased LA content and decreased pH and NH_3-N (Guo et al., 2020; Hu et al., 2020). In the present study, LAB inoculation at ensiling decreased pH and NH_3-N in alfalfa silage. However, no difference in LA content was detected among all treatments, and the AA content in the Control

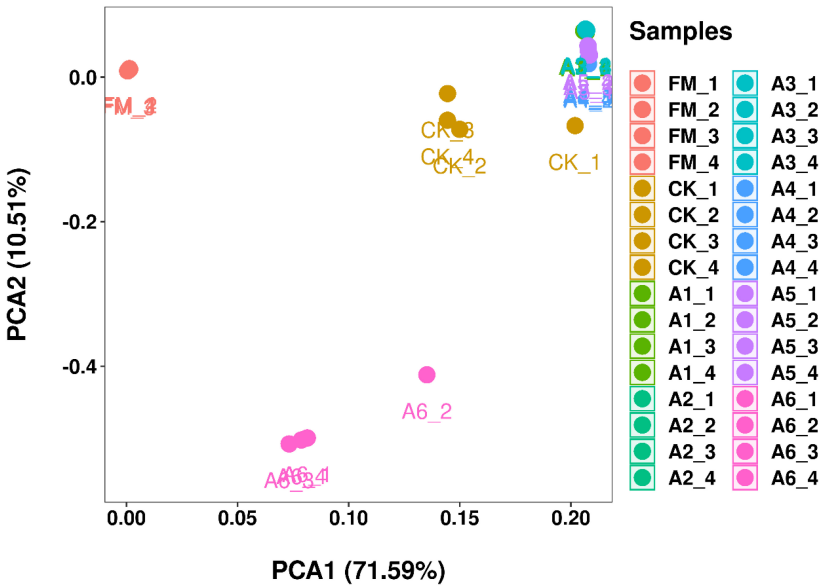

FIGURE 1 | Principal component analysis (PCA) of the bacterial communities in silage and fresh materials (n = 4). FM, fresh materials; CK (Control), 2.00 ml/kg fresh weight (FW) of distilled water; A1, 2 g/t FW of the first additive and 2.00 ml/kg FW of distilled water; A2, 2 g/t FW of the second additive and 2.00 ml/kg FW of distilled water; A3, 5 g/t FW of the third additive and 2.00 ml/kg FW of distilled water; A4, 1 g/t FW of the fourth additive and 2.00 ml/kg FW of distilled water; A5, 1 g/t FW of the fifth additive and 2.00 ml/kg FW of distilled water; A6, 1 g/t FW of the sixth additive and 2.00 ml/kg FW of distilled water.

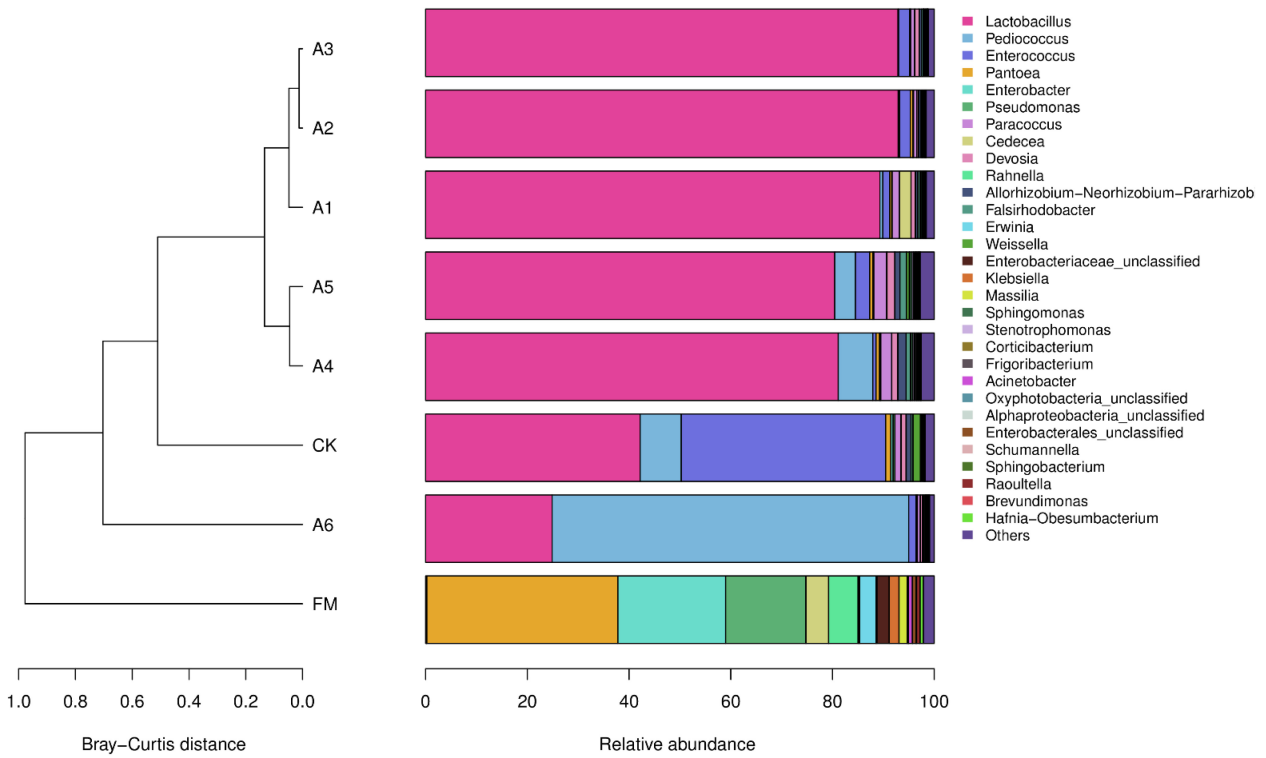

FIGURE 2 | Relative abundance of the bacterial community (genus level) in silages and fresh materials (FM) (n = 4). FM, fresh materials; CK (Control), 2.00 ml/kg fresh weight (FW) of distilled water; A1, 2 g/t FW of the first additive and 2.00 ml/kg FW of distilled water; A2, 2 g/t FW of the second additive and 2.00 ml/kg FW of distilled water; A3, 5 g/t FW of the third additive and 2.00 ml/kg FW of distilled water; A4, 1 g/t FW of the fourth additive and 2.00 ml/kg FW of distilled water; A5, 1 g/t FW of the fifth additive and 2.00 ml/kg FW of distilled water; A6, 1 g/t FW of the sixth additive and 2.00 ml/kg FW of distilled water.

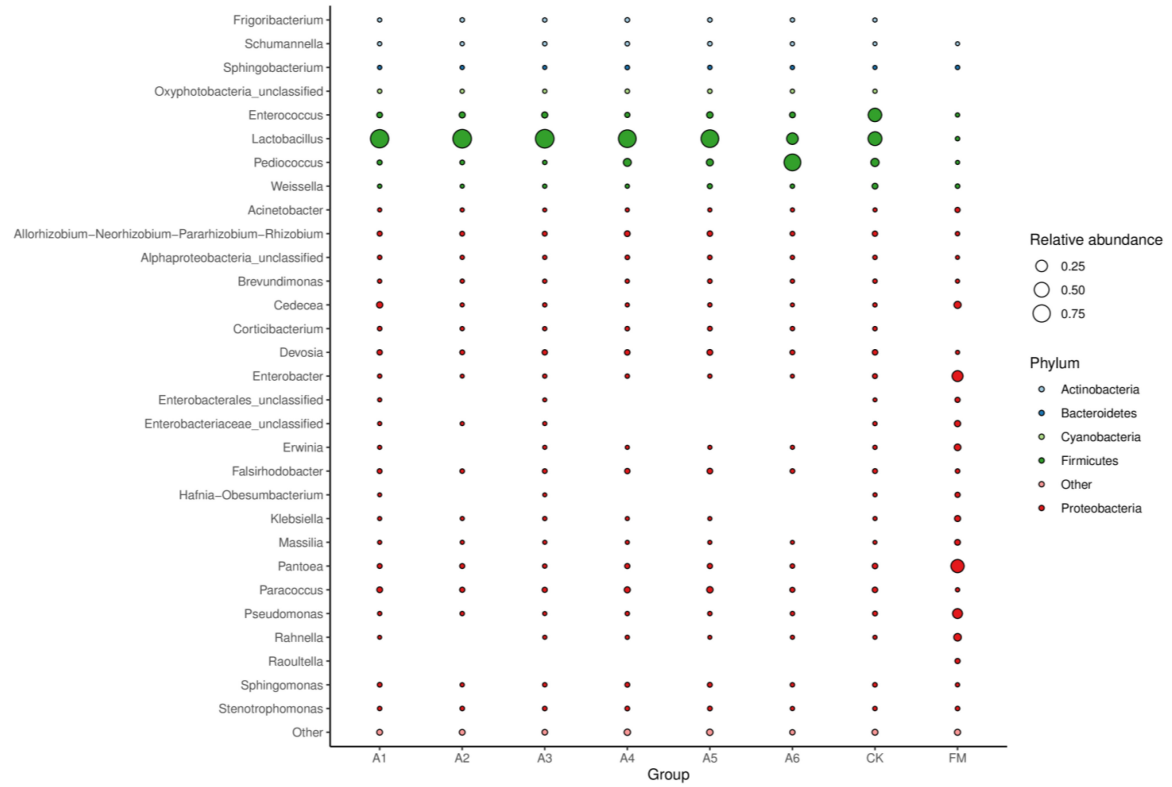

FIGURE 3 | Bubble plot of the bacterial community (genus level) among silages and fresh material ($n = 4$, $p < 0.05$). FM, fresh materials; CK (Control), 2.00 ml/kg fresh weight (FW) of distilled water; A1, 2 g/t FW of the first additive and 2.00 ml/kg FW of distilled water; A2, 2 g/t FW of the second additive and 2.00 ml/kg FW of distilled water; A3, 5 g/t FW of the third additive and 2.00 ml/kg FW of distilled water; A4, 1 g/t FW of the fourth additive and 2.00 ml/kg FW of distilled water; A5, 1 g/t FW of the fifth additive and 2.00 ml/kg FW of distilled water; A6, 1 g/t FW of the sixth additive and 2.00 ml/kg FW of distilled water.

silage and A1-treatment were higher than that in A4-, A5-, and A6-treatments. In addition, Control silage contained more WSC than LAB-treatments (15.1 vs. 2.86–7.11 g/kg DM); however, BC did not differ between Control silage and LAB-treatments (except A4) (**Table 1**). *Lactobacillus*, as the principal component of the additives used in the present study, was negatively correlated with WSC content in alfalfa silage (**Supplementary Figure 1**). The results suggest that other fermentation products (valeric acid, caproic acid, succinic acid, citric acid, ethanol, propanol, and 1,2-propandiol) might be generated during the ensiling process in LAB-treatments, and inoculating LAB at ensiling increased the utilization of WSC in silage during the fermentation process. This phenomenon was also observed in alfalfa silage (Xie et al., 2021), whole-plant corn silage (Jiang et al., 2020), and whole-plant sweet sorghum silage (Diepersloot et al., 2021). Inoculating heterofermentative LAB at ensiling reduces LA/AA in terminal silage by decreasing LA and increasing AA (Kung et al., 2018). The additives (except A5) used in the present study contained heterofermentative LAB (A1, *L. buchneri*; A2, *L. casei*; A3, *L. buchneri*; A4, *L. buchneri*, *L. casei*; A6, *L. buchneri*). However, A5-treatment had no difference in LA relative to other treatments; AA relative to A2-, A3-, A4-, and A6-treatments; and LA/AA relative to other LAB-treatments (**Table 1**). The finding might be due to homofermentative LAB dominating the fermentation process, as reflected by the total abundance of *L. plantarum*, *E. mundtii*, *E. faecium*, and *P. acidilactici* in Control silage and LAB-treatments (58.50, 85.50, 88.76, 73.21, 64.16, 74.42, and 77.28%, respectively) (**Supplementary Table 1**). Moreover, Guo et al. (2020) reported the undifferenced LA and AA concentrations between alfalfa silages treated with homo- and hetero-fermentative LAB.

During fermentation, proteolysis in silage is inevitable, owing to the presence of plant and microbial proteases (Thomas et al., 1980; Hassanat et al., 2007). The NH_3-N, as part of the non-protein, shows the extent of silage preservation during the ensiling process, owing to its low utilization in the rumen (Xue et al., 2017; Yin et al., 2017). In the present study, NH_3-N (41.3 g/kg TN) in Control silage was lower than the suggested concentrations in legume silage (< 120 g/kg TN) (Kung et al., 2018), indicating that the Control silage was well preserved. This finding might be because the higher DM content (480 g/kg) and the ideal anaerobic environment during the ensilage process cause a decrease in the activity of undesired microorganisms during ensiling. Propionic acid, butyric acid, coliforms, and *Clostridia* were not detected in any of the silages (**Table 1** and **Figure 2**). Furthermore, compared with Control silage, the LAB-treatments displayed lower pH (4.70 vs. 4.32– to 4.39), NH_3-N (41.3 vs. 19.2– to 29 g/kg TN), *Enterobacteriaceae* (2.43% vs.

0.1–% to 0.43%), and potentially pathogenic bacteria (6.50% vs. 1.05–% to 1.86%, expect A1) (**Table 1** and **Supplementary Figures 2**, **3**). This finding indicated that the fermentation products of the LAB additives used in the present study contributed more to the preservation of silage and the inhibition of undesired microorganisms during fermentation under low pH, low moisture, and ideal anaerobic conditions. Previous studies reported that LAB inoculation at ensiling decreased NH_3-N content in alfalfa silage (Hu et al., 2020; Sun et al., 2021b).

In the present study, compared with Control silage, LAB-treatments had higher DM content, although the difference did not reach a significant level among Control silage and the A1-, A2-, A3-, and A6-treatments. Moreover, there were no differences in CP concentration among the silages (**Table 2**). Previous studies revealed that inoculating LAB at ensiling alfalfa increases the contents of DM and CP and improves the fermentation quality of terminal silage (Li et al., 2020; Wu et al., 2020). The results suggest that satisfactory fermentation quality contributes to increasing the DM content and preserving the CP of alfalfa silage. The concentrations of NDF, ADF, and hemicellulose in the A5- and A6-treatments were lower than those in the other treatments (**Table 2**), which might be related to the function of *L. plantarum* (producing feruloyl esterase) in A5 and the composition of A6 (β-glucanase and xylanase). Feruloyl esterase can promote cell wall degradation, especially in collaboration with cellulase and hemicellulose, by cleaving the ester or ether linkages between ferulic acid and sugars (Dilokpimol et al., 2016; Duan et al., 2021). Su et al. (2019) reported that inoculating feruloyl esterase-producing *Lactobacillus fermentum* at ensiling alfalfa decreased the NDF and ADF concentrations in terminal silage. β-glucanase, as one type of cellulase, can cleave glycosidic bonds in the amorphous regions of cellulose polymers (Takizawa et al., 2020). Moreover, xylanase contributes to the degradation of hemicellulose (Paës et al., 2012; Vucinic et al., 2021). Collectively, these findings indicate that A5 and A6, as additives, can degrade the cell wall during fermentation in alfalfa silage.

Microbial Counts and Bacterial Community of Silage

Inoculating LAB at ensiling can increase LAB count in the terminal silage (Hu et al., 2020; Zhang et al., 2020). However, in the present study, the Control silage and A2-treatment displayed more LAB counts than the other treatments (**Table 3**). Moreover, the Control silage had the highest pH and aerobic bacterial count (**Tables 1**, **3**). These results suggest that the LAB in the A2-treatment might have better resistance to less moisture and a low pH environment (494 g/kg and 4.26, respectively), and the microorganisms in the Control silage had greater activity under less moisture and weakly acidic conditions (504 g/kg and 4.70, respectively). The A4- and A6-treatments contained lower yeast counts than the other treatments (**Table 3**), which might be related to the presence of *P. acidilactici* in A4 and A6 as additives used in the present study. Previous studies reported that *P. acidilactici* inhibits effects on other microorganisms by producing antimicrobial bacteriocins (Kaya and Simsek, 2020; Surachat et al., 2021).

In the present study, most bacteria were detected in all samples the goods' coverages reached approximately 1 (**Table 3**). The bacterial diversity of the silages was lower than that of the alfalfa pre-ensiling (**Table 3**). Furthermore, the material had a clearly separated bacterial community from the silages (**Figure 1**); similar results were reported by Zheng et al. (2017) and Zhao S. et al. (2021). This finding might be due to the large increasing abundance of LAB genera as the main bacterial taxa in silages (87.90%–95.30%) (**Figures 2**, **3**). The Shannon and Simpson indexes for A4-treatment were higher than those for other LAB-treatments and did not differ from those of Control silage. Moreover, A1- and A2-treatments had lower Shannon and Simpson indexes than the other treatments (**Table 3**). These results suggest that the bacterial diversity was higher in the Control silage and A4-treatment, but lower in the A1- and A2-treatments. Interestingly, the same trend was detected in the number of main bacterial species, with > 10% abundance (**Supplementary Figure 4**). The bacterial communities in the Control silage and A6-treatment separated clearly from those of other treatments (**Figure 1**) due to the less abundance of *Lactobacillus* detected in the former (**Figures 2**, **3**). Moreover, Control silage contained more *Lactobacillus* and *Enterococcus* and less *Pediococcus* than A6-treatment (**Figures 2**, **3**), resulting in the separation of the bacterial communities between them (**Figure 1**).

Inoculation of LAB at ensiling optimizes the bacterial community and improves the fermentation quality of the terminal silage (Zhang et al., 2021; Zhao S. et al., 2021). In general, *Lactobacillus* dominates the bacterial community in well-preserved silage owing to its great capacity to produce acid and reduce pH during ensiling (Zi et al., 2021). In the present study, compared with Control silage, the LAB-treatments contained different bacterial communities (**Figure 1**) and had lower pH and NH_3-N/TN (**Table 1**). Moreover, *Lactobacillus* dominated the bacterial communities in the A1-, A2-, A3-, A4-, and A5-treatments (89.32, 92.93, 92.87, 81.12, and 80.44%, respectively) (**Figure 2**). Nevertheless, the most dominant genus in A6-treatment was *Pediococcus* (70.14%) (**Figure 2**), which also caused lower pH, AA, and NH_3-N/TN, and higher LA/AA than those for Control silage (**Table 1**). Such finding indicates that alfalfa silage is also well-preserved, with *Pediococcus* being the dominant genus. The difference in the most dominant bacterial genus among LAB-treatments might be related to the composition of the commercial additives used in the present study. *Lactobacillus plantarum* had the highest composition in A1, A2, A3, A4, and A5. Additionally, previous studies revealed that silage treated with *L. plantarum* had a greater abundance of *Lactobacillus* than Control silage (Zhang et al., 2020; Zhao C. et al., 2021). These results demonstrate that ensiling forage with *L. plantarum* increases the abundance of *Lactobacillus* in the terminal silage. Inoculating *P. acidilactici*, as one of the two components of A6 (*L. buchneri*, $\geq 7.5 \times 10^{10}$ CFU/g; *P. acidilactici*, $\geq 5.0 \times 10^{10}$ CFU/g), increased the abundance of *Pediococcus* in A6-treatment compared with other LAB-treatments (70.14% vs. 0.11–6.81%) (**Figure 2**). Moreover, the A4-treatment contained higher *Pediococcus* than A1-, A2-, A3-, and A5-treatments (6.81% vs. 0.11–4.07%), and *P. acidilactici* is

one of the four components of A4. These results indicate that ensiling alfalfa with *P. acidilactici* increases the abundance of *Pediococcus* in the terminal silage. *Lactobacillus buchneri*, as one of the components of A1, A3, A4, and A6, was only detected in A6-treatment, with 0.17% abundance (**Supplementary Table 1**). Previous studies reported that *L. buchneri*, as an inoculant at ensiling, was detected as a minor taxon in alfalfa silage (Guo et al., 2020) and whole-plant corn silage (Xu et al., 2020; Netthisinghe et al., 2021). Moreover, *E. faecalis* (one of compositions of A1) was only detected in A1- and A2-treatments, with 0.025 and 0.001% abundances, respectively, and *L. casei* (one of compositions of A2 and A4) was only present in A6-treatment (0.011%) (**Supplementary Table 1**). These results indicate that *L. buchneri*, *L. casei*, and *E. faecalis* might have weaker competitiveness than *L. plantarum* and *P. acidilactici* in alfalfa silage with less moisture and low pH environments (475–497 g/kg and 4.33–4.39, respectively). The role of these LAB as the main components of LAB additives during the fermentation process in silage requires further study. The main bacterial genera in Control silage were *Lactobacillus* (42.18%), *Enterococcus* (40.18%), and *Pediococcus* (8.09%), indicating that the LAB genera dominated the bacterial community in Control silage (DM = 496 g/kg). Previous studies reported that the LAB population dominates the bacterial community in low-moisture alfalfa silage (DM > 400 g/kg) without any treatment (Guo et al., 2018, 2020) and presents as minor taxa in high-moisture alfalfa silage (DM < 270 g/kg) without any inoculants (Yang et al., 2020; Zhao S. et al., 2021). These results suggest that wilting alfalfa pre-ensiling may increase the total abundance of LAB genera in the bacterial community of alfalfa silage.

CONCLUSION

The LAB genera are present as minor taxa in fresh alfalfa. Inoculating commercial LAB additives at ensiling alfalfa improved the fermentation quality, contributed to the preservation, and altered the bacterial community of the terminal silage. *Lactobacillus*, *Enterococcus*, and *Pediococcus* dominated the bacterial community in the Control silage. *Lactobacillus* was the most dominant bacterial genus in the A1-, A2-, A3-, A4-, and A5-treatments, and *Pediococcus* was the most dominant in A6-treatment. Further, addition of A5 and A6 decreased the concentrations of NDF, ADF, and hemicellulose in silage. Overall, the commercial lactic acid bacterial additives used in the present study can be employed to inoculate ensiling alfalfa in Northern China.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

NN, YX, and YT designed the study. NN wrote the manuscript. NN, MQ, NW, LS, YZ, and XW performed the experiments. HX, YX, and YT reviewed and edited the manuscript. NN, MQ, NW, HX, and YZ analyzed the data. YX and YT funded and supervised the experiments. All authors reviewed the manuscript.

FUNDING

This work was funded by the National Natural Science Foundation of China (32160342 and 31772674), the Inner Mongolia Science and Technology Plan (2020GG0049), and Inner Mongolia Agriculture and Animal Husbandry Innovation Fund (2022CYZX04).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.836899/full#supplementary-material

Supplementary Figure 1 | Redundancy analysis of the bacterial community (top 10 genera) and fermentation quality of alfalfa silage (n = 4).

Supplementary Figure 2 | Relative abundance of the bacterial community (family level) in alfalfa silage (n = 4).

Supplementary Figure 3 | Relative abundance of potentially pathogenic bacteria in alfalfa silage (n = 4).

Supplementary Figure 4 | Relative abundance of the bacterial community (species level) in alfalfa silage (n = 4).

Supplementary Table 1 | Relative abundance of lactic acid bacteria (species level) in alfalfa silage (n = 4).

REFERENCES

AOAC (2005). *Official Methods of Analysis*, 18th Edn. Gaithersburg, MD: AOAC International.

Bai, C., Wang, C., Sun, L., Xu, H., Jiang, Y., Na, N., et al. (2021). Dynamics of bacterial and fungal communities and metabolites during aerobic exposure in whole-plant corn silages with two different moisture levels. *Front. Microbiol.* 12:663895. doi: 10.3389/fmicb.2021.663895

Besharati, M., Palangi, V., Ghozalpour, V., Nemati, Z., and Ayaşan, T. (2021). Essential oil and apple pomace affect fermentation and aerobic stability of alfalfa silage. *S. Afr. J. Anim. Sci.* 51, 371–377. doi: 10.4314/sajas.v51i3.11

Cai, Y. (1999). Identification and characterization of *Enterococcus* species isolated from forage crops and their influence on silage fermentation. *J. Dairy Sci.* 82, 2466–2471. doi: 10.3168/jds.S0022-0302(99)75498-6

Diepersloot, E. C., Pupo, M. R., Ghizzi, L. G., Gusmão, J. O., Heinzen, C. Jr., McCary, C. L., et al. (2021). Effects of microbial inoculation and storage length on fermentation profile and nutrient composition of whole-plant sorghum silage of different varieties. *Front. Microbiol.* 12:660567. doi: 10.3389/fmicb.2021.660567

Dilokpimol, A., Makela, M. R., Aguilar-Pontes, M. V., Benoit-Gelber, I., Hilden, K. S., and de Vries, R. P. (2016). Diversity of fungal feruloyl esterases: updated

phylogenetic classification, properties, and industrial applications. *Biotechnol. Biofuels* 9:231. doi: 10.1186/s13068-016-0651-6

Duan, X., Dai, Y., and Zhang, T. (2021). Characterization of feruloyl esterase from Bacillus pumilus SK52.001 and its application in ferulic acid production from de-starched wheat bran. *Foods* 10:1229. doi: 10.3390/foods10061229

Gao, R., Wang, B., Jia, T., Luo, Y., and Yu, Z. (2021). Effects of different carbohydrate sources on alfalfa silage quality at different ensiling days. *Agriculture* 11:58. doi: 10.3390/agriculture11010058

Guan, H., Shuai, Y., Yan, Y., Ran, Q., Wang, X., Li, D., et al. (2020). Microbial community and fermentation dynamics of corn silage prepared with heat-resistant lactic acid bacteria in a hot environment. *Microorganisms* 8:719. doi: 10.3390/microorganisms8050719

Guo, L., Yao, D., Li, D., Lin, Y., Bureenok, S., Ni, K., et al. (2020). Effects of lactic acid microbial community, and *in vitro* digestibility of alfalfa silage. *Front. Microbiol.* 10:2998. doi: 10.3389/fmicb.2019.02998

Guo, X. S., Ke, W. C., Ding, W. R., Ding, L. M., Xu, D. M., Wang, W. W., et al. (2018). Profiling of metabolome and bacterial community dynamics in ensiled *Medicago sativa* inoculated without or with *Lactobacillus plantarum* or *Lactobacillus buchneri*. *Sci. Rep.* 8:357. doi: 10.1038/s41598-017-18348-0

Hartinger, T., Gresner, N., and Südekum, K. (2019). Effect of wilting intensity, dry matter content and sugar addition on nitrogen fractions in lucerne silages. *Agriculture* 9:11. doi: 10.3390/agriculture9010011

Hassanat, F., Mustafa, A. F., and Seguin, P. (2007). Effects of inoculation on ensiling characteristics, chemical composition and aerobic stability of regular and brown midrib millet silages. *Anim. Feed Sci. Technol.* 139, 125–140. doi: 10.1016/j.anifeedsci.2007.01.005

Hu, Z., Niu, H., Tong, Q., Chang, J., Yu, J., Li, S., et al. (2020). The microbiota dynamics of alfalfa silage during ensiling and after air exposure, and the metabolomics after air exposure are affected by *Lactobacillus casei* and cellulase addition. *Front. Microbiol.* 11:519121. doi: 10.3389/fmicb.2020.519121

Jiang, F., Cheng, H., Liu, D., Wei, C., An, W., Wang, Y., et al. (2020). Treatment of whole-plant corn silage with lactic acid bacteria and organic acid enhances quality by elevating acid content, reducing pH, and inhibiting undesirable microorganisms. *Front. Microbiol.* 11:593088. doi: 10.3389/fmicb.2020.593088

Kaya, H. I., and Simsek, O. (2020). Characterization of *Pediococcus acidilactici* PFC69 and *Lactococcus lactis* PFC77 bacteriocins and their antimicrobial activities in tarhana fermentation. *Microorganisms* 8:1083. doi: 10.3390/microorganisms8071083

Kung, L. Jr., Shaver, R. D., Grant, R. J., and Schmidt, R. J. (2018). Silage review: Interpretation of chemical, microbial, and organoleptic components of silages. *J. Dairy Sci.* 101, 4020–4033. doi: 10.3168/jds.2017-13909

Li, F., Ding, Z., Adesogan, A. T., Ke, W., Jiang, Y., Bai, J., et al. (2020). Effects of class IIa bacteriocin-producing *Lactobacillus* species on fermentation quality and aerobic stability of alfalfa silage. *Animals* 10, 1575. doi: 10.3390/ani10091575

Logue, J. B., Stedmon, C. A., Kellerman, A. M., Nielsen, N. J., and Andersson, A. F. (2016). Experimental insights into the importance of aquatic bacterial community composition to the degradation of dissolved organic matter. *ISME J.* 10, 533–545. doi: 10.1038/ismej.2015.131

Luo, R., Zhang, Y., Wang, F., Liu, K., Huang, G., Zheng, N., et al. (2021). Effects of sugar cane molasses addition on the fermentation quality, microbial community, and tastes of alfalfa silage. *Animals* 11:355. doi: 10.3390/ani11020355

McDonald, P., and Henderson, A. R. (1964). Determination of water-soluble carbohydrates in grass. *J. Sci. Food Agr.* 15, 395–398. doi: 10.1002/jsfa.2740150609

McDonald, P., Henderson, A. R., and Heron, S. J. E. (1991). *The Biochemistry of Silage*, 2nd Edn. Wales: Cambrian Printers, Ltd.

Netthisinghe, A., Woosley, P., Rowland, N., Willian, T., Gilfillen, B., and Sistani, K. (2021). Alfalfa forage production and nutritive value, fermentation characteristics and hygienic quality of ensilage, and soil properties after broiler litter amendment. *Agronomy* 11:701. doi: 10.3390/agronomy11040701

Paës, G., Berrin, J. G., and Beaugrand, J. (2012). GH11 xylanases: structure/function/properties relationships and applications. *Biotechnol. Adv.* 30, 564–592. doi: 10.1016/j.biotechadv.2011.10.003

Playne, M. J., and McDonald, P. (1966). The buffering constituents of herbage and silage. *J. Sci. Food Agr.* 17, 264–268. doi: 10.1002/jsfa.2740170609

Schmidt, R. J., Hu, W., Mills, J. A., and Kung, L. Jr. (2009). The development of lactic acid bacteria and Lactobacillus buchneri and their effects on the fermentation of alfalfa silage. *J. Dairy Sci.* 92, 5005–5010. doi: 10.3168/jds.2008-1701

Silva, V. P., Pereira, O. G., Leandro, E. S., Da Silva, T. C., Ribeiro, K. G., Mantovani, H. C., et al. (2016). Effects of lactic acid bacteria with bacteriocinogenic potential on the fermentation profile and chemical composition of alfalfa silage in tropical conditions. *J. Dairy Sci.* 99, 1895–1902. doi: 10.3168/jds.2015-9792

Su, R., Ni, K., Wang, T., Yang, X., Zhang, J., Liu, Y., et al. (2019). Effects of ferulic acid esterase-producing Lactobacillus fermentum and cellulase additives on the fermentation quality and microbial community of alfalfa silage. *PeerJ* 7:e7712. doi: 10.7717/peerj.7712

Sun, L., Bai, C., Xu, H., Na, N., Jiang, Y., Yin, G., et al. (2021a). Succession of bacterial community during the initial aerobic, intense fermentation, and stable phases of whole-plant corn silages treated with lactic acid bacteria suspensions prepared from other silages. *Front. Microbiol.* 12:655095. doi: 10.3389/fmicb.2021.655095

Sun, L., Jiang, Y., Ling, Q., Na, N., Xu, H., Vyas, D., et al. (2021b). Effects of adding pre-fermented fluid prepared from lucerne or red clover on fermentation quality and *in vitro* digestibility of the ensiled wilting-forages. *Agriculture* 11:454. doi: 10.3390/agriculture11050454

Surachat, K., Kantachote, D., Deachamag, P., and Wonglapsuwan, M. (2021). Genomic insight into *Pediococcus acidilactici* HN9, a potential probiotic strain isolated from the traditional Thai-Style fermented beef nhang. *Microorganisms* 9:50. doi: 10.3390/microorganisms9010050

Takizawa, S., Asano, R., Fukuda, Y., Feng, M., Baba, Y., Abe, K., et al. (2020). Change of endoglucanase activity and rumen microbial community during biodegradation of cellulose using rumen microbiota. *Front. Microbiol.* 11:603818. doi: 10.3389/fmicb.2020.603818

Thomas, P. C., Chamberlain, D. G., Kelly, N. C., and Wait, M. K. (1980). The nutritive value of silages digestion of nitrogenous constituents in sheep receiving diets of grass-silage and grass silage and barley. *Br. J. Nutr.* 43, 469–479. doi: 10.1079/BJN19800114

Van Soest, P. J., Roberts, J., and Lewis, B. A. (1991). Methods for dietary fibre neutral detergent fibre and nonstarch polysaccharides in relation to animal nutrition. *J. Dairy Sci.* 74, 3583–3594. doi: 10.3168/jds.S0022-0302(91)78551-2

Vucinic, J., Novikov, G., Montanier, C. Y., Dumon, C., Schiex, T., and Barbe, S. (2021). A comparative study to decipher the structural and dynamics determinants underlying the activity and thermal stability of GH-11 xylanases. *Int. J. Mol. Sci.* 22:5961. doi: 10.3390/ijms22115961

Wang, B., Sun, Z., and Yu, Z. (2020). Pectin degradation is an important determinant for alfalfa silage fermentation through the rescheduling of the bacterial community. *Microorganisms* 8:488. doi: 10.3390/microorganisms8040488

Wang, C., Han, H., Sun, L., Na, N., Xu, H., Chang, S., et al. (2021). Bacterial succession pattern during the fermentation process in whole-plant corn silage processed in different geographical areas of Northern China. *Processes* 9:900. doi: 10.3390/pr9050900

Wang, C., He, L., Xing, Y., Zhou, W., Yang, F., Chen, X., et al. (2019). Effects of mixing Neolamarckia cadamba leaves on fermentation quality, microbial community of high moisture alfalfa and stylo silage. *Microb. Biotechnol.* 12, 869–878. doi: 10.3390/10.1111/1751-7915.13429

Wang, M., Gao, R., Franco, M., Hannaway, D. B., Ke, W., Ding, Z., et al. (2021). Effect of mixing alfalfa with whole-plant corn in different proportions on fermentation characteristics and bacterial community of silage. *Agriculture* 11:174. doi: 10.3390/agriculture11020174

Weissbach, F., and Strubelt, C. (2008). Correcting the dry matter content of grass silages as a substrate for biogas production. *J. Agric. Eng.* 63:210.

Wu, Z., Xu, S., Yun, Y., Jia, T., and Yu, Z. (2020). Effect of 3-phenyllactic acid and 3-phenyllactic acid-producing lactic acid bacteria on the characteristics of alfalfa silage. *Agriculture* 10:10. doi: 10.3390/agriculture10010010

Xie, Y., Bao, J., Li, W., Sun, Z., Gao, R., Wu, Z., et al. (2021). Effects of applying lactic acid bacteria and molasses on the fermentation quality, protein fractions and *in vitro* digestibility of baled alfalfa silage. *Agronomy* 11:91. doi: 10.3390/agronomy11010091

Xu, D., Ding, W., Ke, W., Li, F., Zhang, P., and Guo, X. (2019). Modulation of metabolome and bacterial community in whole crop corn silage by inoculating homofermentative *Lactobacillus plantarum* and heterofermentative *Lactobacillus buchneri*. *Front. Microbiol.* 9:3299. doi: 10.3389/fmicb.2018.03299

Xu, D., Ding, Z., Wang, M., Bai, J., Ke, W., Zhang, Y., et al. (2020). Characterization of the microbial community, metabolome and biotransformation of phenolic compounds of sainfoin (*Onobrychis viciifolia*) silage ensiled with or without inoculation of Lactobacillus plantarum. *Bioresour. Technol.* 316:123910. doi: 10.1016/j.biortech.2020.123910

Xue, Y., Bai, C., Sun, J., Sun, L., Chang, S., Sun, Q., et al. (2017). Effects of locations and growth stages on nutritive value and silage fermentation quality of Leymus chinensis in Eurasian steppe of northern China. *Grassl. Sci.* 64, 40–50. doi: 10.1111/grs.12177

Yang, F., Wang, Y., Zhao, S., and Wang, Y. (2020). *Lactobacillus plantarum* Inoculants delay spoilage of high moisture alfalfa silages by regulating bacterial community composition. *Front. Microbiol.* 11:1989. doi: 10.3389/fmicb.2020.01989

Yang, F., Zhao, S., Wang, Y., Fan, X., Wang, Y., and Feng, C. (2021). Assessment of bacterial community composition and dynamics in alfalfa silages with and without *Lactobacillus plantarum* inoculation using absolute quantification 16S rRNA sequencing. *Front. Microbiol.* 11:629894. doi: 10.3389/fmicb.2020.629894

Yin, G., Bai, C., Sun, J., Sun, L., Xue, Y., Zhang, Y., et al. (2017). Fermentation quality and nutritive value of total mixed ration silages based on desert wormwood (*Artemisia desertorum* Spreng.) combining with early stage corn. *Anim. Sci. J.* 88, 1963–1969. doi: 10.1111/asj.12862

Zhang, G., Fang, X., Feng, G., Li, Y., and Zhang, Y. (2020). Silage fermentation, bacterial community, and aerobic stability of total mixed ration containing wet corn gluten feed and corn stover prepared with different additives. *Animals* 10:1775. doi: 10.3390/ani10101775

Zhang, M., Wang, L., Wu, G., Wang, X., Lv, H., Chen, J., et al. (2021). Effects of *Lactobacillus plantarum* on the fermentation profile and microbiological composition of wheat fermented silage under the freezing and thawing low temperatures. *Front. Microbiol.* 12:671287. doi: 10.3389/fmicb.2021.671287

Zhao, C., Wang, L., Ma, G., Jiang, X., Yang, J., Lv, J., et al. (2021). Cellulase interacts with lactic acid bacteria to affect fermentation quality, microbial community, and ruminal degradability in mixed silage of soybean residue and corn stover. *Animals* 11:334. doi: 10.3390/ani11020334

Zhao, S., Yang, F., Wang, Y., Fan, X., Feng, C., and Wang, Y. (2021). Dynamics of fermentation parameters and bacterial community in high-moisture alfalfa silage with or without lactic acid bacteria. *Microorganisms* 9:1225. doi: 10.3390/microorganisms9061225

Zheng, M. L., Niu, D. Z., Jiang, D., Zuo, S. S., and Xu, C. C. (2017). Dynamics of microbial community during ensiling direct-cut alfalfa with and without LAB inoculant and sugar. *J. Appl. Microbiol.* 122, 1456–1470. doi: 10.1111/jam.13456

Zi, X., Li, M., Chen, Y., Lv, R., Zhou, H., and Tang, J. (2021). Effects of citric acid and lactobacillus plantarum on silage quality and bacterial diversity of king grass silage. *Front. Microbiol.* 12:631096. doi: 10.3389/fmicb.2021.631096

Conflict of Interest: XW and YT were employed by Inner Mongolia Youran Animal Husbandry Co., Ltd.

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.

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 25 April 2022
doi: 10.3389/fmicb.2022.870168

D-Allulose (D-Psicose) Biotransformation From Allitol by a Newly Found NAD(P)-Dependent Alcohol Dehydrogenase From *Gluconobacter frateurii* NBRC 3264 and the Enzyme Characterization

Xin Wen[1], Huibin Lin[2], Yuhang Ning[1], Guangwen Liu[1], Yilin Ren[3], Can Li[4], Chengjia Zhang[1], Jianqun Lin[1], Xin Song[1]* and Jianqiang Lin[1]*

[1] State Key Laboratory of Microbial Technology, Shandong University, Qingdao, China, [2] Shandong Academy of Chinese Medicine, Jinan, China, [3] Qingdao Longding Biotech Limited Company, Qingdao, China, [4] School of Biological Engineering, Qilu University of Technology, Jinan, China

OPEN ACCESS

Edited by:
Zhenhua Ming,
Guangxi University, China

Reviewed by:
Jianmin Xing,
Institute of Process Engineering (CAS), China
Xiuzhen Gao,
Shandong University of Technology, China

***Correspondence:**
Jianqun Lin
jianqunlin@sdu.edu.cn
Xin Song
songx@sdu.edu.cn

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

Received: *06 February 2022*
Accepted: *17 March 2022*
Published: *25 April 2022*

Citation:
Wen X, Lin H, Ning Y, Liu G, Ren Y, Li C, Zhang C, Lin J, Song X and Lin J (2022) D-Allulose (D-Psicose) Biotransformation From Allitol by a Newly Found NAD(P)-Dependent Alcohol Dehydrogenase From Gluconobacter frateurii NBRC 3264 and the Enzyme Characterization. Front. Microbiol. 13:870168. doi: 10.3389/fmicb.2022.870168

The NAD(P)-dependent alcohol dehydrogenase (ADH) gene was cloned from *Gluconobacter frateurii* NBRC 3264 and expressed in *Escherichia coli* BL21 star (DE3). The expressed enzyme was purified and the characteristics were investigated. The results showed that this ADH can convert allitol into D-allulose (D-psicose), which is the first reported enzyme with this catalytic ability. The optimum temperature and pH of this enzyme were 50°C and pH 7.0, respectively, and the enzyme showed a maximal activity in the presence of Co^{2+}. At 1 mM Co^{2+} and allitol concentrations of 50, 150, and 250 mM, the D-allulose yields of 97, 56, and 38%, respectively, were obtained after reaction for 4 h under optimal conditions, which were much higher than that obtained by using the epimerase method of about 30%.

Keywords: D-allulose, allitol, NAD(P)-dependent alcohol dehydrogenase, *Gluconobacter frateurii NBRC 3264*, biotransformation

INTRODUCTION

D-Allulose (D-psicose), an epimer of D-fructose at the C3 position, is a kind of rare sugar according to the definition by the International Society of Rare Sugars (ISRS). D-Allulose is a low-energy sweet and is regarded as a potential substitute for sucrose as it has 70% of the relative sweetness but only 0.3% of the energy of sucrose (Zhang et al., 2015). More importantly, D-allulose has many important physiological functions, for example, blood glucose suppressive effect (EdyLiani et al., 2020), body fat accumulation inhibitive effect (Kim et al., 2017), reactive oxygen species scavenging effect (Li et al., 2018), and neuroprotective effect (Zhao et al., 2021). In addition, it has good properties for food industry applications, such as improving the gelling behavior and producing good flavor (Zhang et al., 2013). Importantly, it has been approved as "generally regarded as safe" (GRAS) by the Food and Drug Administration (FDA) of the United States, and has been allowed to be used as an ingredient in dietary supplements in the United States and some other countries.

In nature, D-allulose is found in very small amounts in the wheat and *Itea* plants. So, it is impractical to extract it from natural resources for mass production of D-allulose. The chemical synthetic method is one choice, but it may produce toxic by-products and is not suitable for food production. Biotransformation is an ideal method and is widely accepted in D-allulose mass

production due to the advantages of easy operation, mild reaction conditions, no toxic by-products, and environmental friendliness. At present, D-allulose was namely biotransformed from D-fructose by using D-psicose 3-epimerase or D-tagatose 3-epimerase (Zhu et al., 2012, 2019c; Li et al., 2018; **Figure 1**). However, the reaction catalyzed by epimerase is limited by thermodynamic equilibrium unfavorable to the D-allulose direction, and the conversion yield of D-allulose is about 30%, which greatly decreases the production efficiency and increases the difficulty in product separation. To overcome the limitation of thermodynamic equilibrium, Kim et al. added boronic acid to the reaction system to form a complex with sugar to increase the D-allulose conversion yield (Kim et al., 2008). As the binding affinity of boric acid to D-allulose is much higher than that of D-fructose, the reaction equilibrium is shifted toward the formation of D-allulose, and that increases the conversion yield of D-allulose (Kim et al., 2008). However, boric acid is toxic and used in large quantities, and the removal of boric acid is difficult. For the above reasons, this method is difficult to be applied in real applications. Alternatively, the thermodynamic equilibrium limitation can also be overcome by combining the D-allulose biocatalytic process with continuous D-allulose separation (Wagner et al., 2015; Li et al., 2021). However, this method is complex and cumbersome and is also difficult to be applied in real applications.

Fortunately, D-allulose can also be biotransformed from allitol by using dehydrogenation reaction using dehydrogenase as the catalyst according to the Izumoring strategy (Izumori, 2006), which can overcome the above limitation of the thermodynamic equilibrium and improve the conversion rate of D-allulose. Moreover, allitol can be prepared easily from low-cost substrates of D-glucose or D-fructose by the biotransformation method (Zhu et al., 2015; Hassanin et al., 2016; Wen et al., 2020a,b). Poonperm et al. (2007) biotransformed allitol into D-allulose by using the resting cells of *Bacillus pallidus* Y25 for the first time. Gullapalli et al. (2007). biotransformed allitol into D-allulose by using *Enterobacter aerogenes* IK7. However, the exact enzyme that catalyzed allitol into D-allulose was unknown.

In this study, the gene encoding NAD(P)-dependent alcohol dehydrogenase (ADH) with protein ID WP_099183078.1 from *Gluconobacter frateurii* NBRC 3264 was cloned and overexpressed in *E. coli*. The ADH was confirmed to convert allitol into D-allulose (D-psicose), which is the first reported enzyme with this catalytic ability. The enzymatic properties, such as optimal pH, temperature, and metal ion, of this ADH were investigated. The activation effect of Co^{2+} on the ADH to increase the enzyme activity and the D-allulose yield was determined, and the kinetics of this enzyme were also investigated. The highest D-allulose conversion yield of 97% was obtained, which was more than twofold higher than the epimerase method. The method developed in this study is expected to be applied to the industrial production of D-allulose.

MATERIALS AND METHODS

Materials and Reagents

The restriction enzymes were obtained from TaKaRa (Beijing, China). The DNA polymerase was obtained from Vazyme (Nanjing, China). T_4 DNA ligase was purchased from Thermo Fisher (United States). Ampicillin and isopropyl-β-D-1-thiogalactopyranoside (IPTG) were purchased from Sangon Biotech (Shanghai, China). Allitol was prepared in our lab as described previously (Wen et al., 2020a,b, 2022).

Construction of Recombinant *E. coli* Expressing Alcohol Dehydrogenase

According to NCBI, the whole genome of *Gluconobacter frateurii* NBRC 3264 was sequenced by Hosoyama et al. and was released into the GenBank National Center for Biotechnology Information (NCBI)[1]. The *adh* gene locus_tag was GFR01_RS14945 and the ADH protein ID number was WP_099183078.1. The optimization and synthesis of the gene encoding NAD(P)-dependent alcohol dehydrogenase (ADH) were made by a company named Boshang (Jinan, China). The *adh* region was initially amplified from the plasmid pETDuet$_{-1}$-*adh* (no 6 × His-tag) using primers *adh*-pET22b-*Nde* I-U and *adh*-pET22b-*Xho* I-D (**Table 1**). A 6 × His-tag sequence was present in the vector to aid protein purification. Then, the *adh* region was inserted into the plasmid pET22b at the *Nde* I and *Xho* I restriction sites to create the recombinant plasmid pET22b-*adh*. The recombinant plasmid pET22b-*adh* was transformed into *E. coli* DH5α and verified correctly by electrophoresis and sequencing. And then, the recombinant plasmid pET22b-*adh* was transformed into *E. coli* BL21 star (DE3) for the expression of ADH. The strains, plasmids, and primers used in this study are listed in **Table 1**.

Media and Cultivation Conditions

The seed culture used in this study was the LB medium containing 10 g/L tryptone, 5 g/L yeast extract, and 10 g/L NaCl. The LB medium supplied with 5 g/L glucose (named LBG medium) was used for the expression of ADH. The cultivation broth of recombinant *E. coli* expressing ADH was inoculated with 1% dose into the LBG medium containing 100 μg/ml ampicillin, and cultivated at 37°C and 200 rpm. After 3 h of cultivation, IPTG was added to the final concentration of 0.2 mM and the cultivation was continued for a further 12 h at 20°C and 100 rpm. The cells of the recombinant *E. coli* expressing ADH were harvested by centrifugation at 4°C and 10,000 × g for 5 min.

Crude Alcohol Dehydrogenase Preparation, Alcohol Dehydrogenase Purification, and Enzyme Assay

The harvested cells were washed three times by using 20 mM Na_2HPO_4-NaH_2PO_4 buffer (pH 7.0). The washed cells were collected by centrifugation, resuspended in 20 mM Na_2HPO_4-NaH_2PO_4 buffer (pH 7.0), and disrupted by sonication at 4°C until the mixture solution became transparent. The supernatant was obtained by centrifugation at 4°C and 10,000 × g for 15 min and was used for crude ADH. The crude ADH was checked by Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis (SDS–PAGE).

[1]https://www.ncbi.nlm.nih.gov/nuccore/1271388588

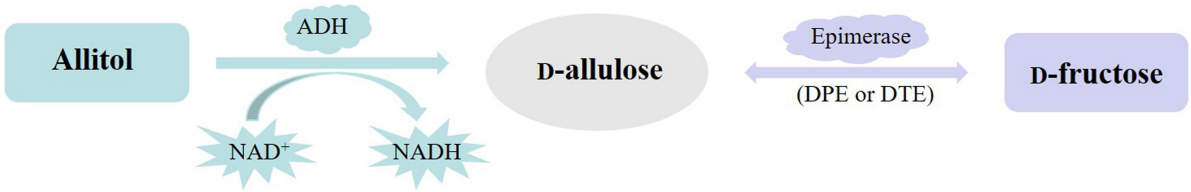

FIGURE 1 | D-Allulose biotransformation from allitol or D-fructose (ADH, NAD(P)-dependent alcohol dehydrogenase; DPE, D-psicose 3-epimerase; DTE, D-tagatose 3-epimerase).

The preparation of crude ADH used for ADH purification is the same as the above except the washing buffer and resuspending buffer were changed to the binding buffer (20 mM NaH_2PO_4, 500 mM NaCl, 30 mM imidazole, pH 7.4). HisTrap^{TM} HP (5 mL) column was used for the purification of the recombinant ADH. The column was washed using double-distilled water and equilibrated with a binding buffer. And then, the collected supernatant was loaded onto the column, and the unbound proteins were washed with the binding buffer, and the ADH was then washed with the elution buffer (20 mM NaH_2PO_4, 500 mM NaCl, 200 mM imidazole, pH 7.4). Finally, the purified ADH was checked by SDS–PAGE and was concentrated by the ultrafiltration tube with the membrane of the cutoff molecular weight of 10 kDa at 4°C and 3,700 × g. All purification steps of ADH were handled at 4°C.

The 1 ml reaction mixture for ADH assay consisted of each of the following reagents unless otherwise specified: 20 mM Na_2HPO_4-NaH_2PO_4 buffer (pH 7.0), 2 mM NAD^+, enzyme solution, and 50 mM allitol, and then incubated at 50°C and 200 rpm shaker for 30 min. One unit of enzyme activity was defined as the amount of D-allulose produced from allitol per minute. The amount of allitol and D-allulose were measured by HPLC using a Carbomix Pb-NP column (7.8 mm × 300 mm, 10 μm, Sepax Technologies) at 78°C and eluted with double-distilled water at a flow rate of 0.5 ml/min.

Effects of pH, Temperature, and Metal Ions on Recombinant Alcohol Dehydrogenase and Kinetic Modeling

Four buffer systems of sodium acetate–acetic acid (20 mM, pH 5.0–6.0), disodium hydrogen phosphate–sodium dihydrogen phosphate (20 mM, pH 6.0–8.0), tris–HCl (20 mM, pH 8.0–9.0), and glycine–NaOH (20 mM, pH 9.0–11.0) were, respectively, used in determining the optimum pH of the recombinant ADH expressed by *E. coli*.

The optimum temperature for the enzyme activity was measured by assaying the enzyme solution over the temperature range of 30–60°C. The thermal stability of the recombinant ADH was investigated by maintaining the enzyme solution in disodium hydrogen phosphate–sodium dihydrogen phosphate (20 mM, pH 7.0) at various temperatures for 3 h and measuring the residual enzyme activities at 0.5-h intervals.

The residual activity of the enzyme was determined as described in the above method in the "Crude ADH preparation, ADH purification, and enzyme assay." The enzyme solution was incubated with the metal ions Co^{2+}, Zn^{2+}, Ni^{2+}, Ca^{2+}, Mg^{2+}, Ba^{2+}, Fe^{3+}, Mn^{2+}, Fe^{2+}, and Cu^{2+} at a final concentration of 1 mM. The measured activities were compared with the activity of the enzyme without the metal ion addition (control) under the same conditions.

Kinetic modeling can help to understand the reaction characteristics of this enzyme and predict the reaction results. The reaction rate is normally affected by the substrate concentration, while it is also strongly affected by Co^{2+} for the ADH under investigation. Here, the D-allulose production kinetics under various substrate concentrations of 50, 150, and 250 mM allitol, respectively, with or without the activator of Co^{2+} addition, were investigated.

D-Allulose Identification

The product was identified by using the HPLC analysis, specific optical rotations, and mass spectrometry. The high performance liquid chromatography (HPLC) analysis method was referred to in "Crude ADH preparation, ADH purification, and enzyme assay." Specific optical rotations were determined by using the polarimeter (INESA WZZ-3, China). Mass spectrum (BRUKER impactHD, Germany) was performed in the negative ion detection mode with the ESI ion source.

TABLE 1 | Plasmids, strains and primers used in this study.

Plasmids, strains and primers	**Relevant characteristics, sources and sequences**	
Plasmids and strains	**Relevant characteristics**	**Sources**
pETDuet_1-MCSII*adh*	*adh* (no His·Tag), Amp^r	Boshang (Jinan, China)
E. coli DH5α	For gene cloning	Weidi (Shanghai, China)
E. coli BL21 star (DE3)	For gene expression	Weidi (Shanghai, China)
pET22b-*adh*	*adh* (His·Tag), Amp^r	This study
E. coli DH5α-pET22b-*adh*	For plasmid cloning	This study
E. coli BL21 star (DE3)-pET22b	Empty plasmid pET22b	This study
E. coli BL21 star (DE3)-pET22b-*adh*	ADH protein	This study
Primers	**Sequences (5′–3′)**	
adh-pET22b-*Nde* I-U	GGGAATTCCATATG*GCCCAGGCCCTGGTGCTGGAAAAG	
adh-pET22b-*Xho* I-D	CCGCTCGAGCAGAACAATCTGCAGTTTAACATC	

**Underlines refer to enzyme restriction sites.*

```
            Nde I
1           CAT ATG GCC CAG GCC CTG GTG CTG GAA AAG AAA GGC GAA CTG AGT CTG CGC GAA ATT GCC
1           His MET Ala Gln Ala Leu Val Leu Glu Lys Lys Gly Glu Leu Ser Leu Arg Glu Ile Ala

61          CTG CCG AGC GAA CTG GGT CCG AAT GAT GTT CGT ATT GCA ATT CAT ACC GTG GGC ATT TGT
21          Leu Pro Ser Glu Leu Gly Pro Asn Asp Val Arg Ile Ala Ile His Thr Val Gly Ile Cys

121         GGC AGC GAT GTG CAT TAT TAT ACC CAT GGT GCC ATT GGT CCG TTT GTT GTT CGC GAA CCG
41          Gly Ser Asp Val His Tyr Tyr Thr His Gly Ala Ile Gly Pro Phe Val Val Arg Glu Pro

181         ATG GTT CTG GGC CAT GAA GCA AGT GGC ACC ATT ACC GAA ATT GGT AGT AAT GTG CGC AGT
61          MET Val Leu Gly His Glu Ala Ser Gly Thr Ile Thr Glu Ile Gly Ser Asn Val Arg Ser

241         CTG AAA GTT GGC GAT CGT GTT TGC ATG GAA CCG GGC ATT CCG GAT CCG CAG AGT CGC GCA
81          Leu Lys Val Gly Asp Arg Val Cys MET Glu Pro Gly Ile Pro Asp Pro Gln Ser Arg Ala

301         ACC CTG ATG GGT CAG TAT AAT GTG GAT CCG GCC GTG CGC TTT TGG GCA ACC CCG CCT ATT
101         Thr Leu MET Gly Gln Tyr Asn Val Asp Pro Ala Val Arg Phe Trp Ala Thr Pro Pro Ile

361         CAT GGT TGC CTG ACC CCG AGT GTG GTG CAT CCG GCA GCA TTC ACT TTT AAA CTG CCG GAT
121         His Gly Cys Leu Thr Pro Ser Val Val His Pro Ala Ala Phe Thr Phe Lys Leu Pro Asp

421         AAT GTT AGT TTT GCC GAA GGC GCC ATG ATT GAA CCG CTG GCA GTG GGT GTG CAT GCA AGC
141         Asn Val Ser Phe Ala Glu Gly Ala MET Ile Glu Pro Leu Ala Val Gly Val His Ala Ser

481         GTG AAA GCA GCC ATT AAG CCG GGT GAC ATT TGT CTG GTG ACC GGC TGC GGT CCG ATT GGT
161         Val Lys Ala Ala Ile Lys Pro Gly Asp Ile Cys Leu Val Thr Gly Cys Gly Pro Ile Gly

541         ATT ATG ACC GCC CTG GCA GCC CTG GCC AGT GGC GCA GGT CAG GTG TTT ATT ACC GAT CTG
181         Ile MET Thr Ala Leu Ala Ala Leu Ala Ser Gly Ala Gly Gln Val Phe Ile Thr Asp Leu

601         GCC CCG GCA AAA CTG GCA ATT GCA GGT CAG TAT GAT GGT ATT CGC CCG ATT AAT GTT CGT
201         Ala Pro Ala Lys Leu Ala Ile Ala Gly Gln Tyr Asp Gly Ile Arg Pro Ile Asn Val Arg

661         GAT GAA AAA CCG CGT GAT GTG GTT GAT GCA ACC TGT GGC AGC GAC TGG GGT GTG GAT GTT
221         Asp Glu Lys Pro Arg Asp Val Val Asp Ala Thr Cys Gly Ser Asp Trp Gly Val Asp Val

721         GTG TTT GAA GCA AGC GGT TTT GCC GGC GCA TAC GAT GAT GCC CTG GCC TGC GTG CGT CCG
241         Val Phe Glu Ala Ser Gly Phe Ala Gly Ala Tyr Asp Asp Ala Leu Ala Cys Val Arg Pro

781         GGC GGT ACC ATT GTG TTT GTG GGT ATG CCG ATT CAG AAA GTG CCG TTT GAT ATT GTG GCC
261         Gly Gly Thr Ile Val Phe Val Gly MET Pro Ile Gln Lys Val Pro Phe Asp Ile Val Ala

841         GCC CAG GCA AAA GAA ATT CGT ATG GAA ACC GTG TTT CGC TAT GCC AAT GTT TAT GAT CGT
281         Ala Gln Ala Lys Glu Ile Arg MET Glu Thr Val Phe Arg Tyr Ala Asn Val Tyr Asp Arg

901         GCA ATT CGC CTG ATT AGT GCA GGC AAA ATT GAT CTG AAA CCG CTG GTG AGC GAA ACC TTT
301         Ala Ile Arg Leu Ile Ser Ala Gly Lys Ile Asp Leu Lys Pro Leu Val Ser Glu Thr Phe

961         CCG TTT GAT CAG GGT ATT GCC GCA TTT GAA CGT GCC GCA GAA GCA CGC CCG AGC GAT GTT
321         Pro Phe Asp Gln Gly Ile Ala Ala Phe Glu Arg Ala Ala Glu Ala Arg Pro Ser Asp Val
                                Xho I
1021        AAA CTG CAG ATT GTT CTG CTC GAG
341         Lys Leu Gln Ile Val Leu Leu Glu
```

FIGURE 2 | The amino acid sequence and optimized gene sequence of ADH.

RESULTS AND DISCUSSION

Cloning, Expression, Purification, and Application of Recombinant *Gluconobacter frateurii* NBRC 3264 Alcohol Dehydrogenase

The *adh* gene was optimized and synthesized and cloned into pET22b to obtain the recombinant plasmid pET22b-*adh*, which was transformed into *E. coli* BL21 star (DE3). The amino acid sequence (345aa) and the optimized gene sequence of ADH are shown in **Figure 2**. The recombinant ADH expression was induced by IPTG. The SDS–PAGE analysis showed a strong extra protein band with a molecular mass of ∼36.5 kDa compared with that of the control *E. coli* BL21 star (DE3)-pET22b and confirmed the soluble property of ADH (**Figure 3A**). The purification of recombinant ADH was carried out by using the HisTrap^{TM} HP (5 mL) column. The result of the ADH purification was

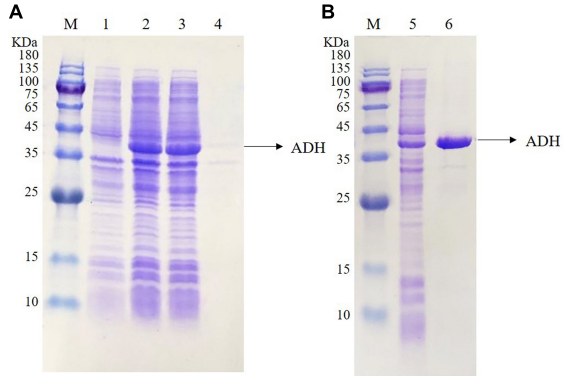

FIGURE 3 | Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis (SDS-PAGE) analysis of the expressed ADH **(A)** and the purified ADH **(B)**. Lane M, protein marker; lane 1, the total proteins of *E. coli* BL21 star (DE3)-pET22b; lane 2, the total proteins of *E. coli* BL21 star (DE3)-pET22b-*adh*; lane 3, the soluble supernatant of *E. coli* BL21 star (DE3)-pET22b-*adh*; lane 4, the inclusion body of *E. coli* BL21 star (DE3)-pET22b-*adh*; lane 5, ADH crude enzyme solution; lane 6, ADH purified enzyme solution.

analyzed by the SDS–PAGE (**Figure 3B**), and the purified ADH was concentrated ten times by ultrafiltration.

The ADHs catalyze interconversions between alcohols and aldehydes or ketones (Maria-Solano et al., 2017; Zheng et al., 2017; Bartsch et al., 2020). For example, alcohol dehydrogenase from *Pyrococcus furiosus* can catalyze 2, 5-hexanedione to 2, 5-hexanediol (Machielsen et al., 2008). In addition, a sorbitol dehydrogenase (340aa), a homologous enzyme to the alcohol dehydrogenase, which had the same amino acid sequence of ADH from 4 to 343aa, catalyzed the conversion of D-sorbitol to D-fructose in the presence of NAD^+ (El-Kabbani et al., 2004). The purified and concentrated ADH was inoculated into the reaction solution containing 20 mM Na_2HPO_4-NaH_2PO_4 buffer (pH 7.0), 2 mM NAD^+, and 50 mM allitol, and reacted at 50°C shaken at 200 rpm. As shown in **Figure 4**, the ADH was preliminary confirmed to catalyze allitol into allulose. Next, specific optical rotations of authentic L-allulose, authentic D-allulose, and the purified product were measured. The specific rotation of authentic L-allulose was negative, while the specific rotation of authentic D-allulose and the purified product was positive which agreed with the reports (Gullapalli et al., 2007; Poonperm et al., 2007). Further, the purified product was analyzed by mass spectrometry with a measured mass of 180.1, which was identical to the molar mass of D-allulose. In conclusion, ADH from *G. frateurii* NBRC 3264 can convert allitol into D-allulose, which is the first reported enzyme with this catalytic ability.

Effect of pH on D-Allulose Biotransformation by Recombinant Alcohol Dehydrogenase

Figure 5A shows that the optimum pH is 7.0, and the relative enzyme activities are above 80% between pH 7.0 and pH 10.0, which indicates that the ADH has a broad pH range. The optimum pH for D-allulose biotransformation from allitol by *Bacillus pallidus* Y25 resting cells was also pH 7.0 (Poonperm et al., 2007). However, the optimum pH for D-allulose biotransformation from allitol by *Enterobacter aerogenes* IK7 was pH 11.0 which was much higher than that of the recombinant ADH (Gullapalli et al., 2007). But, the optimum pH of the enzyme could be different from that of the resting cells in catalyzing the same reaction.

Effect of Temperature on D-Allulose Biotransformation and Enzyme Stability of the Recombinant Alcohol Dehydrogenase

Figure 5B shows that the optimum temperature is 50°C, and the relative enzyme activities are 63.8, 79.3, 83, and 52% at 40, 45, 55, and 60°C, respectively, compared with that at the optimum temperature. The optimum temperature of *Enterobacter aerogenes* IK7 resting cells for D-allulose biotransformation from allitol was 37°C (Gullapalli et al., 2007), which was lower than that of the recombinant ADH. Nevertheless, the optimum temperature of *Bacillus pallidus* Y25 resting cells for D-allulose biotransformation from allitol was 55°C (Poonperm et al., 2007), which was higher than that of the recombinant ADH.

As seen in **Figure 5C**, the enzyme has similar thermal stability at 20, 30, and 40°C, and retains 74.7, 74.4, and 73.2% of its initial activity, respectively, after incubation for 3 h at the above temperatures while the enzyme retained 71.5, 42.4, and 25.6% of its initial activity after incubation at 50°C (**Figure 5C**) for 1, 2, and 3 h, respectively. The results indicated that the ADH had lower thermal stability at a temperature higher than 40°C. Protein engineering is a way to increase the thermal stability of ADH (Magnusson et al., 2019; Zhu et al., 2019b).

Effect of Metal Ions on D-Allulose Biotransformation by the Recombinant Alcohol Dehydrogenase

As shown in **Figure 6**, the addition of Co^{2+}, Zn^{2+}, or Ni^{2+} increases the enzyme activity by 225, 54.1, and 19.1 %, respectively. It was speculated that Co^{2+} or Ni^{2+} was an activator that can bind to the enzyme and change the enzyme configuration to increase the enzyme activity. It was reported that Zn^{2+} plays an important role in the structure and function of alcohol dehydrogenase and sorbitol dehydrogenase (El-Kabbani et al., 2004). The enzyme activity was slightly decreased by 2.6 and 3.3% when the enzyme was incubated with Ca^{2+} and Mg^{2+}, respectively, while the enzyme activity was decreased to 83.1, 69.4, 52.4, 50.3, and 30.3% when the enzyme was incubated with Ba^{2+}, Fe^{3+}, Mn^{2+}, Fe^{2+}, and Cu^{2+}, respectively. About the activity of NAD-dependent sorbitol dehydrogenase from cold-adapted *Pseudomonas mandelii*, the metal ions of Zn^{2+}, Mn^{2+}, and Ca^{2+} had slight activation effects while Ni^{2+} had an inhibition effect (DangThu et al., 2021). Ni^{2+}, Mn^{2+}, Mg^{2+}, and Ca^{2+} can increase the ADH activity which was from *Bartonella apis*, while Zn^{2+}, Li^+, and Mo^{2+} decrease the ADH activity

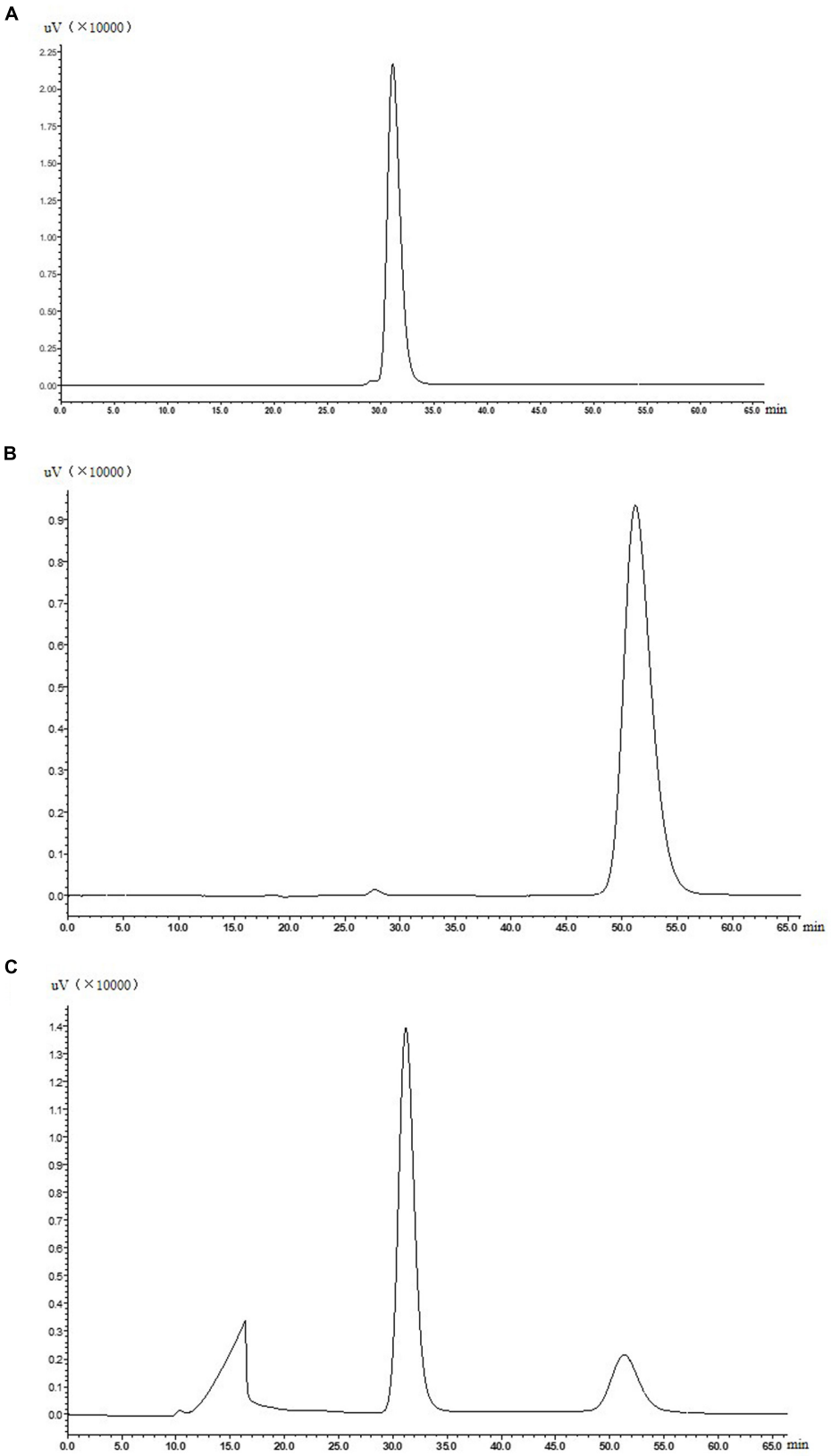

FIGURE 4 | Authentic allitol **(A)**, authentic D-allulose **(B)**, and a sample of reaction solution for the biotransformation of allitol into D-allulose catalyzed by purified ADH **(C)**.

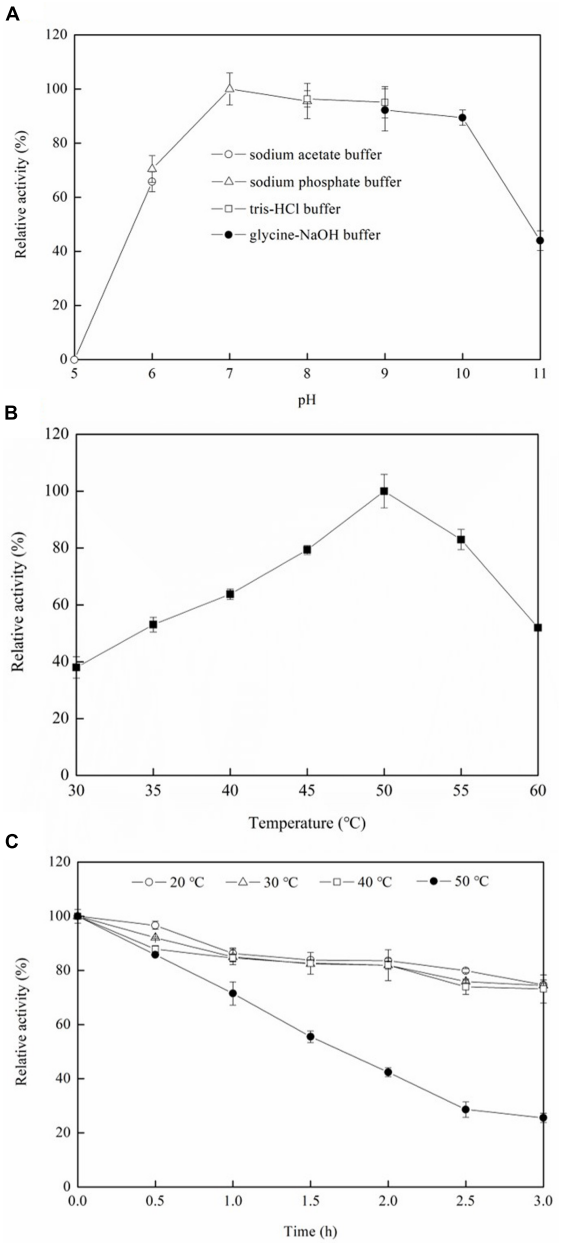

FIGURE 5 | Effects of pH **(A)** and temperature **(B)** on the ADH activities, and the thermal stability of the ADH **(C)**. The conditions for obtaining the highest enzyme activity were set to 100%.

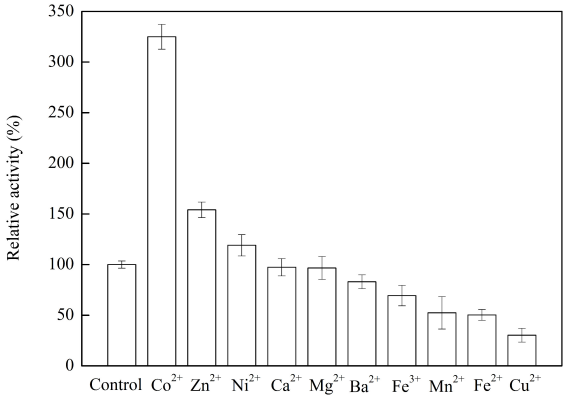

FIGURE 6 | Effects of metal ions on the ADH activity. The conditions for obtaining the highest enzyme activity were set to 100%.

(Zhu et al., 2019a). It indicated that the metal-ion-dependence of ADHs derived from different microorganisms was different.

Effects of Co^{2+} on D-Allulose Biotransformation by the Recombinant Alcohol Dehydrogenase and Kinetic Modeling

The time courses of D-allulose and allitol concentrations in the presence or absence of Co^{2+} at different allitol concentrations are shown by the dots in **Figure 7**. The D-allulose conversion yields of 97, 56, and 38%, from the initial allitol concentrations of 50, 150, and 250 mM, respectively, were obtained at 4 h of reaction with 1 mM Co^{2+} added, which was about 1. 6-, 1. 7-, and 1.7-fold higher, respectively, than that without the Co^{2+} addition.

Then, kinetic modeling was made for D-allulose biotransformation catalyzed by ADH with or without the Co^{2+} addition. Without the Co^{2+} addition, the kinetic equation is shown by Equation (1) and the mass balances are shown by Equations (2) and (3):

$$V = \frac{V_{max}\, S}{\left(k_s + S\right)\ \left(1 + (P/k_i)^{\alpha}\right)} \tag{1}$$

$$\frac{dS}{dt} = -V \tag{2}$$

$$\frac{dP}{dt} = V \tag{3}$$

Where, V_{max}, the maximum reaction rate without Co^{2+}, mmol/L/h; k_s, the substrate affinity constant without Co^{2+}, mM; k_i, the product inhibition constant, mM; α, constant, (-); S, allitol concentration, mM; P, D-allulose concentration, mM. With Co^{2+} addition, the kinetic and mass balance equations are as follows:

$$V' = \frac{V_{max}\, S}{\left(k_s' + S\right)\ \left(1 + (P/k_i)^{\alpha}\right)} \tag{4}$$

$$\frac{dS}{dt} = -V' \tag{5}$$

$$\frac{dP}{dt} = V' \tag{6}$$

Where, k_s' is the substrate affinity constant with Co^{2+}, mM. The differential equations were solved by using the Runge–Kutta method. The model parameters were obtained by optimization using a genetic algorithm (GA) in minimizing the errors between the model predictions and the measured

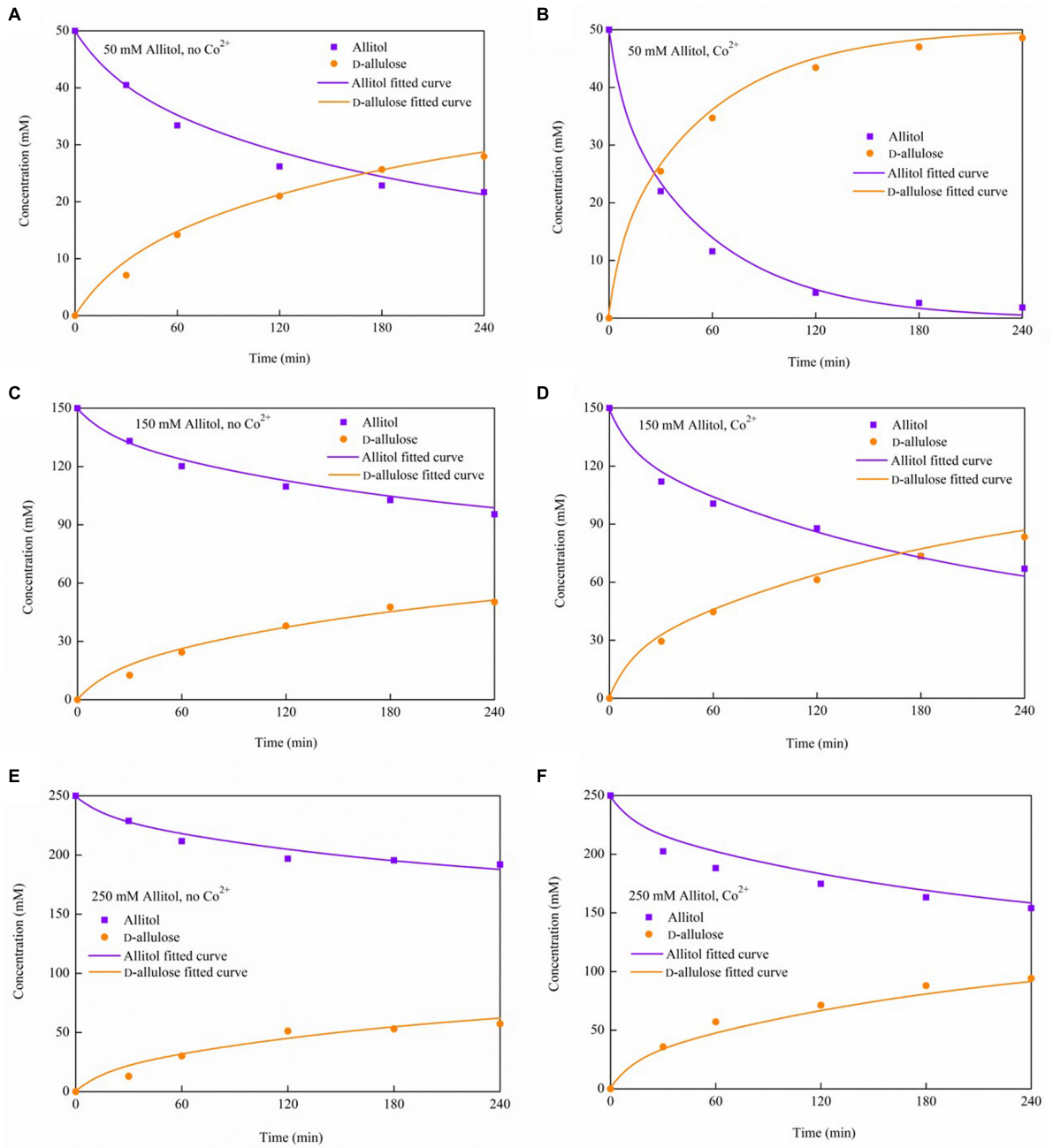

FIGURE 7 | The time courses of allitol and D-allulose concentrations during the biotransformation in the presence or absence of Co^{2+} under various initial allitol concentrations. **(A)** 50 mM allitol, no Co^{2+}; **(B)** 50 mM allitol, Co^{2+}; **(C)** 150 mM allitol, no Co^{2+}; **(D)** 150 mM allitol, Co^{2+}; **(E)** 250 mM allitol, no Co^{2+}; and **(F)** 250 mM allitol, Co^{2+}.

data, and the optimization diagram is shown in **Figure 8**. GA is the optimization algorithm that imitates the biological evolutionary processes, which is efficient in solving sophisticated and nonlinear problems. In optimization of the parameter values using GA, one chromosome codes for five genes, and one gene codes for one parameter value as shown in **Figure 8**. After repeated rounds of biological operations of selection, hybridization (crossover), and mutation until reaching the default termination criteria, the most-fitted chromosome coding for the parameters was obtained to get the optimized parameter values (**Figure 8**). MatLab 2020b (MathWorks, Inc., United States) running on Windows-compatible personal computer was used in the simulation and model parameter optimization. The optimized model parameter values are shown in **Table 2**. By using Equations (1)–(6) and the parameter values listed in **Table 2** as well as the initial values of allitol

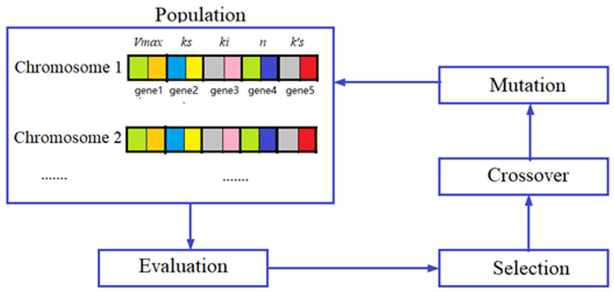

FIGURE 8 | Diagram of the genetic algorithm in parameter value optimization.

TABLE 2 | Model parameter values.

Parameter	Value
V_{max}	3.602 mmol/L/h
ks	321.809 mM
ki	10.740 mM
α	1.326 (–)
k's	21.782 mM

concentrations of 50, 150, and 250 mM, respectively, and the initial value of the D-allulose concentration of 0 mM, computer simulation of the biotransformation processes was made and the results are shown by the lines in **Figure 7**. It showed that the model predictions fitted the experimental data well. It also indicated that the substrate affinity coefficient was much decreased after Co^{2+} addition. Bulut et al. (2020) studied the effect of metal ions on the activity of 10 NAD-dependent formate dehydrogenases and found that there was a clear trend that many metal ions decreased the *Km* values of some FDHs using formate as the substrate, and they estimated that the metal ions could change the protein structure, and the interaction between the substrate or NAD(H) cofactor and the enzyme active site. Therefore, we speculated that the decrease of substrate affinity coefficient after Co^{2+} addition could be the result of the changes of the ADH enzyme structure or the interaction between the substrate of allitol and the active sites of the ADH enzyme. The modeling and simulation results showed that there was product inhibition so that the substrate was hardly completely consumed except in the case at the lowest substrate concentration of 50 mM and at a high enzyme activity with Co^{2+} addition, in which case, the allitol was nearly completely consumed (**Figure 7**). The modeling and simulation work provided numerical results for the reaction process, which are useful in process analyses and optimizations.

In the conventional method of kinetic modeling, the parameter values of the kinetic equation, like the Michaelis–Menten equation, are first obtained by using double-reciprocal linear plotting. And then, the differential equations are solved for the prediction of the reaction progress. In many cases, the predictions are quite different from the experimental measurements, which indicate that the parameter values obtained this way were not accurate. Therefore, a different method by optimization utilized GA was used in this work, which ensures the accurate prediction of the reaction process. The method using GA was ever successfully applied by us (Lin et al., 2004) and other researchers (Dutta et al., 2005; Yarsky, 2021) in the parameter optimization of the biological models.

CONCLUSION

In this study, the gene of NAD(P)-dependent ADH from *G. frateurii* NBRC 3264 was cloned and expressed in *E. coli* BL21 star. The expressed enzyme was purified and was identified for the first time to transform D-allulose from allitol. The effects of pH, temperature, and metal ions on the enzyme activity were determined, and Co^{2+} was found to have a high activation effect on the ADH. A high conversion yield of D-allulose of 97% was obtained at 50 mM allitol with Co^{2+} addition. The kinetics were investigated by modeling and simulation, and product inhibition was found. The enzyme showed enormous potential for application in the high-yield bioconversion of D-allulose and was expected to be applied to the industrial production of D-allulose.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

XW, JianqiL, JianquL, and XS designed the experiments. XW performed the experiments. HL, YN, GL, YR, CL, and CZ gave support to the experiments. XW and JianqiL built the kinetic model, analyzed the data, and wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This research was funded by the China Postdoctoral Science Foundation (2021M701994), the Research and Development Project of China (2017YFC1701502, 2017YFC1701504, and 2017YFC1702701), the Major Increase and Decrease Projects of the Central Government (2060302), the Jinan Agricultural Application Technology Innovation Plan (CX202112), and Ecological Planting and Quality Assurance Project of Genuine Medicinal Materials.

ACKNOWLEDGMENTS

We thank Caiyun Sun from the Core Facilities for Life and Environmental Sciences, State Key Laboratory of Microbial Technology, for help and guidance in the experiments.

REFERENCES

Bartsch, S., Brummund, J., Köpke, S., Straatman, H., Vogel, A., and Schürmann, M. (2020). Optimization of alcohol dehydrogenase for industrial scale oxidation of lactols. *Biotechnol. J.* 15:2000171. doi: 10.1002/biot.202000171

Bulut, H., Valjakka, J., Yuksel, B., Yilmazer, B., Turunen, O., and Binay, B. (2020). Effect of metal ions on the activity of ten NAD-dependent formate dehydrogenases. *Protein J.* 39, 519–530. doi: 10.1007/s10930-020-09924-x

DangThu, Q., Nguyen, T. T., Jang, S. H., and Lee, C. (2021). Molecular cloning and biochemical characterization of a NAD-dependent sorbitol dehydrogenase from cold-adapted *Pseudomonas mandelii*. *FEMS Microbiol. Lett.* 368:fnaa222. doi: 10.1093/femsle/fnaa222

Dutta, J. R., Dutta, P. K., and Banerjee, R. (2005). Modeling and optimization of protease production by a newly isolated *Pseudomonas* sp. using a genetic algorithm. *Process. Biochem.* 40, 879–884. doi: 10.1016/j.procbio.2004.02.013

EdyLiani, D., Yurnaliza, Y., and Saksono, B. (2020). Purification and characterization of D-psicose 3 epimerase (DPEase) from *Escherichia coli* BL21 (DE3) pET21b *dpe*. *Pak. J. Biol. Sci.* 23, 561–566. doi: 10.3923/pjbs.2020.561.566

El-Kabbani, O., Darmanin, C., and Chung, R. P. T. (2004). Sorbitol dehydrogenase: structure, function and ligand design. *Curr. Med. Chem.* 11, 465–476. doi: 10.2174/0929867043455927

Gullapalli, P., Takata, G., Poonperm, W., Rao, D., Morimoto, K., Akimitsu, K., et al. (2007). Bioproduction of D-psicose from allitol with *Enterobacter aerogenes* IK7: a new frontier in rare ketose production. *Biosci. Biotechnol. Biochem.* 71, 3048–3054. doi: 10.1271/bbb.70450

Hassanin, H. A. M., Mu, W., Koko, M. Y. F., Zhang, T., Masamba, K., and Jiang, B. (2016). Allitol: production, properties and applications. *Int. J. Food Sci. Technol.* 52, 91–97. doi: 10.1111/ijfs.13290

Izumori, K. (2006). Izumoring: a strategy for bioproduction of all hexoses. *J. Biotechnol.* 124, 717–722. doi: 10.1016/j.jbiotec.2006.04.016

Kim, N. H., Kim, H. J., Kang, D. I., Jeong, K. W., Lee, J. K., Kim, Y., et al. (2008). Conversion shift of D-fructose to D-psicose for enzyme-catalyzed epimerization by addition of borat. *Appl. Environ. Microbiol.* 74, 3008–3013. doi: 10.1128/AEM.00249-08

Kim, S. E., Kim, S. J., Kim, H. J., and Sung, M. K. (2017). D-Psicose, a sugar substitute, suppresses body fat deposition by altering networks of inflammatory response and lipid metabolism in C57BL/6J-ob/ob mice. *J. Funct. Food.* 28, 265–274. doi: 10.1016/j.jff.2016.11.029

Li, C., Lin, J., Guo, Q., Zhang, C., Du, K., Lin, H., et al. (2018). D-psicose 3-epimerase secretory overexpression, immobilization, and d-psicose biotransformation, separation and crystallization. *J. Chem. Technol. Biotechnol.* 93, 350–357. doi: 10.1002/jctb.5360

Li, Y., Shi, T., Han, P., and You, C. (2021). Thermodynamics-driven production of value-added D-allulose from inexpensive starch by an in vitro enzymatic synthetic biosystem. *ACS Catal.* 11, 5088–5099. doi: 10.1021/acscatal.0c05718

Lin, J., Lee, S., and Koo, Y. (2004). Model development for lactic acid fermentation and parameter optimization using genetic algorithm. *J. Microbiol. Biotechnol.* 14, 1163–1169.

Machielsen, R., Leferink, N. G. H., Hendriks, A., Brouns, S. J. J., Hennemann, H.-G., Dauβmann, T., et al. (2008). Laboratory evolution of *Pyrococcus furiosus* alcohol dehydrogenase to improve the production of (2S,5S)-hexanediol at moderate temperatures. *Extremophiles* 12, 587–594. doi: 10.1007/s00792-008-0164-8

Magnusson, A. O., Szekrenyi, A., Joosten, H. J., Finnigan, J., Charnock, S., and Fessner, W. D. (2019). nanoDSF as screening tool for enzyme libraries and biotechnology development. *FEBS J.* 286, 184–204. doi: 10.1111/febs.14696

Maria-Solano, M. A., Romero-Rivera, A., and Osuna, S. (2017). Exploring the reversal of enantioselectivity on a zinc-dependent alcohol dehydrogenase. *Org. Biomol. Chem.* 15, 4122–4129. doi: 10.1039/c7ob00482f

Poonperm, W., Takata, G., Ando, Y., Sahachaisaree, V., Lumyong, P., Lumyong, S., et al. (2007). Efficient conversion of allitol to D-psicose by *Bacillus pallidus* Y25. *J. Biosci. Bioeng.* 103, 282–285. doi: 10.1263/jbb.103.282

Wagner, N., Bosshart, A., Wahler, S., Failmezger, J., Panke, S., and Bechtold, M. (2015). Model-based cost optimization of a reaction-separation integrated process for the enzymatic production of the rare sugar D-psicose at elevated temperatures. *Chem. Eng. Sci.* 137, 423–435. doi: 10.1016/j.ces.2015.05.058

Wen, X., Lin, H., Ren, Y., Li, C., Zhang, C., Lin, J., et al. (2022). Allitol bioproduction by recombinant *Escherichia coli* with NADH regeneration system co-expressing ribitol dehydrogenase (RDH) and formate dehydrogenase (FDH) in individual or in fusion. *Electron. J. Biotechnol.* 55, 91–98. doi: 10.1016/j.ejbt.2021.11.007

Wen, X., Lin, H., Ren, Y., Li, C., Zhang, C., Song, X., et al. (2020a). Efficient allitol bioproduction from d-fructose catalyzed by recombinant *E. coli* whole cells, and the condition optimization, product purification. *Appl. Biochem. Biotechnol.* 192, 680–697. doi: 10.1007/s12010-020-03359-x

Wen, X., Lin, H., Ren, Y., Li, C., Zhang, C., Song, X., et al. (2020b). Optimization for allitol production from d-glucose by using immobilized glucose isomerase and recombinant *E. coli* expressing d-psicose-3-epimerase, ribitol dehydrogenase and formate dehydrogenase. *Biotechnol. Lett.* 42, 2135–2145. doi: 10.1007/s10529-020-02917-x

Yarsky, P. (2021). Using a genetic algorithm to fit parameters of a COVID-19 SEIR model for US states. *Math. Comput. Simul.* 185, 687–695. doi: 10.1016/j.matcom.2021.01.022

Zhang, W., Fang, D., Zhang, T., Zhou, L., Jiang, B., and Mu, W. (2013). Characterization of a metal-dependent d-psicose 3-epimerase from a novel strain, *Desmospora* sp. 8437. *J. Agric. Food Chem.* 61, 11468–11476. doi: 10.1021/jf4035817

Zhang, W., Li, H., Zhang, T., Jiang, B., Zhou, L., and Mu, W. (2015). Characterization of a d-psicose 3-epimerase from *Dorea* sp. CAG317 with an acidic pH optimum and a high specific activity. *J. Mol. Catal. B Enzym.* 120, 68–74. doi: 10.1016/j.molcatb.2015.05.018

Zhao, J., Wei, H., Chen, J., Li, L., Li, K., and Liu, J. (2021). Efficient biosynthesis of D-allulose in *Bacillus subtilis* through D-psicose 3-epimerase translation modification. *Int. J. Biol. Macromol.* 187, 1–8. doi: 10.1016/j.ijbiomac.2021.07.093

Zheng, Y. G., Yin, H. H., Yu, D. F., Chen, X., Tang, X. L., Zhang, X. J., et al. (2017). Recent advances in biotechnological applications of alcohol dehydrogenases. *Appl. Microbiol. Biotechnol.* 101, 987–1001. doi: 10.1007/s00253-016-8083-6

Zhu, Y., Li, H., Liu, P., Yang, J., Zhang, X., and Sun, Y. (2015). Construction of allitol synthesis pathway by multi-enzyme coexpression in *Escherichia coli* and its application in allitol production. *J. Ind. Microbiol. Biotechnol.* 42, 661–669. doi: 10.1007/s10295-014-1578-1

Zhu, Y., Men, Y., Bai, W., Li, X., Zhang, L., Sun, Y., et al. (2012). Overexpression of d-psicose 3-epimerase from *Ruminococcus* sp. in *Escherichia coli* and its potential application in d-psicose production. *Biotechnol. Lett.* 34, 1901–1906. doi: 10.1007/s10529-012-0986-4

Zhu, Z., Li, C., Liu, X., Gao, D., Wang, X., Tanokura, M., et al. (2019c). Biochemical characterization and biocatalytic application of a novel d-tagatose 3-epimerase from *Sinorhizobium* sp. *RSC Adv.* 9, 2919–2927. doi: 10.1039/c8ra10029b

Zhu, Z., Gao, D., Li, C., Chen, Y., Zhu, M., Liu, X., et al. (2019b). Redesign of a novel D-allulose 3-epimerase from Staphylococcus aureus for thermostability and efficient biocatalytic production of D-allulose. *Microb. Cell Fact.* 18, 59. doi: 10.1186/s12934-019-1107-z

Zhu, Y. H., Liu, C. Y., Cai, S., Guo, L. B., Kim, I. W., Kalia, V. C., et al. (2019a). Cloning, expression and characterization of a highly active alcohol dehydrogenase for production of ethyl (S)-4-chloro-3-hydroxybutyrate. *Indian J. Microbiol.* 59, 225–233. doi: 10.1007/s12088-019-00795-0

Conflict of Interest: YR was employed by the company Qingdao Longding Biotech Limited Company.

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.

frontiers | Frontiers in Microbiology

REVIEW
published: 26 May 2022
doi: 10.3389/fmicb.2022.925454

Machine Learning Advances in Microbiology: A Review of Methods and Applications

Yiru Jiang, Jing Luo, Danqing Huang, Ya Liu and Dan-dan Li**

State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao, China

Microorganisms play an important role in natural material and elemental cycles. Many common and general biology research techniques rely on microorganisms. Machine learning has been gradually integrated with multiple fields of study. Machine learning, including deep learning, aims to use mathematical insights to optimize variational functions to aid microbiology using various types of available data to help humans organize and apply collective knowledge of various research objects in a systematic and scaled manner. Classification and prediction have become the main achievements in the development of microbial community research in the direction of computational biology. This review summarizes the application and development of machine learning and deep learning in the field of microbiology and shows and compares the advantages and disadvantages of different algorithm tools in four fields: microbiome and taxonomy, microbial ecology, pathogen and epidemiology, and drug discovery.

Keywords: microorganisms, machine learning, deep learning, prediction, classification

OPEN ACCESS

Edited by:
Pengfei Ding,
University of Maryland,
Baltimore County, United States

Reviewed by:
Changyou Chen,
Institute of Electrical Engineering
Chinese Academy of Sciences, China
Bingyao Huang,
Guangxi University of Chinese
Medicine, China
Qing Zhang,
Shandong Academy of Agricultural
Sciences, China

****Correspondence:***
Ya Liu
liuya@sdu.edu.cn
Dan-dan Li
dandanli@sdu.edu.cn

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

Received: *21 April 2022*
Accepted: *09 May 2022*
Published: *26 May 2022*

Citation:
Jiang Y, Luo J, Huang D, Liu Y and Li D-d (2022) Machine Learning Advances in Microbiology: A Review of Methods and Applications. Front. Microbiol. 13:925454. doi: 10.3389/fmicb.2022.925454

INTRODUCTION

Microbiology focuses on studying the activity law of microorganisms, exploring the characteristics, culture conditions, and detection methods of microflora, taking its essence (discovering, utilizing, improving, and protecting beneficial microorganisms), and removing its dross (preventing, controlling, or transforming harmful microorganisms). Thus, it is available for science and benefits mankind (Dworkin, 2012; Hanage, 2014; Ha and Devkota, 2020).

Recently, the main research hotspots in microbiology include community classification and its environmental role (Bulgarelli et al., 2013; Zhang et al., 2021), regulation of gut microbiome and host interactions (Turnbaugh et al., 2007; Jones et al., 2014; Malla et al., 2018; Ruff et al., 2020), development of pathogenic microorganisms and drug vaccines (Shahbaaz et al., 2016; Moos et al., 2017; Zhu et al., 2020), and trying to dilute the boundaries between microbiome and genome editing, molecular modification, ecology and resource utilization, biocatalysis, and synthesis (Stres and Kronegger, 2019; Galloway-Pena and Hanson, 2020). In addition, microbiology and multiomics (including genomics, epigenomics, transcriptomics, proteomics, and metabolomics) have combined and developed a variety of multiscale emerging fields (Beck et al., 2021; Liang et al., 2021).

The understanding of microorganisms started from microbial cell morphology and physiological and biochemical characteristics to microbial genotype identification at the nucleic acid and protein levels, and chemical analysis methods based on cell chemical composition analysis and

numerical classification methods relying on the level of computational biology have also been established successively. The rapid progress in the discipline of microbiology is inseparable from the update of observation methods or techniques in the same period (Galloway-Pena and Hanson, 2020). With the advent of the Big Data era, the pressing questions for researchers have gradually evolved into how to quickly and efficiently filter/condense this exponential growth of information to obtain generalized quality data and how to transform the massive data of microbiota into easily understood and visualized knowledge. Compared to traditional research with insufficient data or purely experimental techniques that cause trouble, such as cognitive bias, low reproducibility, and long-time span, the modern microbiology research process is more likely to incorporate new technologies and big data methods to do this better and right.

Artificial intelligence (AI), first proposed by John McCarthy at the Dartmouth Conference in the summer of 1956, concentrates on the simulation of human intelligence extensions and the research and development of theoretical methods, techniques, and applied systems. The entry of AI drives the progress of microbiology and achieves a new paradigm breakthrough (Barredo Arrieta et al., 2020). Combined with the advantages of big data, automation, modeling, and AI, microbiology has evolved toward a multiscale and multidimensional direction, gradually applying to systems biomedicine, systems ecology, etc.

Machine learning (ML), first proposed by Arthur Samuel (Bell Labs, IBM, Stanford) in 1959, is a special branch/subfield of AI that aims to find features from large-scale heterogeneous data. The most basic thing is to use algorithms to parse the data, analyze the patterns in the data automatically, and then utilize these patterns to make predictions and decisions on real-world events (Jordan and Mitchell, 2015). Unlike traditional software programs that are hard-coded to solve specific tasks, ML takes large amounts of data and trains them using algorithms to learn how to accomplish tasks from the data (Domingos, 2012). With the integration of cross-scale and complex microbial communities and multiomics integration, ML can be used to systematically present interactions between microflora or with hosts. The workflow of dimensionality reduction and then extraction of spatial features from high-dimensional datasets generated from large data collections is supportive of exploring the functional potential of microorganisms and expanding the study of microbial technology applications.

Deep learning (DL) is a breakthrough ML approach that models high-level abstractions of data through a deep network with multiple layers of processing units, which are parametric models trained by gradient descent (Lecun et al., 2015). ML is a way to implement AI, and DL is a technology to implement ML (**Figure 1**). Remarkably, there is no obvious boundary separating DL from traditional ML and traditional statistical analysis. To handle complex, high-dimensional microbiome data, ML algorithms have been applied to the frontiers of combining microbiome and computational science, more commonly for classification and prediction (Schmidhuber, 2015).

This paper first briefly introduces the ML methods, data processing steps, and algorithms commonly used in microbial research, summarizes the research on ML-based microbial prediction and application, and discusses the advantages and limitations of the methods and tools, demonstrating the development prospects of computational microbiology from the perspective of ML.

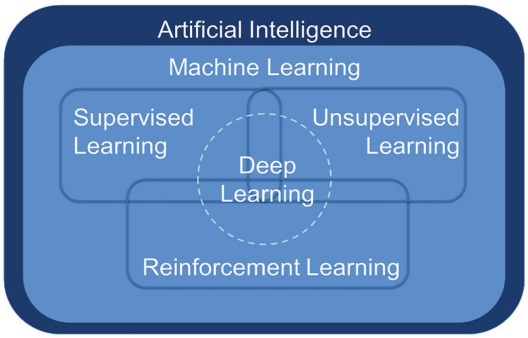

FIGURE 1 | The relationship among artificial intelligence, machine learning, and deep learning.

MACHINE LEARNING

An AI system is supposed to be equipped to learn knowledge from raw data, which is known as ML. Effective features are extracted from raw data by designing targeted pattern recognition algorithms and then using these features with ML algorithms, i.e., distance functions to represent pairwise relationships between objects. The earliest ML algorithms can be traced back to the early 20th century, and a large number of classical methods have been developed within these 100 years (**Figure 2**). This section summarizes the classical algorithms that have appeared in history in four directions: supervised learning, unsupervised learning, DL, and reinforcement learning (RL). Then, we elaborate on the criteria for evaluating the merits of the model and algorithmic workflows.

Supervised Learning

Supervised learning, including regression analysis and statistical classification, refers to a class of methods that use samples from known categories as training sets to train models. Before the concept of ML was introduced, Fisher (1936) invented a supervised data dimensionality reduction algorithm, linear discriminant analysis (LDA). In the 1950s, based on the core idea of Bayes decision theory, which is to select the decision with the highest probability, the Bayes classifier was born and divides the sample into the class with the highest posterior probability. The naive Bayes (NB) model has a simple algorithm with stable classification efficiency, performs well for small-scale data, can handle multiple classification tasks and is suitable for incremental training (Zhang et al., 2009); however, it is required to decide the probability of the posterior by virtue of the prior and data before classification determination. Thus, there is a certain error rate in the classification decision-making,

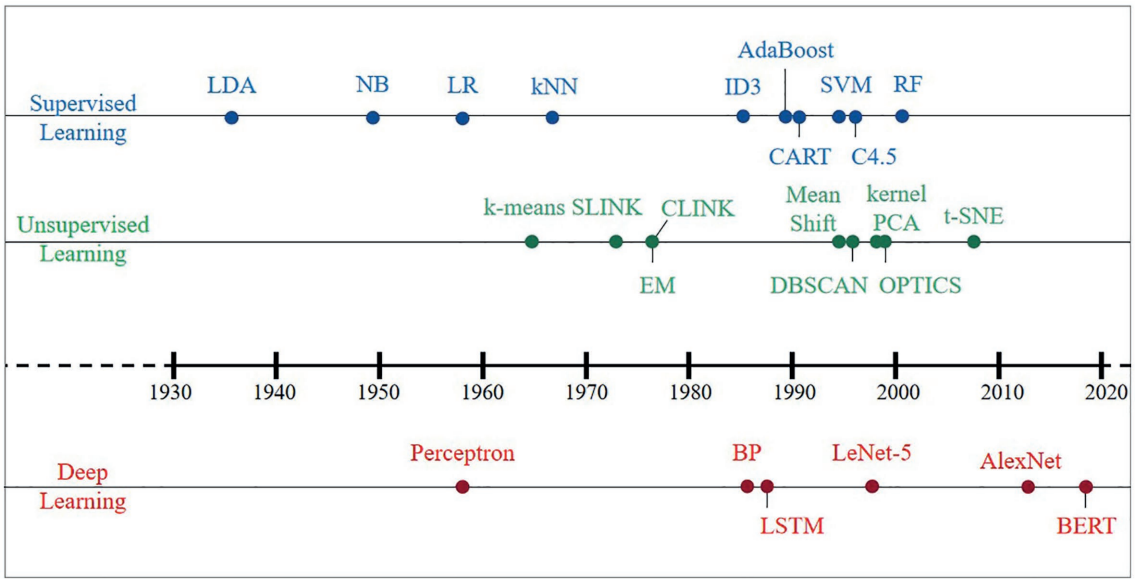

FIGURE 2 | Development history of classical machine learning algorithms since the 1930s.

and it is sensitive to the expression of the input data. Logistic regression (LR) directly predicts the probability of a sample belonging to a positive sample, with a clear model, strong parameter interpretability, and simple and efficient for big data scenarios; however, its performance is easily affected by the correlation between features and the size of the feature space, and it is prone to underfitting problems, resulting in low accuracy (Cox, 1958). The k-nearest neighbor (kNN) algorithm is considered an algorithm based on the idea of template matching that is simple and efficient and can solve both classification and regression problems with high accuracy and insensitivity to outliers; however, its prediction speed is slower than that of LR, especially for dealing with high-dimensional data, which is computationally intensive (Cover and Hart, 1967).

The year 1980 serves as a transition point in the history of ML algorithms, which gradually developed from fragmented and unsystematic enlightenment algorithms into an independent and systematic direction. Various machine learning algorithms have exploded and developed rapidly. In the 1980s and early 1990s, three typical implementations of decision trees (DT): ID3 (Quinlan, 1986), CART (Yeh, 1991), and C4.5 (Quinlan, 1996), had fast computation, high accuracy, and high interpretability, which make DT still used in some problems today, but their characteristic of easy-overfitting leads to easy neglect of the relevance of attributes in the dataset. Two classical algorithms, support vector machine (SVM) based on statistical learning theory (Cortes and Vapnik, 1995) and AdaBoost (Freund, 1990), were developed in the 1990s. The former (SVM) uses kernel functions that can be mapped to a high-dimensional space to solve nonlinear classification problems with uncomplicated classification ideas (maximizing the interval between samples and decision surfaces) and presents better classification performance; however, the method is difficult to solve the multiclassification problem, sensitive to missing data, and thus challenging to achieve large-scale training samples. The latter (AdaBoost) can integrate the use of simple weak classifiers, which does not require either *a priori* knowledge of weak classifiers or filtering of features, and can significantly improve learning accuracy regardless of whether the data are artificial or real; nevertheless, it is susceptible to noise interference and has a long training time.

The random forest (RF) and AdaBoost algorithms belong to integrated learning, with high accuracy, and can effectively run on large datasets and strong resistance to noise (Breiman, 2001); however, the number of decision trees will lead to a very long training time, and overfitting occurs in noisy classification or regression problems. Up until the rise of DL in 2012, supervised learning was rapidly developed, and various ideas and methods emerged one after another, yet no one ML algorithm achieved an overwhelming advantage.

Unsupervised Learning

Unsupervised learning is a method to learn the commonality in the input data to determine whether such commonality exists in the new data, and the research thinking can be divided into two categories: clustering and data dimensionality reduction. The hierarchical clustering algorithm emerged early (Ward, 1963), and some of its implementations are still in use today, including SLINK (Sibson, 1973) and CLINK (Defays, 1977). The K-means clustering algorithm was then born, and the algorithm is simple and easy to implement (Macqueen, 1965), whereas there are the following drawbacks: (1) the number of class clusters needs to be specified by the user in advance; (2) the clustering results are more sensitive to the selection of the initial class cluster centers; (3) it is easy to fall into a local optimum; and (4) only spherical class clusters can be found;

since then, it has been continuously improved and grows into the clustering algorithm with the most variants and improvements. The expectation-maximum (EM) algorithm (Dempster et al., 1977) has been used to solve various extreme likelihood estimation problems in ML with missing data and is commonly used to learn the variational inference of LDA topic models, parameters of the Gaussian mixture model (GMM), and hidden Markov model (HMM). Other density-based clustering algorithms in the 1990s include mean shift (Cheng, 1995), density-based spatial clustering of applications with noise (DBSCAN) algorithm (Ester et al., 1996), and ordering points to identify the clustering structure (OPTICS) algorithm (Ankerst et al., 1999). They are not based on various distances but on density.

A new idea of clustering was born in the early 21st century: transforming the clustering problem into the graph cutting problem, and the representative algorithm covering this new idea is spectral clustering. The data dimension reduction algorithm originated very early, and the advantages of the classic principal component analysis (PCA) algorithm are the complete absence of parameter restrictions, the removal of data redundancy and noise, the compression and preprocessing of the data to make the dataset easier to use, and the results easier to understand (Pearson, 1901). PCA can eliminate the correlation between variables, but the nonlinear dependence between samples may be lost if linear dimensionality reduction is performed *via* PCA. The heavyweight result innovation, kernel PCA (Scholkopf et al., 1998), was based on the kernel technique, combined with PCA and transforming PCA into a nonlinear dimensionality reduction algorithm. Since then, a wave of nonlinear methods has been set in motion, e.g., locally linear embedding (LLE), Laplacian eigenmaps, locality preserving projections, and isometric mapping (Roweis and Saul, 2000; Tenenbaum et al., 2000; Belkin and Niyogi, 2003; He and Niyogi, 2003). Then, *t*-distributed stochastic neighbor embedding (t-SNE) was developed (Van Der Maaten and Hinton, 2008), mainly for visualizing and exploring high-dimensional data, which follows nonlinearity and has the best visualization effect compared with other dimensionality reduction algorithms. The relative similarity of the original data at the time of dimensionality reduction is excellent; however, the results of each run will change slightly for each run due to its random nature. Unsupervised learning, although relatively slow in development and with few breakthroughs, has occupied a dominant role in human and animal learning and is a necessary path to explore strong artificial intelligence.

Deep Learning

Deep learning, compared to traditional ML, is more highly dimensional and targeted to capture as many/complete relationships as possible in the raw data. DL can be classified into supervised, unsupervised and hybrid DL models according to whether labeled data are required or not, where hybrid models usually refer to the use of unsupervised model results as input data or important auxiliary to supervised models. The predecessor and technical essence of DL is artificial neural networks (ANNs). In 1958, the predecessor of ANN, the Perceptron model, was launched (Rosenblatt, 1958), but it was not of practical value because it was too simple and could only handle linear classification problems, not even solving the XOR problem. Therefore, it does not have practical value but mainly lays the ideological foundation for the later algorithms. Research on neural networks entered a bottleneck until the 1980s, for instance, the back propagation (BP) algorithm for training multilayer neural networks/multilayer perceptrons using sigmoid functions for nonlinear mapping (Rumelhart et al., 1986). Based on the forward propagation of traditional neural networks, the BP algorithm adds a backward propagation process of errors, continuously adjusting the weights and thresholds between neurons until the output error reaches a reduction to within the allowed range or reaches a predetermined number of training times. It effectively solves the problem of nonlinear classification and learning and is the basis for improving and applying neural networks.

However, as the scale of the neural network increases, the BP algorithm suffers from the problem of "gradient disappearance." Meanwhile, the limited hardware level of computers led to poor computing power, which could not help the further development of BP algorithm, plus the effect of classification and regression application of shallow ML such as SVM in the same period was continuously proved, and DL thus entered the second winter period. Even during the winter period, algorithms such as convolutional neural networks (CNN) and long short-term memory (LSTM) were developed and are still adopted today to process vision tasks (Lecun et al., 1989). Among them, LeNet-5 was proposed by Lecun et al. (1998) and has become the prototype of most deep convolutional neural networks (DCNNs).

Until Hinton and Salakhutdinov (2006) proposed the concept of DL, the problem of "gradient disappearance" was solved, i.e., the algorithm was trained layer by layer by unsupervised learning and then tuned using a supervised back-propagation algorithm. Hinton and his student Alex Krizhevsky used AlexNet to win the ImageNet competition (Smirnov et al., 2013), which became the pioneer of the current wave of deep learning. Its top 5 accuracy rate of 84.6% has an error rate of only 15.3%, and the network is characterized by (1) the use of the ReLU method to speed up training; (2) the use of dropout to prevent overfitting; and (3) GPU parallel computing technology to solve the problem of long optimization time for deep networks with many parameters. Moreover, some neural network architectures, such as variational autoencoders (VAEs) and generative adversarial networks (GANs), have recently attracted much attention in the DL community. The bidirectional encoder representation from transformers (BERT) model proposed by Devlin et al. (2019) has built a transformer network structure with a self-attention mechanism as the core. Excellent performance is presented in many tasks in natural language processing (NLP) due to its versatility. Essentially, DL is a statistical technique with advantages and limitations that are maturing in the areas of computer vision, natural language processing, and speech recognition.

Reinforcement Learning

Reinforcement learning is a special class of ML algorithms, the most important feature of which is learning from interaction (Keerthi and Ravindran, 1994; Kaelbling et al., 1996). On the basis of interaction, we constantly judge whether the action is related to the goal, corresponding to the generation of rewards or penalties, and repeatedly execute it to finally maximize the expected benefits, an "automatic scoring and escalation" process. Deep reinforcement learning (DRL), a new research hotspot, combines the perceptual capability of deep learning with the decision-making capability of reinforcement learning to achieve direct control from raw input to output through end-to-end learning for applications in robot control, computer vision, natural language processing, and medical care (Erev and Roth, 1998; Frank et al., 2004; Kober et al., 2013; Mnih et al., 2015).

Evaluation Criteria and Algorithmic Workflows

Different algorithms have their own advantages and disadvantages, and there is no superiority or inferiority. What needs to be done is to fully interpret the input data based on different demand scenarios and then build suitable models to continuously adjust to achieve the best performance. Moreover, the belief that "as long as the most advanced and complex model is used, the scientific problem will be solved" is not objective. In essence, computer technology only assists people in making decisions or automates the human decision-making process and improves efficiency. Therefore, the choice of model should be the most suitable one, rather than pursuing the most complex one. There are four criteria used to judge the merits of machine learning algorithms (Greener et al., 2022). (1) Correctness, the most important criterion for judging the merits of an algorithm. (2) Robustness, i.e., fault tolerance, representing the algorithm's ability to respond to and address illegal data input. (3) Readability, easy-to-understand algorithms means a less time-consuming process of debugging, modification, and expansion. (4) Temporality, i.e., time complexity and space complexity, represent the computational effort and memory space required to execute the algorithm, respectively.

The use of ML as a technical tool to solve scientific problems can generally comply with the following five steps in **Figure 3** (Greener et al., 2022). (1) Define the problem, prepare and process the data, and determine the assessment method. The data were split into three groups: training set, validation set, and test set. The training set is given to build the model, the validation set and the test set both refer to the data samples retained when training the model, and the ability of the model to use the training data should be evaluated successively. The data also undergo targeted preprocessing before use, such as vectorization, value normalization, and feature engineering needed for non-DL. Then, we select the most representative evaluation metrics and validate the evaluation method for the problem. Commonly used performance metrics are confusion matrix, precision, recall, specificity, F1 score, precision-recall curve, ROC, AUC, etc. Common evaluation methods include simple leave-out validation, *k*-fold cross-validation, repeated k-fold validation with disrupted data, and bootstrapping. (2) Build the model. Develop models that are more optimized than the benchmark, with the ultimate goal of balancing the dichotomy between optimization and generalization: find the line between underfitting and overfitting and maximize generalization capabilities. (3) Validating the model. Models with statistical efficacy tend to require scaling up the model first, and a threshold of overfitting for monitoring training losses and validation losses will be required. (4) Testing the model. The goal is to evaluate the predictive capability of the model in completely new data, as opposed to validating the data. It is essential to evaluate all aspects of the model, for instance, to check whether the output of the program meets the expected correct values and whether the model results meet the expected evaluation requirements (accuracy or error). (5) Tuning the model. Boosting the performance of the algorithm with more data, different features, or tuned parameters. The previous steps are repeated continuously, with model regularization and tuning of hyperparameters (parameters to control the behavior of the algorithm when building the model) depending on the performance of the model on the validation set until the desired performance is achieved.

Machine learning methods tend to require a combination of mathematical knowledge concerning statistical probability theory, linear algebra, and algorithmic complexity theory, combined with the diversity of microbial data, which makes it intractable for researchers in the field of microbiology to construct and utilize complex ML models independently. In response to the nature and volume of experimental data specific to various research directions, experts in big data science propose ideas and technical support on approaches to leveraging existing data for effective ML, facilitating the emergence of new cross-cutting areas. With the widespread adoption of ML and DL algorithms, humans have been presented with a whole new world of microorganisms, especially in the fields of classification and prediction.

CLASSIFICATION AND PREDICTION

Next, we will characterize the impact of ML on the microbiology field and specific application cases. The application of ML in microbial species and community classification and prediction mainly includes microbiome and taxonomy, microbial ecology, pathogen and epidemiology, and drug discovery (with a particular focus on antibiotics/antimicrobial peptides).

Microbiome and Taxonomy

The microbiome refers to an ecological community of microorganisms with different characteristics and functions that coexist in a given environment, including the genomes and environmental/habitat conditions of the members (Lederberg and McCray, 2001; Berg et al., 2020). The application usually combines one or more of the multiomics techniques, which to some extent is more accurate and precise in classification than single-omics data studies and facilitates the exploration of the influential factors in microbiomics network mechanisms.

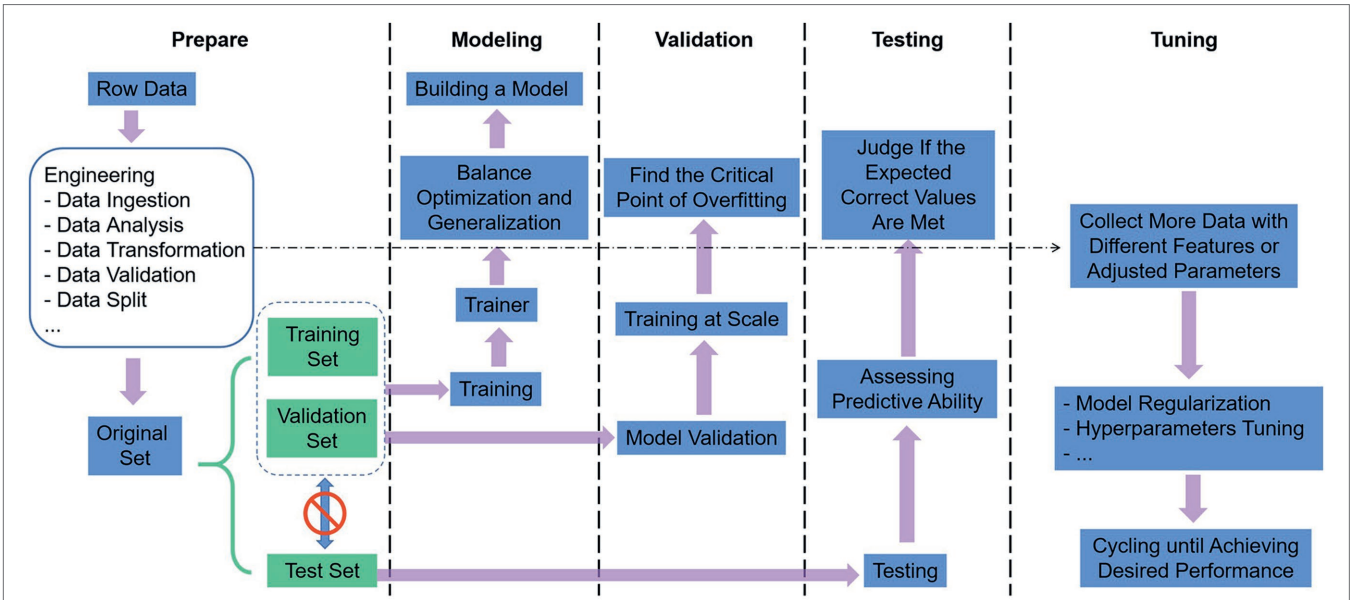

FIGURE 3 | Machine learning workflow.

The strategy of microbial taxonomy is to distribute target sequences to microbial communities at different taxonomic levels utilizing various tools (Marchesi and Ravel, 2015). ML, especially classification and clustering algorithms, allows classification based on data representing the characteristics/functionality of the target community, reflecting similar relationships. With upgrades in sequencing technology, hundreds of millions of short sequencing reads have been generated from merely a single sample, which consequently generates high-dimensional microbiome data (Luz Calle, 2019). Therefore, linear or nonlinear dimensionality reduction algorithms are advantageous in handling complex and multivariate sparse microbiome data to achieve dimensionality reduction (Kostic et al., 2015) and visualization (Song et al., 2019) of the data space.

Common supervised classification algorithms are particularly valuable in identifying highly complex datasets, as in the case of human microbiota surveys (Knights et al., 2011). ML and statistical techniques are in place to build predictive models of taxonomic units (Knights et al., 2011) or functions (White et al., 2009) to distinguish between distinct sample groups. The selection of classification techniques requires considering the characteristics of different microbial communities and extracting the data with different features, encoding the extracted data with labels, and rendering them available for model training (Knights et al., 2011). A study as early as 2012 demonstrated that random forests enable effective and accurate classification of human microbial community samples and allow the identification of key components (OTUs or species) that differentiate between groups (Knights et al., 2011; Yatsunenko et al., 2012). A series of base classifiers are trained separately and independently, and the results of each base classifier training are integrated by adopting a certain rule. This is the idea of ensemble learning, which will obtain better classification results than a single classifier (Wang et al., 2007; Wu and Zhang, 2008). Subsequent studies have targeted the oral microbiota in saliva and classified them with the algorithm of SVM, ANN, and DT (Nakano et al., 2014). Xu et al. (2020) constructed classifiers and classified new samples using LR, SVM and DT based on the dimensionality reduction space generated by t-SNE with Aitchison distance, compared the classification performance of the same classifiers in the original dimension and the dimensionality reduction space, and demonstrated that compared with the t-SNE dimensionality reduction technique using Euclidean distance, Aitchison distance increases the classification accuracy (ACC) of the classifier in general.

Unsupervised learning relies on the strategy of sequencing depth information or OTU clustering of sample data instead of known information (Sangwan et al., 2016). MetaBAT quantifies the similarity of sequences using sequence similarity and information about the sequencing depth of the sample data, using the calculation of the distance between overlapping clusters, and then clustering (Kang et al., 2015). COmposition, read CoverAge, CO-alignment, and paired-end read LinkAge (COCACOLA) calculates the distance with L distance instead of the traditional Euclidean distance (Lu et al., 2017). Strategies for OTU-based clustering inevitably take into account the setting of thresholds, feature extraction, and the choice of specific clustering methods. Cai and Sun (2011) proposed a hierarchical clustering method, i.e., first filtering a large number of unnecessary sequence comparisons with k-mers and then launching the hcluster algorithm in the clustering process to achieve similar accuracy as the standard hierarchical clustering algorithm.

The design of sequence classification methods based on deep learning is not rare. A study proposed a sequence classification technique based on k-mer and two DL architectures—CNN

to discriminate and deep belief network (DBN) to generate—for the bacterial taxonomy of macrogenomic data (Fiannaca et al., 2018). The ANN classifier can optimize the classification effectiveness and confidence of the target community after feature analysis. A study developed a pipeline (cell type recognition and CellCognize) based on multidimensional flow cytometry (FCM) data *via* ANN to enable quantification of cell type diversity and subsequent differentiation and classification of microbiota of known composition (Duygan et al., 2020). Composed of a feed-forward back-propagation algorithm, an input layer, a hidden layer, and an output layer, ANNs have been trained to classify either five or 32 standard multiparameter FCM datasets and forecast cell type attribution of FCM data from poorly trained microbial samples of known or unknown composition.

Given the characteristics of high dimensionality, multinoise, data sparsity, and heterogeneity of histological data, as well as the problem of unbalanced datasets in experiments, the integration of complex and large-scale histological data imposes high demands on the analysis capability of algorithmic models and computing platforms. Currently, the main methods are dimensionality reduction and noise reduction through PCA or autoencoder and transformation of sparse and heterogeneous data through regression methods. However, all of these methods have their drawbacks, and a substantial amount of research on these issues will be necessary in the future.

Microbial Ecology

Microbial ecology, with its origins in environmental microbiome studies, takes as its starting point the study of target microbiota, with the long-term goal of capturing the diversity of multiple species interactions (competition, predation, facilitation, and symbiosis), as well as uncovering their distribution patterns/networks and maintenance mechanisms. Mechanisms of microbiota–microbiota and host–microbiota interactions are critical to our understanding of microbial network structure and function of homeostasis in a given habitat (Broberg et al., 2018; Hassani et al., 2018; Van De Guchte et al., 2018). Advances and applications of new experimental and computational methods will drive the integration of microbial ecology research with leading technologies in integrated multiomics, computational biology, and other fields.

The purpose of constructing ecological networks is to model all interactions between species and their environment. Faisal et al. (2010) used four widely used statistical/ML methods, graphical Gaussian models (GGMs), L1-regularized regression with the least absolute shrinkage and selection operator (LASSO), sparse Bayesian regression (SBR), and Bayesian networks (BNs), to validate their usefulness in identifying community interactions in microecological networks. These methods enable simulated restoration of food web structure and contribute to modeling and predicting the effects of bioclimatic variables. However, since the complete ecological knowledge of the true interaction network between species is hardly visible, assessing the success of the modeling solely relies on known or possible relationships. Although pairwise interactions are the basis for the study of complex ecological networks, higher-order interactions involving three or more taxonomic units increase the variability and stochasticity in the study of community composition, on which the prediction of microbiota-associated biological phenotypes is based (De'ath and Fabricius, 2000).

It is necessary to simplify scientific problems by switching predictive strategies based on species characteristics when predicting relationships (natural diversity, life cycles, interactions, and coevolution) across species or with their environment. Leite et al. (2018) explored several machine learning techniques (kNN, RF, SVM, and ANN) to predict/identify the presence of a given phage–bacterial pair interaction after 10-fold cross-validation based on accuracy, F-score, specificity, and sensitivity criteria to filter the most predictive algorithms and their parameter values. The theoretical basis of its prediction lays in the interaction between a given phage–bacterial pair of encoded proteins, allowing the work to be converted into protein–protein interaction (PPI) prediction (Cusick et al., 2009). Accordingly, two features, the domain–domain interaction score and protein-level structural information, were selected in the feature extraction phase.

The intersection of genetics and ecology is established on the basis of the population concept. Stupp et al. (2021) proposed supervised ML-based phylogenetic profiling (MLPP) to predict functional interactions between human genes and the interaction environment in which they occur (i.e., biological functions) and established a web server containing functional interaction predictions for all human genes. They predicted the probability of all possible gene pairs in each tag using the lightGBM model, which is related to RF after comparing it with the DT, LR, and NB algorithms. Based on simulations and real data, Pichler et al. (2020) compared generalized linear models (GLM) with ML models (RF, boosted regression trees, deep neural networks, CNN, SVM, NB, and kNN) to measure their capability to predict species interactions based on traits and to extrapolate trait combinations that are causally relevant to species interactions. In a global crop–pollination database, they found that RF had the best predictive performance, predicting species interactions in plant-pollinator networks remarkably well.

However, the reality that most microbial species within communities are not culturable makes the prediction of interspecies interactions in natural microbial communities challenging. This comes from the fact that the accuracy of deep learning (especially deep neural networks) depends on the reliability of the training database. Moreover, there is still space for investigators to design joint experimental and modeling studies to uncover interaction mechanisms that have not yet been fully investigated (Lee et al., 2020).

Pathogen and Epidemiology

Epidemics have the characteristics of being contagious, epidemic, and recurrent. Infestations of previously unknown pathogenic microorganisms pose a continuous threat to food security and human health. To address the medical challenges of epidemiology, the identification, and characterization of pathogens, and the screening and prediction of diseases have emerged as major concerns for professional biomedical scientists. ML, as well as

DL, which dominates in batch image classification, has led to a significant reduction in the time and computational cost spent on dataset analysis due to its extremely efficient, cost-effective, accurate and high-throughput advantages (Ghosh et al., 2022).

Disease epidemiology studies examine the patterns of temporal and spatial dynamics of diseases at the population level under different environmental conditions. Research on issues such as diseases caused by plant and animal viruses provides a large dataset on gene expression, chromosome conformation, genetic variation, traits, and diseases. The relevance of the viral genome allows for screening with the help of macrogenomics. The application of ML enables the integration of multiomics data and significantly improves macrogenomic data analysis. ML assists in classifying these viral sequences and thus deepens our understanding of virus evolution. VirFinder is a k-mer-based platform to identify prokaryotic virus sequences from mixed macrogenomic data, accelerating the screening of pathogens at the genetic level in plant and animal virome studies (Ren et al., 2017). The synergistic application of ML and hyperspectral imaging (HSI) provides a new methodological idea for image detection of viral diseases. While the high-dimensional data generated by HSI contain redundant information, ML reduces the dimensionality of HSI data by determining the effective specific wavelength range through data preprocessing. For instance, multilayer perceptrons (MLPs), ANNs, and CNNs enable the detection and classification of color images by hidden image features with high accuracy of >96.00% (Lowe et al., 2017). Compared to traditional ML, which requires feature extraction techniques tailored to the nature of the data and model, DL supports automatic feature extraction, reducing computational time, and the burden of reliance on professional expertise. Training a model (classifier) with live images is a case in point (Ferentinos, 2018). This implies that determining the reliability of the classification relies to some extent on the abundance of available images based on the scene. For example, VGGNet, obtained by Chen et al. (2020), achieved an average accuracy of 92.00% for predicting rice plant image categories based on ImageNet and Inception module pre-training.

Phages are the most abundant organisms on Earth and have been considered as natural enemies of bacteria. Several ML algorithm models aiming to improve the automatic recovery and prediction of phages already exist. For instance, VirSorter searches the database of expected proteins up front using probabilistic similarity and reference homology to identify viral signals, but the disadvantage is that it does not fully distinguish between virus and nonvirus Pfam annotations (Roux et al., 2015). Kaelin et al. (2022) employed VirSorter v.1.0.5 to identify circular contigs of candidate viruses. Another tool, Meta-genomic Analysis and Retrieval of Viral Elements (MARVEL), which aggregates annotation and sequence signature information of previously identified phages, was developed to identify and predict double-stranded DNA phage sequences in macrogenomic boxes. Given the excellent recall, Braga et al. (2020) used MARVEL to identify phage bins for prediction. According to the authors' statement, comparing the performance of MARVEL, VirFinder, and VirSorter, all three performed comparably in terms of specificity, with MARVEL outperforming in terms of sensitivity (Amgarten et al., 2018). VIBRANT, the first computational tool to utilize neural networks and protein similarity methods, had a particularly strong performance score (94%) in the automatic recovery of microbial viruses, which was stronger than the first three (Kieft et al., 2020). Luo et al. (2022) filtered ≥ 1 kb contigs to identify viral contigs and related reads *via* VIBRANT. We summarize the available data and materials, which are shown in **Table 1**.

To date, most of the results generated from the intersection of pathogen research and machine learning in epidemiology have been prospective and feasible. Comparing different stages of classifier innovation, we found that feature extraction and ranking that include multiple layers of information enhance the prediction accuracy of the model. The embedding of DL refreshes our knowledge of pathogen features.

Drug Discovery (With a Particular Focus on Antibiotics/Antimicrobial Peptides)

The abuse of antibiotics has led to a worsening problem of drug resistance in pathogenic bacteria, which has been an enormous threat to human health. Screening for secondary metabolites in soil microorganisms that inhibit the growth of pathogenic bacteria is regarded as the traditional primary means of antibiotic discovery (Wright, 2017). The current dilemma of decreasing the rate of discovery of new antibiotics urgently needs to be addressed. In addition, the administrative costs of screening approaches based on large synthetic chemical libraries and the high rate of antibiotic design attrition have increased the necessity for new antibiotic discovery methods to improve the rate of new antibiotic discovery. Modern drug discovery has entered the era of big data. AI modeling of the dynamic, heterogeneous, and large-scale nature of drug datasets continues to drive paradigm innovation in drug discovery (Zhu, 2020).

Techniques to identify and predict new antibiotic structural classes with the help of ML are largely mature and widely adopted (Camacho et al., 2018). DL accelerates the screening process of compounds with antibiotic properties from existing chemical libraries (Dimasi et al., 2016). Antimicrobial peptides

TABLE 1 | The available data and materials for prediction of pathogens and epidemiology.

Tools	Availability of data and materials	References
VirSorter	https://github.com/simroux/VirSorter	Roux et al., 2015
VirSorter2	https://bitbucket.org/MAVERICLab/VirSorter2	Guo et al., 2021
VirFinder	https://github.com/jessieren/VirFinder	Ren et al., 2017
DeepVirFinder	https://github.com/jessieren/DeepVirFinder	Ren et al., 2020
MARVEL	https://github.com/LaboratorioBioinformatica/MARVEL	Amgarten et al., 2018
VIBRANT	https://github.com/AnantharamanLab/VIBRANT/	Kieft et al., 2020

(AMPs) are candidates for coping with antibiotic resistance. Researchers have successively established several antimicrobial peptide databases containing data on various types of AMPs from various sources, such as APD, CAMP, and AVPDB, which greatly facilitate mining and forecasting of AMPs. Fu et al. (2020) designed a DL model for high-throughput antibacterial peptide recognition (ACEP), which is innovative in that it introduces an amino acid embedding tensor to capture the similarity between amino acids, constructed a "convolution and concatenation" (CVCA) layer using the attention mechanism of natural language processing (NLP) to fuse various heterogeneous information or features, and quantified the contribution of different components of the model to the final prediction using the attention scores of different parts of the peptide sequence. Capecchi et al. (2021) trained recurrent neural networks (RNNs) using sequence information from DBAASP v.2 (Database of Antimicrobial Activity and Structure of Peptides, now updated to DBAASP v.3; Pirtskhalava et al., 2021), including AMP and non-AMP datasets, and hemolytic and non-hemolytic data, mixing the use of supervised and unsupervised learning for the first time, maximizing the utilization of highly selected posterior data. The study also synthesized and tested 28 sequences generated and selected, yielding 12 new active AMPs, eight of which were also non-hemolytic. Das et al. (2021) designed a fully automated computational framework for molecular targeting and screening, in which conditional latent (attribute) space sampling (CLaSS) was designed for target generation, which is more efficient and easily reusable than other ML methods. The framework generates a potential space of molecular information *via* deep generative autoencoder modeling, utilizes a classifier for training guidance, and filters the generated molecules through deep learning classifiers based on the physicochemical features obtained in high-throughput molecular dynamics simulations. This study reported 20 CLaSS-generated AMP sequences and 11 non-AMP sequences obtained *via* the above screening method, which was shown to be less prone to false negatives. Wang (2022) combined various NLP neural network models (NNMs), built a pipeline containing LSTM, attention, and BERT, and established a DL method that adapts to learn AMP sequence features to mine and identify novel AMPs. Among a total of 2,349 sequences identified as candidate AMPs, 216 were chemically synthesized, including 181 indicative of antibacterial activity (>83% positivity). The code availability is shown in **Table 2**.

Overall, the time is ripe for modern ML/DL applications for antibiotic discovery (Cardoso et al., 2020). Their effective contribution to the bulk filtering and prediction of antimicrobial peptides is alleviating concerns about the high risks and low returns associated with antibiotic development. Notably, the high success rate of deep neural network model-guided antibiotic development is heavily dependent on the combination of model prediction and appropriate experimental design, and this wet-dry combination strategy is a scientific idea that has started to be popularized after the discovery of complementary information and experimental practices.

TABLE 2 | The code availability for prediction of antimicrobial peptide (AMP) discovery.

Tools	Code availability	References
ACEP	https://github.com/Fuhaoyi/ACEP	Fu et al., 2020
RNN	https://github.com/reymond-group/MLpeptide	Capecchi et al., 2021
CLaSS	https://github.com/IBM/controlled-peptide-generation	Das et al., 2021
AMP prediction pipeline with NNMs	https://github.com/mayuefine/c_AMPs-prediction	Wang, 2022

CONCLUSION

Research in machine learning and deep learning is evolving rapidly, with architectures, algorithm combinations, and computational strategies changing rapidly. The ultimate goal is not only to predict the accuracy of the task but also to uncover the underlying biological processes in the scientific problem. The perception that "deep learning may eventually eliminate all other machine learning algorithms" is limited and one-sided. Deep learning modeling requires a large amount of training data to demonstrate fantastic performance, but realistic colony research frequently encounters problems with small sample datasets. At this point, deep learning methods fail to attack them, but traditional machine learning methods are capable of handling them. The development of effective analytical tools, including software for data mining and machine learning, ensures data validity, proper annotation, and open sharing, allowing most studies arising from the intersection of microbiology and machine learning to show promising findings. After bioinformatics and multiomics integration, ML and DL will lead the next wave of technologies to uncover biological regularity.

AUTHOR CONTRIBUTIONS

YJ drafted the manuscript. JL and DH modified the English content of this manuscript. YL and D-dL conceived the idea. All authors contributed to the article and approved the submitted version.

FUNDING

The work was financially supported by the China Postdoctoral Science Foundation (2021M701987; D-dL) and (2021M700084; YL).

ACKNOWLEDGMENTS

We thank Zhuo Pan and Yiran Cheng for advice on the technical basis in a draft of this manuscript.

REFERENCES

Amgarten, D., Braga, L. P. P., Da Silva, A. M., and Setubal, J. C. (2018). MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Front. Genet.* 9:304. doi: 10.3389/fgene.2018.00304

Ankerst, M., Breunig, M. M., Kriegel, H. P., and Sander, J. (1999). "OPTICS: ordering points to identify the clustering structure." in *1999 ACM SIGMOD International Conference on Management of Data.* June, 1-3; Philadelphia, PA, 49–60.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Beck, L. C., Granger, C. L., Masi, A. C., and Stewart, C. J. (2021). Use of omic technologies in early life gastrointestinal health and disease: from bench to bedside. *Expert Rev. Proteomics* 18, 247–259. doi: 10.1080/14789450.2021.1922278

Belkin, M., and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* 15, 1373–1396. doi: 10.1162/089976603321780317

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Verges, M.-C. C., Charles, T., et al. (2020). Microbiome definition re-visited: old concepts and new challenges. *Microbiome* 8:103. doi: 10.1186/s40168-020-00875-0

Braga, L. P. P., Spor, A., Kot, W., Breuil, M. C., Hansen, L. H., Setubal, J. C., et al. (2020). Impact of phages on soil bacterial communities and nitrogen availability under different assembly scenarios. *Microbiome* 8:52. doi: 10.1186/s40168-020-00822-z

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Broberg, M., Doonan, J., Mundt, F., Denman, S., and Mcdonald, J. E. (2018). Integrated multi-omic analysis of host-microbiota interactions in acute oak decline. *Microbiome* 6:21. doi: 10.1186/s40168-018-0408-5

Bulgarelli, D., Schlaeppi, K., Spaepen, S., Van Themaat, E. V. L., and Schulze-Lefert, P. (2013). Structure and functions of the bacterial microbiota of plants. *Annu. Rev. Plant Biol.* 64, 807–838. doi: 10.1146/annurev-arplant-050312-120106

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

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell* 173, 1581–1592. doi: 10.1016/j.cell.2018.05.015

Capecchi, A., Cai, X., Personne, H., Kohler, T., Van Delden, C., Reymond, J. L., et al. (2021). Machine learning designs non-hemolytic antimicrobial peptides. *Chem. Sci.* 12, 9221–9232. doi: 10.1039/d1sc01713f

Cardoso, M. H., Orozco, R. Q., Rezende, S. B., Rodrigues, G., Oshiro, K. G. N., Cândido, E. S., et al. (2020). Computer-aided design of antimicrobial peptides: are we generating effective drug candidates? *Front. Microbiol.* 10:3097. doi: 10.3389/fmicb.2019.03097

Chen, J. D., Chen, J. X., Zhang, D. F., Sun, Y. D., and Nanehkaran, Y. A. (2020). Using deep transfer learning for image-based plant disease identification. *Comput. Electron. Agric.* 173:105393. doi: 10.1016/j.compag.2020.105393

Cheng, Y. Z. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* 17, 790–799. doi: 10.1109/34.400568

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1023/A:1022627411411

Cover, T. M., and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27. doi: 10.1109/tit.1967.1053964

Cox, D. R. (1958). The regression-analysis of binary sequences. *J. Royal Statist. Soc, Ser. B Statist. Method.* 20, 215–232. doi: 10.1111/j.2517-6161.1958.tb00292.x

Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.-R., Simonis, N., et al. (2009). Literature-curated protein interaction datasets. *Nat. Methods* 6, 39–46. doi: 10.1038/nmeth.1284

Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., et al. (2021). Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* 5, 613–623. doi: 10.1038/s41551-021-00689-x

De'ath, G., and Fabricius, K. E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81, 3178–3192. doi: 10.2307/177409

Defays, D. (1977). Efficient algorithm for a complete link method. *Comput. J.* 20, 364–366. doi: 10.1093/comjnl/20.4.364

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via Em algorithm. *J. Royal Statist. Soc. Ser. B. Methodol.* 39, 1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technologies.* June 2-7; Minneapolis, 4171–4186.

Dimasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* 47, 20–33. doi: 10.1016/j.jhealeco.2016.01.012

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM* 55, 78–87. doi: 10.1145/2347736.2347755

Duygan, B. D. O., Hadadi, N., Babu, A. F., Seyfried, M., and Van Der Meer, J. R. (2020). Rapid detection of microbiota cell type diversity using machine-learned classification of flow cytometry data. *Commun. Biol.* 3:379. doi: 10.1038/s42003-020-1106-y

Dworkin, M. (2012). Sergei Winogradsky: a founder of modern microbiology and the first microbial ecologist. *FEMS Microbiol. Rev.* 36, 364–379. doi: 10.1111/j.1574-6976.2011.00299.x

Erev, I., and Roth, A. E. (1998). Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev.* 88, 848–881.

Ester, M., Kriegel, H.P., Sander, J., and Xiaowei, X. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise." in *KDD-96 Proceedings. Second International Conference on Knowledge Discovery and Data Mining*, August 2-4; 226–231.

Faisal, A., Dondelinger, F., Husmeier, D., and Beale, C. M. (2010). Inferring species interaction networks from species abundance data: a comparative evaluation of various statistical and machine learning methods. *Ecol. Inform.* 5, 451–464. doi: 10.1016/j.ecoinf.2010.06.005

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agric.* 145, 311–318. doi: 10.1016/j.compag.2018.01.009

Fiannaca, A., La Paglia, L., La Rosa, M., Lo Bosco, G., Renda, G., Rizzo, R., et al. (2018). Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinformatics* 19:198. doi: 10.1186/s12859-018-2182-6

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* 7, 179–188. doi: 10.1111/j.1469-1809.1936.tb02137.x

Frank, M. J., Seeberger, L. C., and O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* 306, 1940–1943. doi: 10.1126/science.1102941

Freund, Y. (1990). "Boosting a weak learning algorithm by majority." in *Proceedings of the Third Annual Workshop on Computational Learning Theory*, 202–216.

Fu, H., Cao, Z., Li, M., and Wang, S. (2020). ACEP: improving antimicrobial peptides recognition through automatic feature fusion and amino acid embedding. *BMC Genomics* 21:597. doi: 10.1186/s12864-020-06978-0

Galloway-Pena, J., and Hanson, B. (2020). Tools for analysis of the microbiome. *Dig. Dis. Sci.* 65, 674–685. doi: 10.1007/s10620-020-06091-y

Ghosh, D., Chakraborty, S., Kodamana, H., and Chakraborty, S. (2022). Application of machine learning in understanding plant virus pathogenesis: trends and perspectives on emergence, diagnosis, host-virus interplay and management. *Virol. J.* 19:42. doi: 10.1186/s12985-022-01767-5

Greener, J. G., Kandathil, S. M., Moffat, L., and Jones, D. T. (2022). A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* 23, 40–55. doi: 10.1038/s41580-021-00407-0

Guo, J. R., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., et al. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 9:37. doi: 10.1186/s40168-020-00990-y

Ha, C. W. Y., and Devkota, S. (2020). The new microbiology: cultivating the future of microbiome-directed medicine. *Am. J. Physiol. Gastrointest. Liver Physiol.* 319, G639–G645. doi: 10.1152/ajpgi.00093.2020

Hanage, W. P. (2014). Microbiology: microbiome science needs a healthy dose of scepticism. *Nature* 512, 247–248. doi: 10.1038/512247a

Hassani, M. A., Duran, P., and Hacquard, S. (2018). Microbial interactions within the plant holobiont. *Microbiome* 6:58. doi: 10.1186/s40168-018-0445-0

He, X. F., and Niyogi, P. (2003). "Locality preserving projections." in *17th Annual Conference on Neural Information Processing Systems (NIPS)*. December 8-13; Canada, 153–160.

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Jones, M. L., Ganopolsky, J. G., Martoni, C. J., Labbe, A., and Prakash, S. (2014). Emerging science of the human microbiome. *Gut Microbes* 5, 446–457. doi: 10.4161/gmic.29810

Jordan, M. I., and Mitchell, T. M. (2015). Machine learning: trends, perspectives, and prospects. *Science* 349, 255–260. doi: 10.1126/science.aaa8415

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: a survey. *J. Artif. Intell. Res.* 4, 237–285. doi: 10.1613/jair.301

Kaelin, E. A., Rodriguez, C., Hall-Moore, C., Hoffmann, J. A., Linneman, L. A., Ndao, I. M., et al. (2022). Longitudinal gut virome analysis identifies specific viral signatures that precede necrotizing enterocolitis onset in preterm infants. *Nat. Microbiol.* 7, 653–662. doi: 10.1038/s41564-022-01096-x

Kang, D. D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3:e1165. doi: 10.7717/peerj.1165

Keerthi, S. S., and Ravindran, B. (1994). A tutorial survey of reinforcement learning. *Sadhana Acad. Proc. Eng. Sci.* 19, 851–889.

Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 8:90. doi: 10.1186/s40168-020-00867-0

Knights, D., Costello, E. K., and Knight, R. (2011). Supervised classification of human microbiota. *FEMS Microbiol. Rev.* 35, 343–359. doi: 10.1111/j.1574-6976.2010.00251.x

Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: a survey. *Int. J. Robot. Res.* 32, 1238–1274. doi: 10.1177/0278364913495721

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyotylainen, T., Hämäläinen, A. M., et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* 17, 260–273. doi: 10.1016/j.chom.2015.01.001

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

Lecun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 541–551. doi: 10.1162/neco.1989.1.4.541

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Lederberg, J., and Mccray, A. T. (2001). 'Ome sweet 'omics—A genealogical treasury of words. *Scientist* 15:8.

Lee, J.-Y., Sadler, N. C., Egbert, R. G., Anderton, C. R., Hofmockel, K. S., Jansson, J. K., et al. (2020). Deep learning predicts microbial interactions from self-organized spatiotemporal patterns. *Comput. Struct. Biotechnol. J.* 18, 1259–1269. doi: 10.1016/j.csbj.2020.05.023

Leite, D. M. C., Brochet, X., Resch, G., Que, Y.-A., Neves, A., and Peña-Reyes, C. (2018). Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC Bioinformatics* 19:420. doi: 10.1186/s12859-018-2388-7

Liang, J., Luo, W., Yu, K., Xu, Y., Chen, J., Deng, C., et al. (2021). Multi-Omics revealing the response patterns of symbiotic microorganisms and host metabolism in Scleractinian coral *Pavona minuta* to temperature stresses. *Meta* 12:18. doi: 10.3390/metabo12010018

Lowe, A., Harrison, N., and French, A. P. (2017). Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress. *Plant Methods* 13:80. doi: 10.1186/s13007-017-0233-z

Lu, Y. Y., Chen, T., Fuhrman, J. A., and Sun, F. (2017). COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics* 33, btw290–btw798. doi: 10.1093/bioinformatics/btw290

Luo, E., Leu, A. O., Eppley, J. M., Karl, D. M., and Delong, E. F. (2022). Diversity and origins of bacterial and archaeal viruses on sinking particles reaching the abyssal ocean. *ISME J.* doi: 10.1038/s41396-022-01202-1 [Epub ahead of print].

Luz Calle, M. (2019). Statistical analysis of metagenomics data. *Genom. Inform.* 17:e6. doi: 10.5808/GI.2019.17.1.e6

Macqueen, J. (1965). On convergence of K-means and partitions with minimum average variance. *Ann. Math. Stat.* 36:1084.

Malla, M. A., Dubey, A., Kumar, A., Yadav, S., Hashem, A., and Abd_Allah, E. F. (2018). Exploring the human microbiome: the potential future role of next-generation sequencing in disease diagnosis and treatment. *Front. Immunol.* 9:2868. doi: 10.3389/fimmu.2018.02868

Marchesi, J. R., and Ravel, J. (2015). The vocabulary of microbiome research: a proposal. *Microbiome* 3:31. doi: 10.1186/s40168-015-0094-5

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. doi: 10.1038/nature14236

Moos, W. H., Pinkert, C. A., Irwin, M. H., Faller, D. V., Kodukula, K., Glavas, I. P., et al. (2017). Epigenetic treatment of persistent viral infections. *Drug Dev. Res.* 78, 24–36. doi: 10.1002/ddr.21366

Nakano, Y., Takeshita, T., Kamio, N., Shiota, S., Shibata, Y., Suzuki, N., et al. (2014). Supervised machine learning-based classification of oral malodor based on the microbiota in saliva samples. *Artif. Intell. Med.* 60, 97–101. doi: 10.1016/j.artmed.2013.12.001

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philos. Mag.* 2, 559–572. doi: 10.1080/14786440109462720

Pichler, M., Boreux, V., Klein, A. M., Schleuning, M., and Hartig, F. (2020). Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods Ecol. Evol.* 11, 281–293. doi: 10.1111/2041-210x.13329

Pirtskhalava, M., Amstrong, A. A., Grigolava, M., Chubinidze, M., Alimbarashvili, E., Vishnepolsky, B., et al. (2021). DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.* 49, D288–D297. doi: 10.1093/nar/gkaa991

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.* 1, 81–106. doi: 10.1023/a:1022643204877

Quinlan, J. R. (1996). "Bagging, boosting, and C4.5." in *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*; August 4-8; 1, 725-730.

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 5:69. doi: 10.1186/s40168-017-0283-5

Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., et al. (2020). Identifying viruses from metagenomic data using deep learning. *Quantitat. Biol.* 8, 64–77. doi: 10.1007/s40484-019-0187-4

Rosenblatt, F. (1958). The perceptron—a probabilistic model for information-storage and organization in the brain. *Psychol. Rev.* 65, 386–408. doi: 10.1037/h0042519

Roux, S., Enault, F., Hurwitz, B. L., and Sullivan, M. B. (2015). VirSorter: mining viral signal from microbial genomic data. *PeerJ* 3:e985. doi: 10.7717/peerj.985

Roweis, S. T., and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* 290:2323. doi: 10.1126/science.290.5500.2323

Ruff, W. E., Greiling, T. M., and Kriegel, M. A. (2020). Host-microbiota interactions in immune-mediated diseases. *Nat. Rev. Microbiol.* 18, 521–538. doi: 10.1038/s41579-020-0367-2

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Sangwan, N., Xia, F., and Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome* 4:8. doi: 10.1186/s40168-016-0154-5

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Scholkopf, B., Smola, A., and Muller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* 10, 1299–1319. doi: 10.1162/089976698300017467

Shahbaaz, M., Bisetty, K., Ahmad, F., and Hassan, M. I. (2016). Current advances in the identification and characterization of putative drug and vaccine targets

in the bacterial genomes. *Curr. Top. Med. Chem.* 16, 1040–1069. doi: 10.2174/1568026615666150825143307

Sibson, R. (1973). Slink—optimally efficient algorithm for single-link cluster method. *Comput. J.* 16, 30–34. doi: 10.1093/comjnl/16.1.30

Smirnov, E. A., Timoshenko, D. M., and Andrianov, S. N. (2013). "Comparison of Regularization Methods for ImageNet Classification with Deep Convolutional Neural Networks." in *2nd AASRI Conference on Computational Intelligence and Bioinformatics (CIB).* Jeju Island, December 27-28; 89–94.

Song, W., Wang, L., Liu, P., and Choo, K.-K. R. (2019). Improved t-SNE based manifold dimensional reduction for remote sensing data processing. *Multimed. Tools Appl.* 78, 4311–4326. doi: 10.1007/s11042-018-5715-0

Stres, B., and Kronegger, L. (2019). Shift in the paradigm towards next-generation microbiology. *FEMS Microbiol. Lett.* 366:fnz159. doi: 10.1093/femsle/fnz159

Stupp, D., Sharon, E., Bloch, I., Zitnik, M., Zuk, O., and Tabach, Y. (2021). Co-evolution based machine-learning for predicting functional interactions between human genes. *Nat. Commun.* 12:6454. doi: 10.1038/s41467-021-26792-w

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290:2319. doi: 10.1126/science.290.5500.2319

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449, 804–810. doi: 10.1038/nature06244

Van De Guchte, M., Blottiere, H. M., and Dore, J. (2018). Humans as holobionts: implications for prevention and therapy. *Microbiome* 6:81. doi: 10.1186/s40168-018-0466-8

Van Der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Wang, J. (2022). Identification of anti-microbial peptides from the human gut microbiome using deep learning. *Nat. Biotechnol.* doi: 10.1038/s41587-022-01226-0 [Epub ahead of print].

Wang, J. Y., Lee, H. M., and Ahmad, S. (2007). SVM-cabins: prediction of solvent accessibility using accumulation cutoff set and support vector machine. *Proteins* 68, 82–91. doi: 10.1002/prot.21422

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244. doi: 10.2307/2282967

White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* 5:e1000352. doi: 10.1371/journal.pcbi.1000352

Wright, G. D. (2017). Opportunities for natural products in 21st century antibiotic discovery. *Nat. Prod. Rep.* 34, 694–701. doi: 10.1039/c7np00019g

Wu, S., and Zhang, Y. (2008). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics* 24, 924–931. doi: 10.1093/bioinformatics/btn069

Xu, X., Xie, Z., Yang, Z., Li, D., and Xu, X. (2020). A t-SNE based classification approach to compositional microbiome data. *Front. Genet.* 11:620143. doi: 10.3389/fgene.2020.620143

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227. doi: 10.1038/nature11053

Yeh, C. H. (1991). Classification and regression trees (Cart). *Chemom. Intell. Lab. Syst.* 12, 95–96. doi: 10.1016/0169-7439(91)80113-5

Zhang, J., Liu, Y. X., Guo, X., Qin, Y., Garrido-Oter, R., Schulze-Lefert, P., et al. (2021). High-throughput cultivation and identification of bacteria from the plant root microbiota. *Nat. Protoc.* 16, 988–1012. doi: 10.1038/s41596-020-00444-7

Zhang, M.-L., Pena, J. M., and Robles, V. (2009). Feature selection for multi-label naive Bayes classification. *Inf. Sci.* 179, 3218–3229. doi: 10.1016/j.ins.2009.06.010

Zhu, H. (2020). Big data and artificial intelligence modeling for drug discovery. *Annu. Rev. Pharmacol. Toxicol.* 60, 573–589. doi: 10.1146/annurev-pharmtox-010919-023324

Zhu, X., Yan, S., Yuan, F., and Wan, S. (2020). The applications of nanopore sequencing technology in pathogenic microorganism detection. *Canad. J. Infect. Dis. Med. Microbiol.* 2020, 1–8. doi: 10.1155/2020/6675206

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

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

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 02 June 2022
doi: 10.3389/fmicb.2022.930981

Isolation, Screening, and Active Metabolites Identification of Anti-*Vibrio* Fungal Strains Derived From the Beibu Gulf Coral

Bingyao Huang†, Shuai Peng†, Shifang Liu, Yanting Zhang, Yuxiao Wei, Xinya Xu, Chenghai Gao, Yonghong Liu and Xiaowei Luo**

Institute of Marine Drugs, Guangxi University of Chinese Medicine, Nanning, China

OPEN ACCESS

Edited by:
Zheng Zhang,
Shandong University, China

Reviewed by:
Riming Huang,
South China Agricultural University, China
Tian Yong Qi,
Fuzhou University, China
Xian-Wen Yang,
Ministry of Natural Resources, China

***Correspondence:**
Xiaowei Luo
luoxiaowei1991@126.com
Yonghong Liu
yonghongliu@scsio.ac.cn

†These authors have contributed equally to this work

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

Received: *28 April 2022*
Accepted: *12 May 2022*
Published: *02 June 2022*

Citation:
Huang B, Peng S, Liu S, Zhang Y, Wei Y, Xu X, Gao C, Liu Y and Luo X (2022) Isolation, Screening, and Active Metabolites Identification of Anti-Vibrio Fungal Strains Derived From the Beibu Gulf Coral. Front. Microbiol. 13:930981. doi: 10.3389/fmicb.2022.930981

The Beibu Gulf harbors abundant underexplored marine microbial resources, which are rich in diversified secondary metabolites. The genera *Vibrio* is a well-known pathogenic bacterium of aquatic animals. In this study, 22 fungal strains were isolated and identified from the Beibu Gulf coral *via* the serial dilution method and internal transcribed spacer (ITS) sequence analysis, which were further divided into three branches by phylogenetic tree analysis. The crude extracts of them *via* small-scale fermentation were selected for the screening of inhibitory activity against *Vibrio alginalyticus*, *Vibrio coralliilyticus*, *Vibrio harveyi*, *Vibrio parahaemolyticus*, *Vibrio owensii*, and *Vibrio shilonii*. The results showed that eight fungal extracts displayed anti-*Vibrio* activity *via* the filter paper disk assay. Several of them showed strong inhibitory effects. Then, two tetramic acid alkaloids, equisetin (**1**) and 5′-epiequisetin (**2**), were identified from *Fusarium equiseti* BBG10 by bioassay-guided isolation, both of which inhibited the growth of *Vibrio* spp. with the MIC values of 86–132 μg/ml. The scanning electron microscope results showed that cell membranes of *Vibrio* became corrugated, distorted or ruptured after treatment with **1** and **2**. Taken together, this study provided eight fungal isolates with anti-*Vibrio* potentials, and two alkaloid-type antibiotics were found with anti-*Vibrio* effects from the bioactive strain *F. equiseti* BBG10. Our findings highlight the importance of exploring promising microbes from the Beibu Gulf for the identification of anti-*Vibrio* for future antibiotic development.

Keywords: the Beibu Gulf, coral-derived fungi, anti-*Vibrio*, phylogenetic tree, equisetin

INTRODUCTION

As an important biocenosis in marine ecosystems, a coral reef ecosystem represents an extraordinarily diverse biota in tropical environments, among which corals often constitute a dominant part of the reef biomass. Scleractinian corals harbor diverse and abundant microbial symbionts with different types of interactions, which function as the primary reef ecosystem engineers, constructing the framework and shaping the resource availability for other coral reef-associated organisms (Tang et al., 2020). Manipulation of the coral-associated microbiome was postulated as a key strategy to improve the resilience of reef-building corals. The genera *Vibrio*, known as *Vibrio coralliilyticus* and *Vibrio mediterranei*, are important coral pathogens

capable of inducing serious coral damage, which has seriously impacted reef-building corals throughout the oceans as well as global warming (Esther et al., 2020). Recently, coral and its associated microorganisms have been evidenced as promising producers of structurally diverse compounds with a wide range of potent bioactivities, such as anti-inflammatory, cytotoxic, antimicrobial, antivirus, and antifouling activities (Hou et al., 2015; Sang et al., 2019).

The Beibu Gulf, located in the north of the South China Sea, harbors abundant biodiversity in both marine organisms and microorganisms and is regarded as a potential source of new species, new genes, new drugs, and new biological materials (Xu et al., 2020). However, there are relatively few reports about marine natural products from the Beibu Gulf (Carroll et al., 2019, 2020, 2021, 2022; Wang et al., 2019). In continuation of our research program aiming at the discovery of bioactive metabolites from the Beibu Gulf-derived marine fungi, a series of new bioactive compounds with diversified structures have been obtained recently (Luo et al., 2021; Lu et al., 2022; Zhang et al., 2022). Therefore, the main objective of this study was to investigate the anti-*Vibrio* potential of fungi isolated from scleractinian corals collected from the Weizhou Islands coral reef in the Beibu Gulf and to obtain and evaluate the potential exploitable anti-*Vibrio* alkaloids, equisetin and 5′-epiequisetin, from *Fusarium equiseti* BBG10.

MATERIALS AND METHODS

Sampling and Isolation of Fungi

The corals *Porites lutea* were collected from the Weizhou Islands coral reef in Guangxi Zhuang Autonomous Region, China, in July 2019. The samples were stored in sterilized polythene bags, transported to the laboratory, and processed immediately for the isolation and cultivation of fungi. The fungi were isolated by the serial dilution method (1:10, 1:100, and 1:1,000) using potato dextrose agar (PDA) medium supplemented with sea salt (20 g/L) and chloramphenicol (20 mg/L). The inoculated plates were cultured at 25°C for 1–3 weeks and observed the growth of fungi intermittently. Fungal isolates were chosen and transferred into another blank agar plates based on their morphological traits. The isolated strains were deposited at the Institute of Marine Drugs, Guangxi University of Chinese Medicine, Nanning, China.

Identification of Fungi

The fungal strains were cultured in PDA medium at 28°C for 5 days. The genomic DNA of the fungal strains was isolated by using the protocol described previously (Shen et al., 2020). The internal transcribed spacer (ITS) sequences were checked and amplified using ITS1-(5′-TCCGTAGGTGAACCTGCGG-3′) and ITS4-(5′-TCCTCCGCTTATTGATATGC-3′) primers. The fungi were identified mainly by analysis of the ITS sequences (as shown in **Supplementary Material**) in the NCBI BLAST program. The phylogenetic tree was created based on the ITS sequences by MEGA7.

Screening of Fungi Fermentation and Extracts

The small-scale fermentations of 22 fungal isolates (BBG1–BBG22) were carried out in rice solid medium (50 g rice, 1.2 g artificial sea salt, and 60 mL H_2O) employing with 1-L Erlenmeyer flasks at room temperature for 30 days. The fermented cultures were overlaid and extracted three times with EtOAc. All the fungal extracts (10 mg/mL dissolved in methanol) were analyzed by high-performance liquid chromatography (HPLC; Shimadzu Prominence-i LC-2030) using a PDA detector and an ODS column (YMC-pack ODS-A, 4.6 mm × 250 mm, 5 μm). In addition, the organic extracts were combined and evaporated *in vacuo* as a total crude extract for further anti-*Vibrio* assays.

Anti-*Vibrio* Assay

All the fungal extracts, along with two isolated tetramic acid alkaloids, equisetin (**1**) and 5′-epiequisetin (**2**), were screened for antibacterial activity against *Vibrio alginalyticus*, *Vibrio coralliilyticus*, *Vibrio harveyi*, *Vibrio parahaemolyticus*, *Vibrio owensii*, and *Vibrio shilonii*, by using a K–B disk agar diffusion method (Zhang et al., 2022). The strain *Vibrio parahaemolyticus* was kindly provided by Prof. Nan Li (Nanning Normal University, Nanning, China), while other Vibrio strains were kindly provided by Prof. Chang Chen (South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, China). Each fungal extract was dissolved in dimethyl sulfoxide (DMSO) at a final concentration of 25 mg/ml, and the two compounds (**1**–**2**) were prepared at a concentration of 10 mg/ml. The positive control chloramphenicol was prepared at a concentration of 150 μg/mL in DMSO. The *Vibrio* strains were cultivated in Luria Bertani (LB) broth medium (10 g/L peptone, 5 g/L yeast extract, and 20 g/L NaCl, pH adjusted to 7.0) and were incubated at 28°C, and the cultures were incubated to logarithmic phase with the optical density at 600 nm (OD_{600}) reaching 0.8. The exponential-phase cells were added to LB agar medium (40°C–45°C) at a final concentration of 5×10^4 cfu/mL. After solidification for 20 min, sterile filter paper impregnated with 3 μL of sample solution was placed on the plates and incubated at 28°C for 12 h. The anti-*Vibrio* effects were checked and recorded.

The minimal inhibitory concentration (MIC) assay of equisetin (**1**) and 5′-epiequisetin (**2**) toward these *Vibrio* strains was further determined with minor modification as described previously (Zhang et al., 2022). The *Vibrio* strains were cultivated in LB broth medium at 28°C, and the culture was incubated to exponential phase with the optical density at 600 nm (OD_{600}) reaching 0.8. Briefly, the OD_{600} of exponential-phase cells of *Vibrio* was adjusted to 0.01 with LB broth medium. Thereafter, 150 μL of the *Vibrio* cells suspension was transferred into the wells of a 96-well microplate with different concentrations of **1** or **2**. DMSO (1%, v/v) was served as the negative control. The microplate was incubated at 28°C for 16 h and checked after incubation (Wiegand et al., 2008). The MIC values were defined as the lowest concentrations of **1** or **2** that inhibited the growth of *Vibrio* spp. ($OD_{600} < 0.05$).

Scanning electron microscopy (SEM) was further performed to investigate the morphological changes of *Vibrio* treated with compounds **1** and **2**. The strain *V. parahaemolyticus* was incubated on the LB as described above. Thereafter, compounds **1** and **2**

(10 mg/mL) were added to the *Vibrio* cells suspension at a final concentration of 1.0 × MIC and incubated at 28°C for 12 h. The treated cells were collected for the SEM assay. The treated cells were washed three times by PBS, and fixed in 2.5% glutaraldehyde for 2 h. Then, the treated cells dehydrated successively in an ethanol series of 30%, 50%, 70%, 80%, 90%, and 100% tert-butyl alcohol for 10 min at each stage. The freeze-dried samples were analyzed with an SEM (Zeiss, Sigma 300) operated at 3 kV.

Isolation and Structure Characterization of Equisetin (1) and 5′-Epiequisetin (2)

The fungal strain *F. equiseti* BBG10 was cultured on Müller Hinton broth (MB) agar plates (malt extract 15 g, artificial sea salt 15 g, and agar 20 g) at 25°C for 7 days. Then, it was inoculated in the seed medium (malt extract 15 g and artificial sea salt 15 g in 1.0-L tap distilled H_2O, pH 7.4–7.8) at 25°C on a rotary platform shaker at 180 rpm for 48 h. Subsequently, a large-scale fermentation of *F. equiseti* BBG10 was carried out in modified rice solid medium (150 g rice, 3.0 g artificial sea salt, and 180 mL H_2O) employing with 1 L × 20 Erlenmeyer flasks at room temperature for 30 days. The whole fermented cultures were extracted with EtOAc three times to provide a brown extract (50 g). The EtOAc crude extract was fractionated by medium pressure liquid chromatography (MPLC) using a step gradient elution with petroleum ether/CH_2Cl_2/MeOH (petroleum ether/CH_2Cl_2, 1:0–0:1; CH_2Cl_2/methanol, 1:0–1:1, v/v), which afforded 10 fractions (Frs.1 ~ 10) based on thin-layer chromatography (TLC) analysis. Fr.6 was further separated by semipreparative high performance liquid chromatography (HPLC) with MeCN/H_2O (80:20, v/v, 5.0 mL/min) to yield compounds **1** (t_R = 38 min, 300 mg) and **2** (t_R = 43 min, 350 mg). Their structures were confirmed by high resolution-electron spray ionization mass spectrometry (HR-ESIMS) and high-performance liquid chromatography-diode array detection (HPLC-DAD) data analysis as well as comparison with the standard reference substances. HR-ESIMS spectra were collected on a Waters Xevo G2-S TOF mass spectrometer (Waters Corporation, United States).

RESULTS

Identification and Phylogenetic Tree Analysis of Fungal Strains Derived From the Weizhou Islands Coral

Twenty-two candidate fungal strains were identified on the basis of the molecular protocol by amplification and sequencing of the DNA sequences of the ITS region of the rDNA gene. A classification of 22 strains based on the species name of the closely related species is shown in **Figure 1**. The predominant genera were *Aspergillus* and *Trichoderma*. Based on the ITS sequences, a phylogenetic tree was created using the neighbor-joining method to analyze the genetic relationship between 22 strains. The results showed that these strains could be divided into three major branches, while *Annulohypoxylon stygium* BBG22 is the most distant from other strains and belongs to a relatively independent branch. Moreover, the morphological property of the potential strain BBG10 was further collected to confirm the identification by using scanning electron microscopy (**Supplementary Figure 1**).

Anti-*Vibrio* Activity of the Fungal Extracts

Cultivation of fungi from the Weizhou Islands coral yielded a total of 30 isolates. Reduplicate isolates were excluded under the guidance of observation of morphological differences, including the visible examination of growth characteristics, mycelia, and diffusible pigment. As a result, 22 independent strains (BBG1–BBG22) were selected for the screening of anti-bacterial activity against 6 strains of *Vibrio* spp. As shown in **Table 1** and **Supplementary Figure 2**, the fungal extracts of 8 isolates (36.3%) displayed potential growth inhibition against *Vibrio* in the filter paper disk assay (75 μg extracts/per piece). It is worth noting that the fungal extracts of two isolates, *Trichoderma virens* BBG4 and *Trichoderma harzianum* BBG6, exhibited strong anti-*Vibrio* activity, and the sizes of the inhibition zone were larger than that of the positive control, chloramphenicol (150 μg/ml). A previous study reported that *T. virens* could produce gliotoxin with strong antimicrobial activity (Vargas et al., 2014). Besides, *T. harzianum* was reported to be a biocontrol bacterium for plant diseases (Altomare et al., 1999). The results also showed that different *Vibrio* species have divergent susceptibilities to extracts. Additionally, 22 fungal extracts were further analyzed by HPLC for the diversity of secondary metabolites. Therefore, *F. equiseti* BBG10 with interesting HPLC-DAD profiles (**Supplementary Figure 3**) of its crude extract was selected as the bioactive target strain to identify the active constituents.

Production of Bioactive Metabolites

To investigate the bioactive constituents of *F. equiseti* BBG10, a large-scale fermentation was performed in 3 kg of rice solid medium. After harvest, its organic extract was further separated by MPLC and HPLC. Their structures were confirmed by HR-ESIMS and HPLC-DAD data analysis as well as comparison with the standard reference substances, which were identified as equisetin (**1**) and 5′-epiequisetin (**2**), respectively (**Figure 2**).

Anti-*Vibrio* Activities of Compounds 1 and 2

Equisetin has been reported to have various biological actions, including antibacterial (Vesonder et al., 1979), anti-HIV (Hazuda et al., 1999; Clercq, 2000), antiobesity, and selective cytotoxicity to mammalian cells (Yin et al., 2013). However, this is the first report on the anti-*Vibrio* activity of equisetin (**1**) and 5′-epiequisetin (**2**) against *V. alginalyticus*, *V. coralliilyticus*, *V. harveyi*, *V. parahaemolyticus*, *V. owensii*, and *V. shilonii*. The filter paper disk assay showed that equisetin and 5′-epiequisetin exhibited weak bacteriostatic activity against *V. alginalyticus*, *V. parahaemolyticus*, *V. owensii*, and *V. shilonii* (**Figure 3**; **Supplementary Table 1**).

In addition, the MIC assay was further used to test the bacteriostatic ability of equisetin and 5′-epiequisetin. Both equisetin and 5′-epiequisetin showed inhibitory effects on six strains of *Vibrio* with the MIC values ranging from 86 to 132 μg/ml (**Table 2**). In order to clearly reflect the effect of compounds

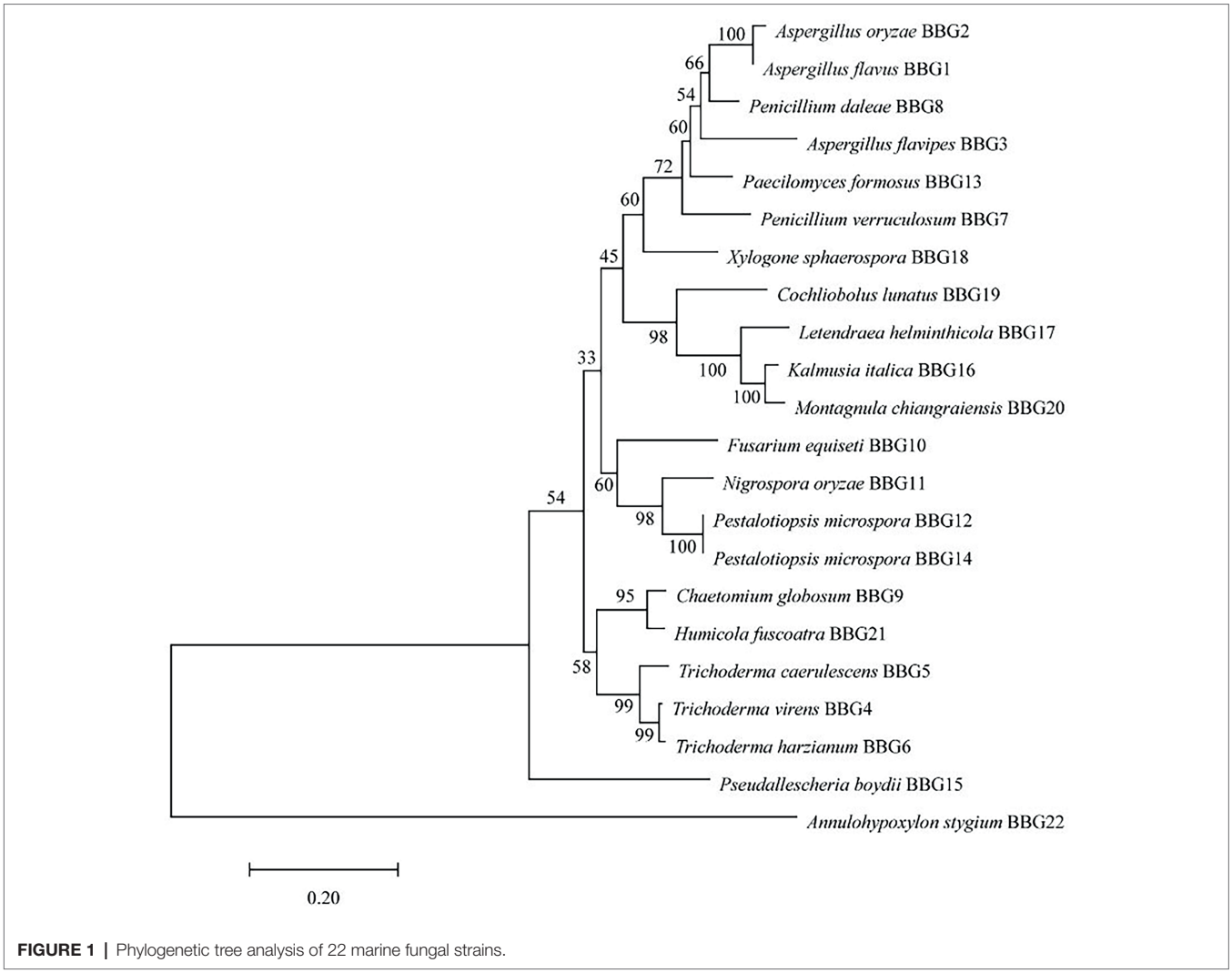

FIGURE 1 | Phylogenetic tree analysis of 22 marine fungal strains.

TABLE 1 | The anti-*Vibrio* activity of fungal extracts (diameter of inhibition zone, mm).

	Vibrio alginalyticus	*Vibrio coralliilyticus*	*Vibrio harveyi*	*Vibrio parahaemolyticus*	*Vibrio shilonii*	*Vibrio owensii*
Chl	1.69 ± 0.10	2.15 ± 0.22	2.00 ± 0.04	1.53 ± 0.07	2.01 ± 0.20	1.66 ± 0.20
Aspergillus oryzae BBG2	0.78 ± 0.06	0.93 ± 0.08	0	0.87 ± 0.19	0.90 ± 0.10	0.89 ± 0.09
Aspergillus flavipes BBG3	0	1.23 ± 0.21	0	0.75 ± 0.02	0.84 ± 0.07	0.72 ± 0.05
T. virens BBG4	1.73 ± 0.27	2.58 ± 0.03	1.96 ± 0.10	2.28 ± 0.07	2.62 ± 0.04	2.26 ± 0.36
T. harzianum BBG6	2.10 ± 0.25	2.20 ± 0.10	1.73 ± 0.15	2.36 ± 0.07	2.32 ± 0.12	1.19 ± 0.01
Chaetomium globosum BBG9	0.80 ± 0.04	1.64 ± 0.05	0.75 ± 0.03	0.81 ± 0.06	1.26 ± 0.07	1.33 ± 0.02
F. equiseti BBG10	0.80 ± 0.04	1.13 ± 0.08	0.71 ± 0.03	0.77 ± 0.03	0.94 ± 0.03	0.78 ± 0.10
Nigrospora oryzae BBG11	0.88 ± 0.07	1.02 ± 0.18	0.82 ± 0.12	0.85 ± 0.02	1.31 ± 0.06	1.28 ± 0.32
A. stygium BBG22	0	0	0.75 ± 0.06	0.85 ± 0.17	0.83 ± 0.06	0.92 ± 0.13

Chl, chloramphenicol.

on the growth of tested bacteria, one of the Vibrio strain, *V. parahaemolyticus*, was selected to further investigate the growth curves of **1** and **2** at different concentrations (0.5 × MIC, 1.0 × MIC, and 2.0 × MIC), while the OD_{600} values were recorded within 16 h. As shown in **Supplementary Figure 4**, the growth of *V. parahaemolyticus* in the negative control and 0.5 × MIC (**1** and **2**) treatment groups entered the logarithmic growth period after 2 h, and the number of bacterial colonies kept growing within 16 h. Notably, the growth of *V. parahaemolyticus* was almost completely stagnant at the treatments of 1.0 × MIC and 2.0 × MIC, which suggested the bacteria were completely inhibited or even killed after treatment of compounds **1** and **2**. The above results indicated that anti-*Vibrio* effect of compounds **1** and **2** was in a dose-dependent manner.

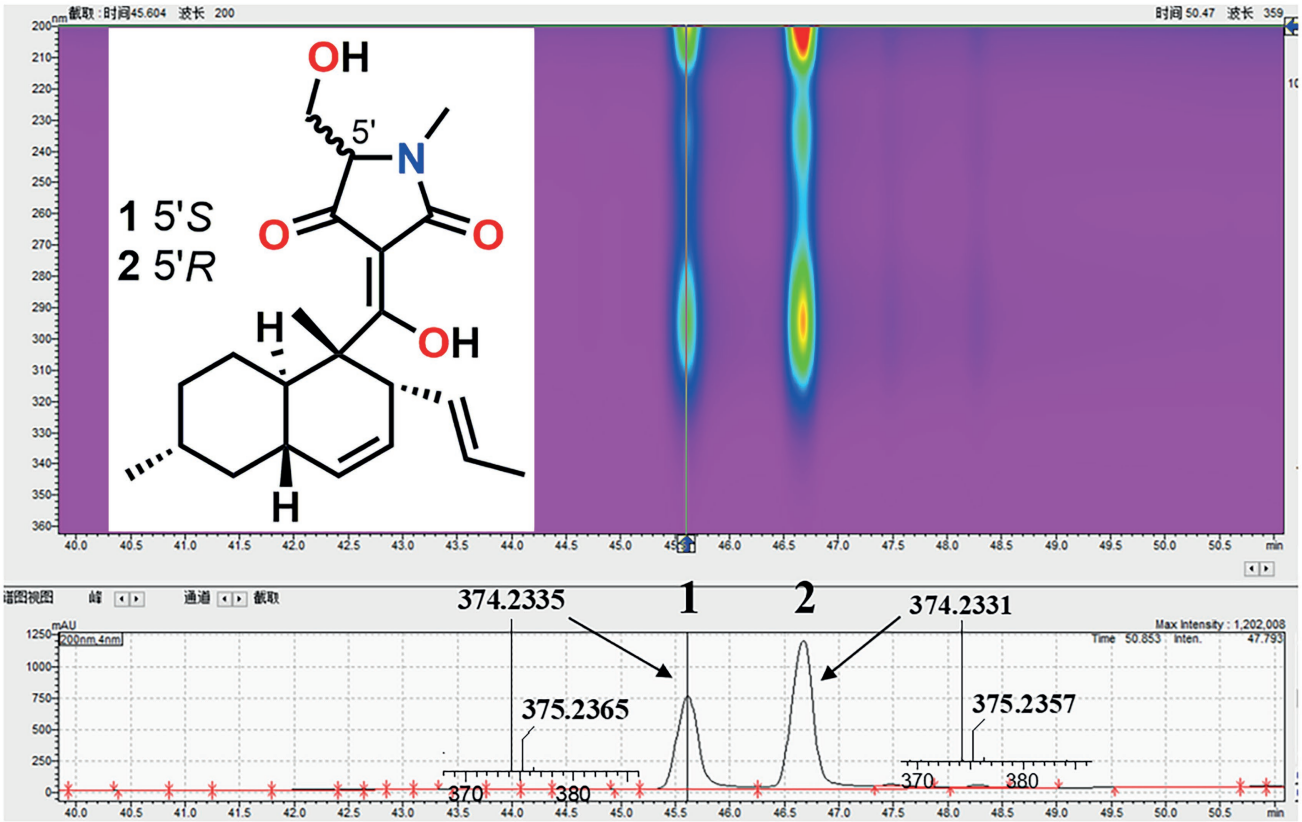

FIGURE 2 | The isolation and structures of equisetin (**1**) and 5′-epiequisetin (**2**).

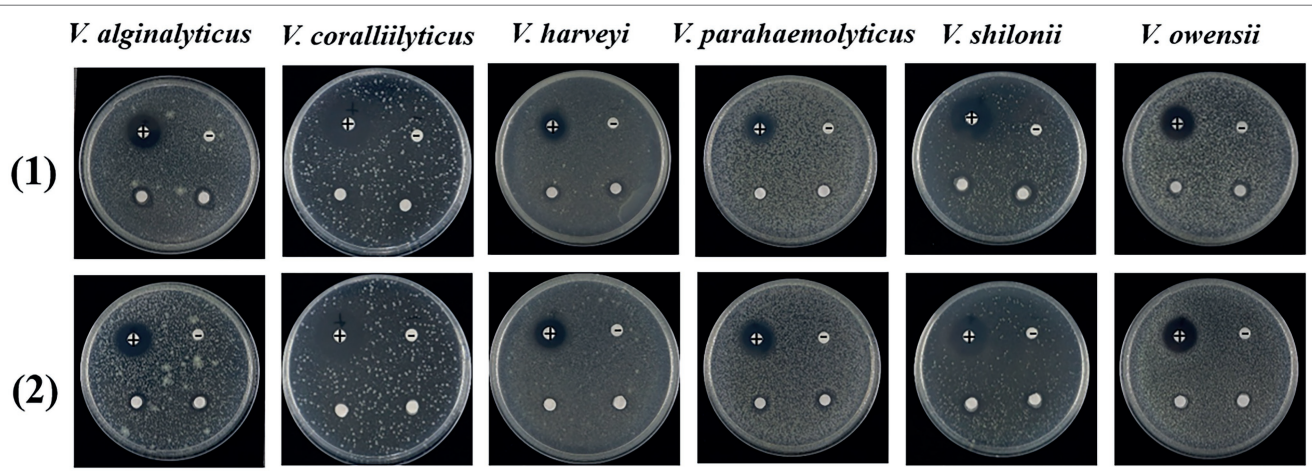

FIGURE 3 | Inhibitory activities of **1** and **2** against *Vibrio*. Each plate contains four pieces of paper disk, positive control marks as "+," negative control marks as "–," and the other two pieces of paper disk contains compound.

Thereafter, SEM was performed to investigate the morphological changes of *V. parahaemolyticus* treated with equisetin and 5′-epiequisetin. The results showed that the cell surfaces of the control group were smooth and that the *Vibrio* cells were plump and round. In contrast, the cell membranes became corrugated, distorted or ruptured after treatment with equisetin or 5′-epiequisetin. Interestingly, the destructive ability of equisetin and 5′-epiequisetin toward *Vibrio* cells were stronger than that of chloramphenicol. These results indicated that equisetin and 5′-epiequisetin can destroy the structure of *Vibrio* cell membranes (**Figure 4**).

TABLE 2 | Anti-*Vibrio* activity of compounds **1** and **2** (MIC, μg/ml).

	Vibrio alginalyticus	*Vibrio coralliilyticus*	*Vibrio harveyi*	*Vibrio parahaemolyticus*	*Vibrio shilonii*	*Vibrio owensii*
1	119	119	132	119	119	119
2	86	106	106	106	106	86
Chl	1	1	1.2	1.2	1	1.2

Chl, chloramphenicol.

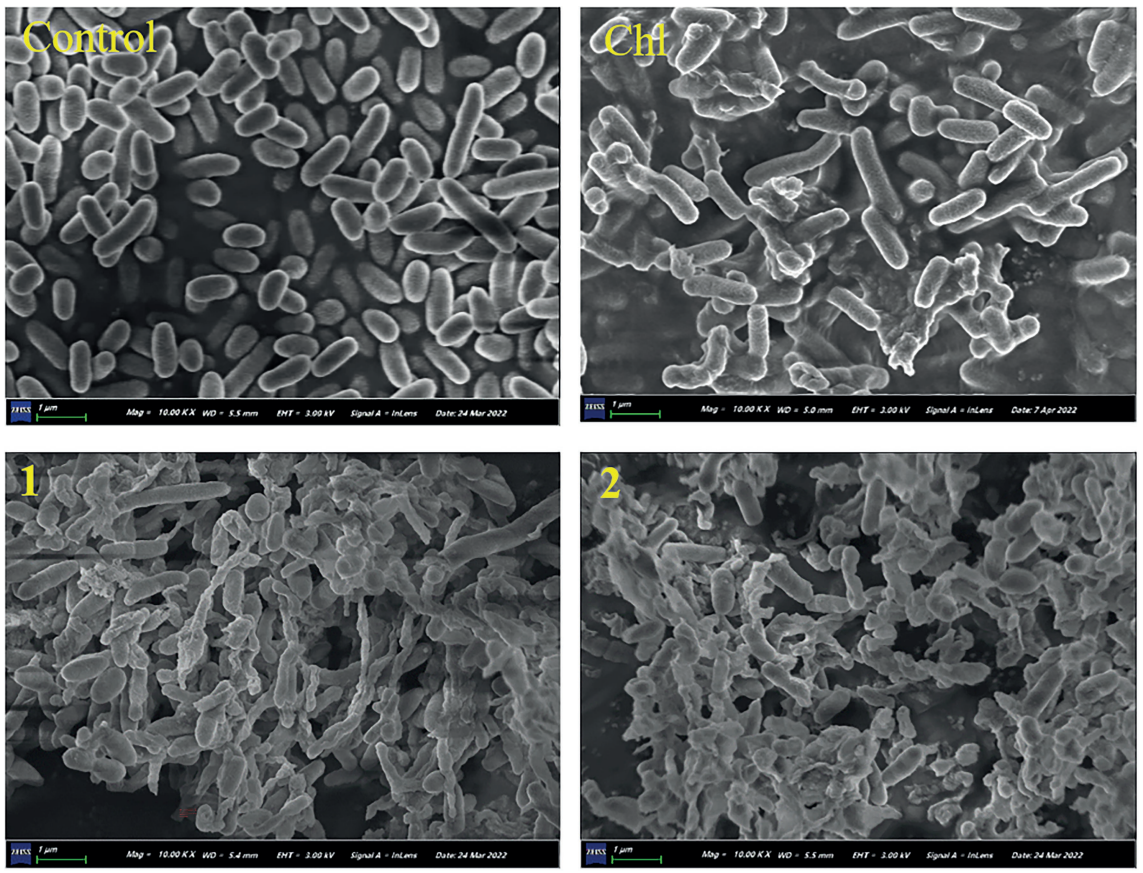

FIGURE 4 | Electron microscopic observation of morphological changes in *Vibrio parahaemolyticus* cells following treatment with **1** and **2**. (Chl, chloramphenicol).

DISCUSSION

The marine environment harbors a vast number of underexplored microbial resources. From a natural products perspective, marine microbes are better resources for novel anti-*Vibrio* lead compounds. Marine natural products represent a rich source of diverse molecules for drug development (Altmann, 2017). According to the Marinlit database, more than 36,000 compounds with diverse structures have been hitherto isolated from marine organisms, while over 1,000 new compounds have been isolated per year in the last decades. Notably, the proportion of novel compounds derived from marine microorganisms is gradually increasing (Carroll et al., 2022). To our knowledge, 15 marine drugs have been approved for marketing, including well-known cephalosporin and rifamycin from marine microorganisms. The Beibu Gulf is located in the tropics and subtropics, which is one of the regions with the most abundant microbial diversity in China. However, the research on microbial resources from the Beibu Gulf is relatively sparse. Therefore, it is promising to isolate and screen microbial resources and their bioactive metabolites from the Beibu Gulf.

Vibrio is a Gram-negative bacterium that is one of the main pathogenic bacteria of fish, shrimp, shellfish, and other marine animals (Letchumanan et al., 2015). Humans can also be infected by eating contaminated seafood, contact with seawater, etc. *Vibrio* pathogenicity mainly includes *V. parahaemolyticus*, *V. alginolyticus*, *V. vulnificus*, and *V. anguillarum* (Yu et al., 2012). The application of antibiotics is an effective method of *vibrio* control. There is an urgent need to find novel antibiotics against *Vibrio* (Preetha et al., 2015). In this work, we tried to

screen new anti-*Vibrio* natural products from the fungal resources from the Beibu Gulf. We screened 22 fungal crude extracts, and eight fungal crude extracts showed different degrees of anti-*Vibrio* activity. Among them, the crude extracts of *T. virens* BBG4 and *T. harzianum* BBG6 exhibited particularly anti-*Vibrio* activity, and the size of the inhibition zone was larger than that of chloramphenicol. Two active components, equisetin and 5′-epiequisetin, were further identified from one of the bioactive strain, *F. equiseti* BBG10.

Equisetin and related derivatives have long been recognized for their wide biological activity against eukaryotic and bacterial cells, including antibacterial (Vesonder et al., 1979), anti-HIV (Hazuda et al., 1999; Clercq, 2000), anti-obesity, and selective cytotoxicity effects on mammalian cells (Yin et al., 2013). Previous reports indicate that equisetin functions in eukaryotic cells by affecting mitochondrial metabolism (Freiberg et al., 2004; Quek et al., 2013). Equisetin could affect malonyl-CoA synthesis as an acetyl-CoA carboxylase inhibitor (Freiberg et al., 2004; Larson et al., 2020). HIV integrase is inhibited by equisetin based upon its metal-binding property (Hazuda et al., 1999; Clercq, 2000).

In this work, equisetin and 5′-epiequisetin were identified from *F. equiseti* BBG10, and the crude extracts exhibited anti-*Vibrio* activity. The filter paper disk assay and MIC assay showed that equisetin and 5′-epiequisetin exhibited slight anti-*Vibrio* activity. Interestingly, the SEM results showed that the cell membranes became corrugated, distorted or ruptured after treatment with equisetin or 5′-epiequisetin. In contrast, the number of cells destroyed by chloramphenicol was less than those of equisetin and 5′-epiequisetin. These results indicated that equisetin and 5′-epiequisetin can more significantly damage the structure of *Vibrio* cell membranes than chloramphenicol, suggesting that the mechanism by which equisetin inhibits cell growth and kills cells is distinct from that of chloramphenicol. Moreover, *T. virens* BBG4 and *T. harzianum* BBG6 are potential strains for finding more potent anti-*Vibrio* compounds. This will be the focus of our future research.

CONCLUSION

In summary, 22 fungal strains were isolated and identified from the Beibu Gulf coral *via* the serial dilution method and ITS sequence analysis, which were further divided into three branches by phylogenetic tree analysis, while eight fungal extracts were screened with potential anti-*Vibrio* activity *via* the filter paper disk assay. Further chemical investigation of the extracts of the target strain *F. equiseti* BBG10 *via* bioassay-guided isolation led to the characterization of two alkaloid-type antibiotics, equisetin and 5′-epiequisetin, which displayed anti-*Vibrio* activities against *V. alginalyticus*, *V. coralliilyticus*, *V. harveyi*, *V. parahaemolyticus*, *V. owensii*, and *V. shilonii*. Our research highlights the coral-derived microorganisms may be a large reservoir of bioactive natural products for future agrochemical development, and equisetin and 5′-epiequisetin could be promising lead compounds for the further development of novel anti-*Vibrio* agents.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

XL, BH, and YL conceived and designed the experiments, analyzed the data, wrote the manuscript, and prepared the figures and supplementary materials. XL, SP, and YZ isolated and purified fungi and compounds. XL, XX, and CG identified compounds. BH, SL, and YW performed anti-*Vibrio* assay and ITS sequencing. BH performed scanning electron microscopy. All authors commented on and approved the manuscript.

FUNDING

This work was financially supported by the Specific Research Project of Guangxi for Research Bases and Talents (AD19110013), the Natural Science Foundation of Guangxi (2021GXNSF DA075010 and 2020GXNSFGA297002), the Guangxi Young and Middle-aged University Teachers' Scientific Research Ability Enhancement Project (2021KY0315), the Special Fund for Bagui Scholars of Guangxi (YL), the National Natural Science Foundation of China (U20A20101), and the Scientific Research Foundation of Guangxi University of Chinese Medicine (2018006, 2018BS042, 2020QN025, 2019BS021, and 2018ZD005).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.930981/full#supplementary-material

REFERENCES

Altmann, K.-H. (2017). Drugs from the Oceans: marine natural products as leads for drug discovery. *Chimia* 71, 646–652. doi: 10.2533/chimia.2017.646

Altomare, C., Norvell, W., Björkman, T., and Harman, G. (1999). Solubilization of phosphates and micronutrients by the plant-growth-promoting and biocontrol fungus *Trichoderma harzianum* Rifai 1295-22. *Appl. Environ. Microbiol.* 65, 2926–2933. doi: 10.1128/AEM.65.7.2926-2933.1999

Carroll, A. R., Copp, B. R., Davis, R. A., Keyzers, R. A., and Prinsep, M. R. (2019). Marine natural products. *Nat. Prod. Rep.* 36, 122–173. doi: 10.1039/c8np00092a

Carroll, A. R., Copp, B. R., Davis, R. A., Keyzers, R. A., and Prinsep, M. R. (2020). Marine natural products. *Nat. Prod. Rep.* 37, 175–223. doi: 10.1039/c9np00069k

Carroll, A. R., Copp, B. R., Davis, R. A., Keyzers, R. A., and Prinsep, M. R. (2021). Marine natural products. *Nat. Prod. Rep.* 38, 362–413. doi: 10.1039/D0NP00089B

Carroll, A. R., Copp, B. R., Davis, R. A., Keyzers, R. A., and Prinsep, M. R. (2022). Marine natural products. *Nat. Prod. Rep.* doi: 10.1039/D1NP00076D

Clercq, E. (2000). Current lead natural products for the chemotherapy of human immunodeficiency virus (HIV) infection. *Med. Res. Rev.* 20, 323–349. doi: 10.1002/1098-1128(200009)20:5<323::AID-MED1>3.0.CO;2-A

Esther, R.-P., Martin-Cuadrado, A. B., Caraballo-Rodríguez, A. M., Rohwer, F., Dorrestein, P. C., and Antón, J. (2020). Virulence as a side effect of interspecies interaction in *Vibrio* coral pathogens. *mBio* 11, e00201–e002020. doi: 10.1128/mBio.00201-20

Freiberg, C., Brunner, N., Schiffer, G., Lampe, T., Pohlmann, J., Brands, M., et al. (2004). Identification and characterization of the first class of potent bacterial acetyl-CoA carboxylase inhibitors with antibacterial activity. *J. Biol. Chem.* 279, 26066–26073. doi: 10.1074/jbc.M402989200

Hazuda, D., Blau, C., Felock, P., Hastings, J., Pramanik, B., Wolfe, A., et al. (1999). Isolation and characterization of novel human immunodeficiency virus Integrase inhibitors from fungal metabolites. *Antivir. Chem. Chemother.* 10, 63–70. doi: 10.1177/095632029901000202

Hou, X.-M., Xu, R.-F., Gu, Y.-C., Wang, C.-Y., and Shao, C.-L. (2015). Biological and chemical diversity of coral-derived microorganisms. *Curr. Med. Chem.* 22, 3707–3762. doi: 10.2174/0929867322666151006093755

Larson, E., Lim, A., Pond, C., Craft, M., Cavuzic, M., Waldrop, G., et al. (2020). Pyrrolocin C and equisetin inhibit bacterial acetyl-CoA carboxylase. *PLoS One* 15:e0233485. doi: 10.1371/journal.pone.0233485

Letchumanan, V., Pusparajah, P., Teng Hern, T., Yin, W.-F., Lee, L. H., and Chan, K.-G. (2015). Occurrence and antibiotic resistance of *Vibrio parahaemolyticus* from shellfish in Selangor, Malaysia. *Front. Microbiol.* 6:1417. doi: 10.3389/fmicb.2015.01417

Lu, H. M., Tan, Y. H., Zhang, Y. T., Li, Z. H., Chen, J. Y., Gao, C. H., et al. (2022). Osteoclastogenesis inhibitory phenolic derivatives produced by the Beibu Gulf coral-associated fungus Acremonium sclerotigenum GXIMD 02501. *Fitoterapia* 159:105201. doi: 10.1016/j.fitote.2022.105201

Luo, X. W., Cai, G. D., Guo, Y. F., Gao, C. H., Huang, W. F., Zhang, Z. H., et al. (2021). Exploring marine-derived ascochlorins as novel human dihydroorotate dehydrogenase inhibitors for treatment of triple-negative breast cancer. *J. Med. Chem.* 64, 13918–13932. doi: 10.1021/acs.jmedchem.1c01402

Preetha, R., Vijayan, K. K., Jayapraksh, N. S., Alavandi, S. V., Santiago, T. C., and Bright Singh, I. S. (2015). Optimization of culture conditions for mass production of the probiotics Pseudomonas MCCB 102 and 103 antagonistic to pathogenic Vibrios in aquaculture. *Probiotics Antimicrob. Proteins* 7, 137–145. doi: 10.1007/s12602-015-9185-7

Quek, N., Matthews, J., Bloor, S., Jones, D., Bircham, P., Heathcott, R., et al. (2013). The novel equisetin-like compound, TA-289, causes aberrant mitochondrial morphology which is independent of the production of reactive oxygen species in *Saccharomyces cerevisiae*. *Mol. BioSyst.* 9, 2125–2133. doi: 10.1039/c3mb70056a

Sang, V. T., Dat, T. T. H., Vinh, L. B., Cuong, L. C. V., Oanh, P. T. T., Ha, H., et al. (2019). Coral and coral-associated microorganisms: A prolific source of potential bioactive natural products. *Mar. Drugs* 17:468. doi: 10.3390/md17080468

Shen, N. X., Liang, Z. Y., Liu, Q., Tu, C. D., Dong, K. M., Wang, C. Y., et al. (2020). Antifungal secondary metabolites isolated from mangrove Rhizosphere soil-derived Penicillium Fungi. *J. Ocean Univ. China* 19, 717–721. doi: 10.1007/s11802-020-4360-1

Tang, K. H., Zhan, W. N., Zhou, Y. Q., Xu, T., Chen, X. Q., Wang, W. Q., et al. (2020). Antagonism between coral pathogen Vibrio coralliilyticus and other bacteria in the gastric cavity of scleractinian coral *Galaxea fascicularis*. *Sci. China: Earth Sci.* 63, 157–166. doi: 10.1007/s11430-019-9388-3

Vargas, W., Mukherjee, P., Laughlin, D., Wiest, A., Moran-Diez, M., and Kenerley, C. (2014). Role of gliotoxin in the symbiotic and pathogenic interactions of *Trichoderma virens*. *Microbiology* 160, 2319–2330. doi: 10.1099/mic.0.079210-0

Vesonder, R., Tjarks, L., Rohwedder, W., Burmeister, H., and Laugal, J. (1979). Equisetin, an antibiotic from *Fusarium equiseti* NRRL 5537, identified as a derivative of N-methyl-2,4-pyrrolidone. *J. Antibiot.* 32, 759–761. doi: 10.7164/antibiotics.32.759

Wang, C., Su, Y., Gao, Y. K., Tan, X. C., Lei, F. H., Li, H., et al. (2019). The isolation, screening and the active metabolites identification of bioactive fungal strains with antimicrobial activity from samples collected in the Guangxi Beibu gulf region. *J. Int. Pharm. Res.* 46, 270–276. doi: 10.13220/j.cnki.jipr.2019.04.005

Wiegand, I., Hilpert, K., and Hancock, R. (2008). Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substance. *Nat. Protoc.* 3, 163–175. doi: 10.1038/nprot.2007.521

Xu, X. Y., Yang, H., Ning, X. Q., Yi, X. X., Liu, Y. H., and Gao, C. H. (2020). Research Progress of marine microbial diversity and chemical diversity in Beibu gulf. *Guangxi Sci.* 27, 433–450, 461. doi: 10.13656/j.cnki.gxkx.20201217.001

Yin, J., Kong, L. L., Wang, C., Shi, Y. B., Cai, S. J., and Gao, S. H. (2013). Biomimetic synthesis of Equisetin and (+)-Fusarisetin A. *Chem. Eur. J.* 19, 13040–13046. doi: 10.1002/chem.201302163

Yu, L.-P., Hu, Y.-H., Sun, J., and Sun, L. (2012). C312M: An attenuated *Vibrio anguillarum* strain that induces immunoprotection as an oral and immersion vaccine. *Dis. Aquat. Org.* 102, 33–42. doi: 10.3354/dao02544

Zhang, Y. T., Li, Z. C., Huang, B. Y., Liu, K., Peng, S., Liu, X. M., et al. (2022). Anti-Osteoclastogenic and antibacterial effects of chlorinated polyketides from the Beibu gulf coral-derived fungus *Aspergillus unguis* GXIMD 02505. *Mar. Drugs* 20:178. doi: 10.3390/md20030178

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

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

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 13 June 2022
doi: 10.3389/fmicb.2022.931065

Diversity and Biogeography of Human Oral Saliva Microbial Communities Revealed by the Earth Microbiome Project

Jinlan Wang[1], Jianqing Feng[2], Yongbao Zhu[1], Dandan Li[3], Jianing Wang[3]* and Weiwei Chi[1]**

[1] National Administration of Health Data, Jinan, China, [2] 96608 Army Hospital of Chinese People's Liberation Army, Hanzhong, China, [3] State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao, China

The oral cavity is an important window for microbial communication between the environment and the human body. The oral microbiome plays an important role in human health. However, compared to the gut microbiome, the oral microbiome has been poorly explored. Here, we analyzed 404 datasets from human oral saliva samples published by the Earth Microbiome Project (EMP) and compared them with 815 samples from the human gut, nose/pharynx, and skin. The diversity of the human saliva microbiome varied significantly among individuals, and the community compositions were complex and diverse. The saliva microbiome showed the lowest species diversity among the four environment types. Human oral habitats shared a small core bacterial community containing only 14 operational taxonomic units (OTUs) under 5 phyla, which occupied over 75% of the sequence abundance. For the four habitats, the core taxa of the saliva microbiome had the greatest impact on saliva habitats than other habitats and were mostly unique. In addition, the saliva microbiome showed significant differences in the populations of different regions, which may be determined by the living environment and lifestyle/dietary habits. Finally, the correlation analysis showed high similarity between the saliva microbiome and the microbiomes of Aerosol (non-saline) and Surface (non-saline), i.e., two environment types closely related to human, suggesting that contact and shared environment being the driving factors of microbial transmission. Together, these findings expand our understanding of human oral diversity and biogeography.

Keywords: oral cavity, saliva, microbiome, earth microbiome project, microbial diversity, environmental microbiome

OPEN ACCESS

Edited by:
Pengfei Ding, University of Maryland, Baltimore County, United States

Reviewed by:
Lu Liu, Qingdao Institute of Bioenergy and Bioprocess Technology (CAS), China
Ran Peng, Institute of Materia Medica, China

****Correspondence:***
Jinlan Wang
jinlan726@163.com
Jianing Wang
wangjianing@sdu.edu.cn
Weiwei Chi
15605316509@163.com

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

Received: *28 April 2022*
Accepted: *25 May 2022*
Published: *13 June 2022*

Citation:
Wang J, Feng J, Zhu Y, Li D, Wang J and Chi W (2022) Diversity and Biogeography of Human Oral Saliva Microbial Communities Revealed by the Earth Microbiome Project. Front. Microbiol. 13:931065. doi: 10.3389/fmicb.2022.931065

INTRODUCTION

The oral cavity is an important place for the delivery and exchange of substances inside and outside the human body and is also a gateway for pathogens and toxic substances to invade the body. The microbes found in the human oral cavity are collectively referred to as the oral microbiome (Gao et al., 2018; Mark Welch et al., 2019, 2020). The oral cavity connects the external environment with the digestive tract and respiratory tract, and the complex and variable interaction of oral microbes helps the body fight against undesirable external stimuli. Imbalances in the microbial community can lead to oral diseases such as dental caries, periodontitis, oral mucosal

diseases, and even some other diseases, such as autoimmune diseases, cardiovascular diseases, diabetes, cancers and neurodegenerative disorders (Jorth et al., 2014; Atarashi et al., 2017; Peters et al., 2017; Blod et al., 2018; Gao et al., 2018; Lira-Junior and Bostrom, 2018; Philip et al., 2018; Plaza-Diaz et al., 2018; Reddy et al., 2018; Saikaly et al., 2018; Wasfi et al., 2018; Bacali et al., 2022). Therefore, the oral microbiome plays an important role in maintaining the balance between human microbial communities and human health, and also in the onset and progression of several localized and systemic diseases including those of bacterial, viral and fungal origin (Zarco et al., 2012; He et al., 2015; Soffritti et al., 2021). However, compared to the gut microbiome, the oral microbiome has received little attention.

There are multiple microenvironments in the oral cavity that communicate with each other through saliva. The composition of the oral microbiome is complex, and the expanded Human Oral Microbiome Database (eHOMD) includes 770 microbial species of 230 genera in 16 bacterial and archaeal phyla (Escapa et al., 2018). Of all the species in this database, 57% are officially named, 13% are unnamed but cultivated and 30% are known only as uncultivated phylotypes. There is no difference among the oral, gut, and skin microbiomes of newborn babies, but the composition of their oral microbiomes will change significantly as age increases and dentition changes (Dominguez-Bello et al., 2010). The differences in the oral microbiomes at different time points for the same individual are significantly lower than those in the gut, skin and other body parts (Costello et al., 2009). The effects of the early living environment on shaping oral microbes are much greater than those of genetic factors (Shaw et al., 2017). In addition, lifestyle habits, social factors, and oral pH value also affect the composition of the oral microbiome (Willis et al., 2018).

The Earth Microbiome Project (EMP) aims to collect as many of the Earth's microbial communities as possible to promote our understanding of the relationship between microbes and the environment, including plants, animals and humans (Gilbert et al., 2010, 2014). The first data published by EMP contained 27,751 samples from 97 independent studies representing different environmental types, geographic locations, and chemical reactions (Thompson et al., 2017). All samples were subjected to DNA extraction and sequencing, and the bacterial and archaeal parts of the entire database were analyzed. Here, using sequencing data from 404 human oral saliva samples published by EMP, we explored the characteristics, core taxa of human oral microbiome and their association with the environmental microbiome, comparative analysis among them with human gut, nose/pharynx and skin microbes.

RESULTS AND DISCUSSION

Prokaryotic Composition in the Human Oral Saliva Habitat

We analyzed the sequenced data of 404 human oral saliva samples from 5 independent studies (**Supplementary Data**). For further calculation, 5000 observed sequences were randomly extracted from each sample. All samples were subjected to the Deblur algorithm to remove erroneous sequences and to calculate operational taxonomic units (OTUs) at single nucleotide precision.

The results showed that the average number of observed bacterial and archaeal OTUs was 71.25 $\pm$ 26.40 in human oral samples, with a maximum of 216 OTUs and a minimum of 25 OTUs in a single sample. The Chao1 index is relatively sensitive to low-abundance species. The average Chao1 index for human oral samples was 85.83 $\pm$ 34.12, ranging from 28.75 to 258.00. The Shannon index can simultaneously reflect species diversity and community uniformity. The average Shannon index for human oral samples was 3.61 $\pm$ 0.76, varying from 0.97 to 5.34. Faith's PD value (Faith's phylogenetic diversity) is a good measure of phylogenetic diversity, and the average of Faith's PD value for human oral samples was 11.87 $\pm$ 2.90, varying between 6.03 and 29.84. These results indicated that the diversity of the human oral microbiome was significantly different among individuals.

The predominant phyla of the human oral saliva microbiome were *Firmicutes*, *Proteobacteria*, *Bacteroidetes*, *Fusobacteria*, and *Actinobacteria*, with average relative abundances of 36.38, 31.00, 17.97, 9.11, and 4.88%, respectively (**Figure 1A**). The total relative abundance of the 10 predominant genera (>1%) was 83.88%, and *Streptococcus* (22.62%), *Neisseria* (13.86%), and *Haemophilus* (13.76%) were the top three genera in terms of average relative abundance (**Figure 1B**). These 10 high-abundance genera in human oral samples were distributed in multiple bacterial phyla, of which *Firmicutes*, *Proteobacteria*, *Bacteroidetes*, *Fusobacterium*, and *Actinomycetes* each contained two genera. Therefore, the human oral microbial communities show high complexity in community composition.

The human oral microbial samples were obtained from 3 regions, including 56 in Italy, 79 in Puerto Rico, and 269 in the United States. We found that the diversity of the oral microbiome was significantly different (Analysis of Variance, ANOVA, $p < 0.01$) among the populations of these four regions, and there were also obvious differences in community composition. A principal coordinate analysis (PCoA) based on Bray-Curtis distance showed that oral samples from the three regions were clearly distinguishable (**Figure 1C**). Therefore, although the individual differences in human oral microbiota were obvious, significant regional differences could still be observed, which might be related to the differences in living environment, dietary habits and other factors of populations in different regions.

Comparative Analyses of Prokaryotic Biodiversity Among Four Human Habitats

In addition to the oral cavity, the gut, nose/pharynx and skin are also important habitats for human microbial colonization. Using the data published by EMP, we compared the differences between the human oral microbiome and the gut, nasal/pharyngeal, and skin microbiomes. Among them, gut microbial data were from 216 samples, nasal/pharyngeal data were from 253 samples, and skin data were from 346 samples.

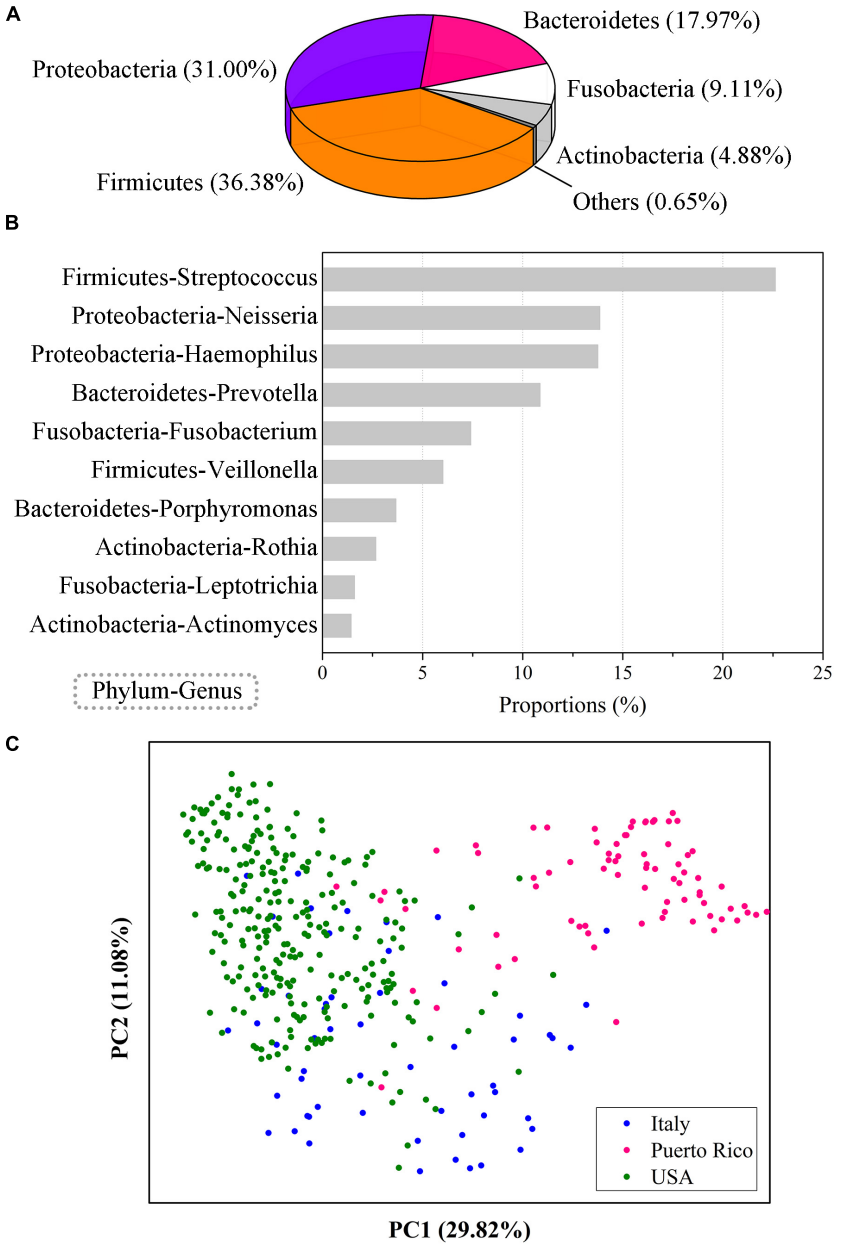

FIGURE 1 | Community composition of the human oral saliva microbiome. **(A)** Community composition of the human saliva microbiome at the phylum level. **(B)** Ten human saliva microbial genera with abundances over 1%. **(C)** PCoA analysis of the oral microbes from populations in different regions. The results were computed based on 404 EMP saliva samples from 3 regions, including 56 in Italy, 79 in Puerto Rico, and 269 in the United States.

The results showed that the human oral microbiome diversity was significantly (ANOVA, $p < 0.01$) lower than that of the gut, nasal/pharyngeal, and skin microbiomes (**Figure 2A**). The average number of observed bacterial and archaeal OTUs was 117 ± 40 in human gut samples, 289 ± 285 in human nasal/pharyngeal samples, and 297 ± 177 in human skin samples, each of which was significantly higher than the average value observed in human oral samples. The average values of the Chao1 index for human gut, nasal/pharyngeal and skin samples were 140.29 ± 49.75, 449.81 ± 469.85, and 422.42 ± 271.39, respectively, which were significantly higher than 85.83 ± 34.12 for oral samples. The average Shannon indexes for human gut, nasal/pharyngeal and skin samples were 4.45 ± 0.81, 4.27 ± 2.00, and 4.85 ± 1.60, respectively, which were significantly higher than 3.61 ± 0.76 for the oral sample. In addition, the average Faith's PD values for human gut, nasal/pharyngeal and skin samples were 15.30 ± 4.35, 30.26 ± 21.22, and 31.30 ± 14.53, respectively, which were also significantly higher than 11.87 ± 2.90 for oral samples.

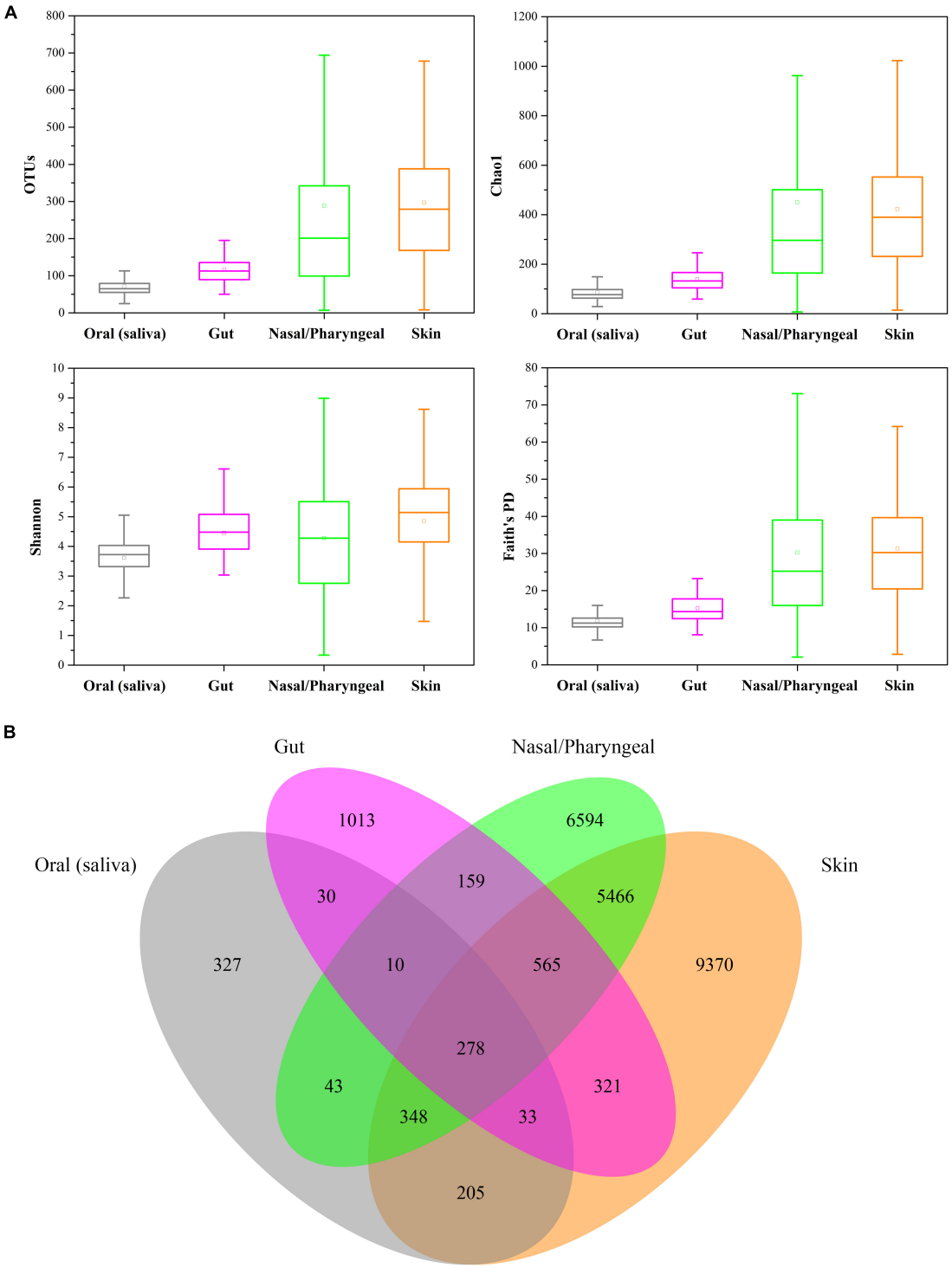

FIGURE 2 | Comparisons of the diversity of human oral saliva, gut, nasal/pharyngeal and skin microbiomes. **(A)** Comparison of alpha diversity, with diversity calculated sequentially as observed OTU, Shannon index, Chao1 index, and Faith's PD value. **(B)** Number of unique and cross-habitat distributed OTUs. Analysis was performed based on 216 gut, 253 nasal/pharyngeal and 346 skin samples from the EMP.

Although the oral saliva habitat contained the most samples, only 1274 OTUs were detected in all 404 samples, which was much lower than 2409 OTUs in gut samples, 13,463 OTUs in nasal/pharyngeal samples and 16,586 OTUs in skin samples, indicating that the low diversity of the human oral saliva microbiome once again (**Figure 2B**). Moreover, for

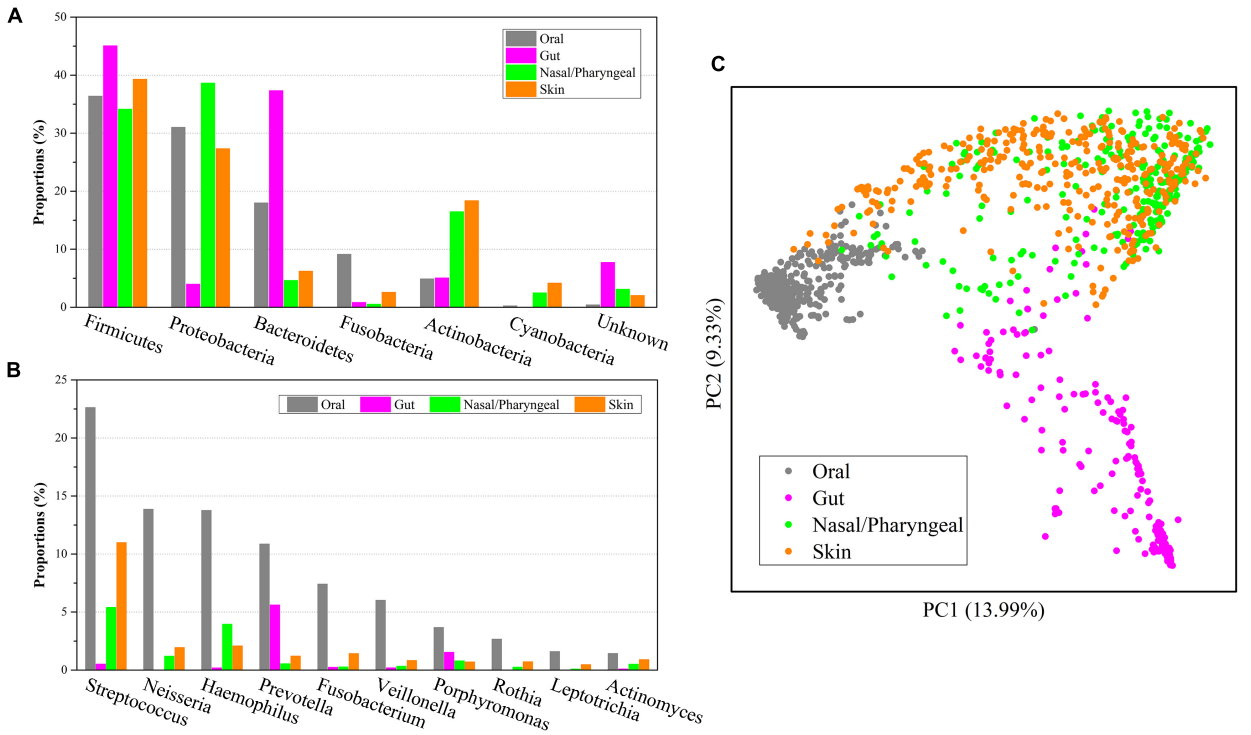

FIGURE 3 | Composition comparisons of human oral saliva, gut, nasal/pharyngeal and skin microbiomes. Distributions of six microbial phyla with the highest abundance **(A)** and 10 genera with the highest abundance **(B)** of the oral microbiome in other body parts. PCoA analysis **(C)** of microbial composition from four different parts of the human oral saliva, gut, nose/pharynx and skin.

these four habitats, only 25.67% of the OTUs in oral saliva habitats were unique to it and did not exist in the other three types of habitats, while the corresponding values of human gut, nasal/pharyngeal and skin habitats were 42.05, 48.98, and 56.50%, respectively. Notably, 278 OTUs could be detected in all four types of habitats, accounting for 21.82% of all OTUs in the oral cavity.

Prokaryotic Composition Differences Among Four Human Habitats

Firmicutes was not only the most abundant microbe at the phylum level in the oral microbiome but also had more than 30% abundance in other body locations, and its abundance in the gut microbiome was as high as 45.06% (**Figure 3A**). *Proteobacteria* had an abundance of more than 25% in the oral, nasal/pharyngeal and skin microbiomes but only 3.98% in the gut microbiome. *Bacteroidetes* accounted for 17.97% in the oral microbiome and 37.33% in the gut microbiome but only 4.64% and 6.22% in the nasal/pharyngeal and skin microbiomes, respectively. The abundance of *Fusobacteria* in the oral (9.11%) microbiome was significantly higher than that in the gut (0.84%), nasal/pharyngeal (0.52%) and skin (2.58%) microbiomes. The abundance of *Actinobacteria* in the oral (4.88%) microbiome was close to that of the gut (5.06%) microbiome but significantly lower than that of the nasal/pharyngeal (16.47%) and skin (18.38%) microbiomes. The abundance of *Cyanobacteria* in the oral (0.24%) microbiome was higher than that in the gut (0.02%) microbiome but lower than that in the nasal/pharyngeal (2.49%) and skin (4.17%) microbiomes. For the 10 genera with more than 1% abundance in the oral microbiome, their abundances in the gut, nasal/pharyngeal and skin microbiomes were significantly lower than those in the oral cavity (**Figure 3B**). For example, the abundance of *Neisseria* in the oral cavity was 13.86%, while the abundances in the gut, nasal/pharyngeal and skin microbiomes were only 0.003, 1.20, and 1.94%, respectively.

Furthermore, we performed a PCoA analysis based on the Bray-Curtis distance for 1,219 samples from the human oral cavity, gut, nose/pharynx, and skin, and displayed them in a scatter plot (**Figure 3C**). The results showed that the microbiome of oral samples could be well distinguished from the microbiomes of other body part samples, indicating that the oral microbiome was significantly different from other parts in community composition. Similarly, the microbiome for gut samples could also be well distinguished from the microbiomes of other body part samples. However, there was considerable overlap for the microbiomes between the nasal/pharyngeal and skin samples. The lowest dispersion of the oral microbiome among the four microbiomes suggested the lowest diversity, which was consistent with the alpha diversity index. The clustering analysis indicated that the nasal/pharyngeal and skin microbiomes were most similar, while the oral microbiome was more similar

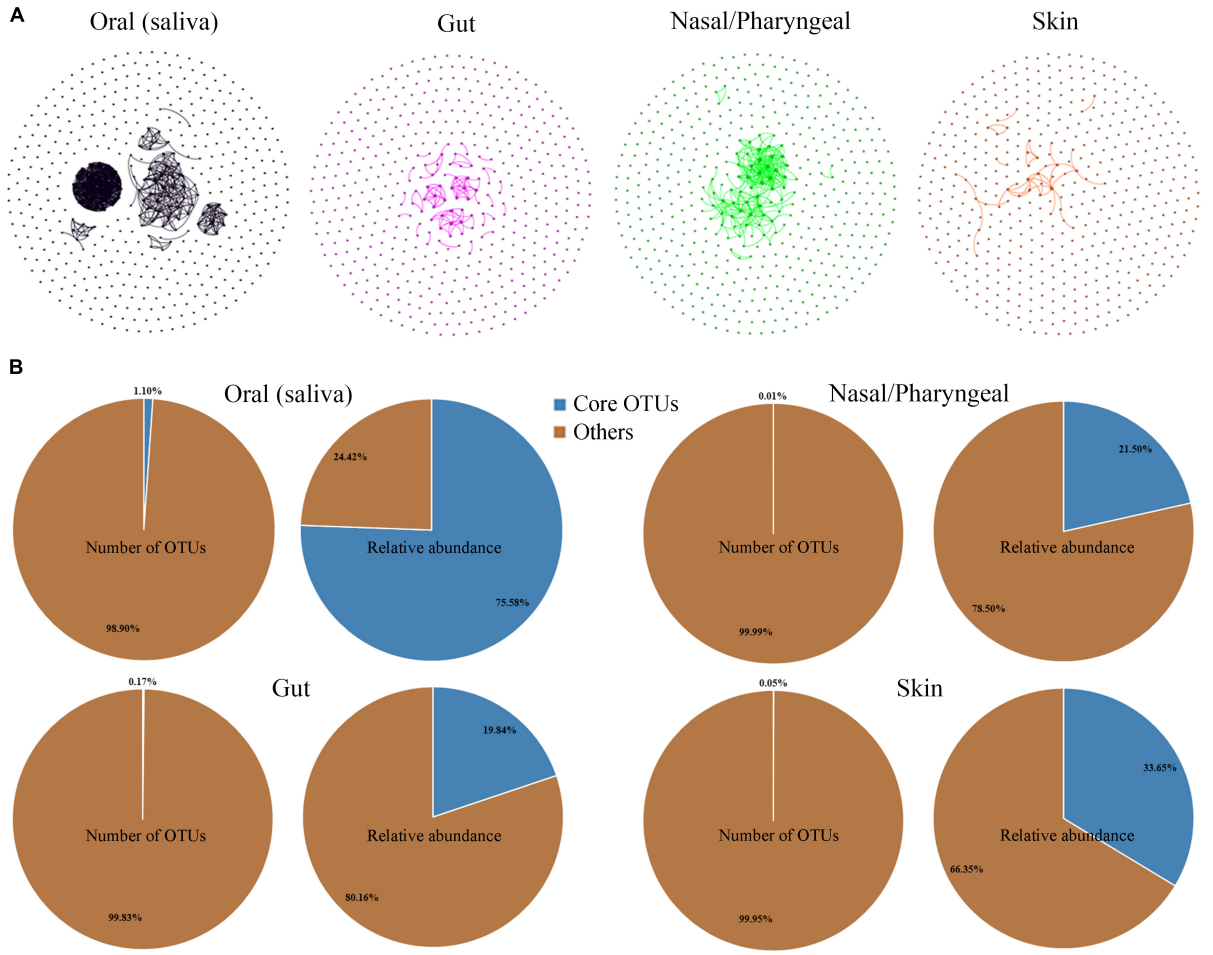

FIGURE 4 | Comparison of the dominant taxa and core taxa in the four habitats of human oral saliva, gut, nose/pharynx, and skin. **(A)** Co-occurrence network analysis of the dominant bacterial OTUs (top 500) of the different habitats. Each node represents an OTU, and the line indicates a significant correlation between the two OTUs (Pearson test, $r > 0.5$, $p < 0.05$). **(B)** Proportions of OTU number and abundance of core taxa in the four habitats of human oral saliva, gut, nose/pharynx, and skin.

to the nasal/pharyngeal and skin microbiomes relative to the gut microbiome.

Core Operational Taxonomic Units of the Human Oral Saliva Microbial Communities

Most microorganisms do not live in isolation; they thrive in communities with large numbers and develop close interactions that generate increased benefits for the group. Network inference techniques have frequently been applied to microbial interactions (Faust and Raes, 2012). To analyze the degree of interactions among dominant microbial taxa in different habitats, OTUs with the top 500 abundance were selected from gut, nasal/pharyngeal, oral saliva and skin samples to construct a co-occurrence network, respectively. The total relative abundance of these OTUs reached 97.42, 91.96, 99.74, and 90.52% in gut, nasal/pharyngeal, oral and skin samples, respectively, suggesting that they occupy the majority of the microbial community. The results showed that the aggregation of the microbial community network was significantly different among different habitats, indicating that there were significant differences in the interaction degree of dominant microbial taxa (**Figure 4A**). Specifically, oral showed the highest degree of network aggregation among the four habitats, followed by nose/pharynx, gut and skin. The parameters representing the correlation-based network topological structures were calculated; these parameters also showed that the edge and average degree were highest in the oral cavity, followed by the nose/pharynx, gut and skin.

We identified core taxa of prokaryotes in the habitat based on the criteria that they were present in at least 80% of the samples and had a total abundance of not less than 1% in all samples. The results showed that the core taxa of the human oral habitat contained only 14 OTUs, accounting for 1.10% of all oral OTUs, but with a total sequence abundance of 75.58% (**Figure 4B**). The origins of these core taxa members were extensive, with 6 OTUs classified as *Firmicutes*, 4 in *Bacteroidetes*, 2 in *Proteobacteria*, 1 in *Fusobacteria*, and 1 in *Actinobacteria*. Similarly, we also analyzed

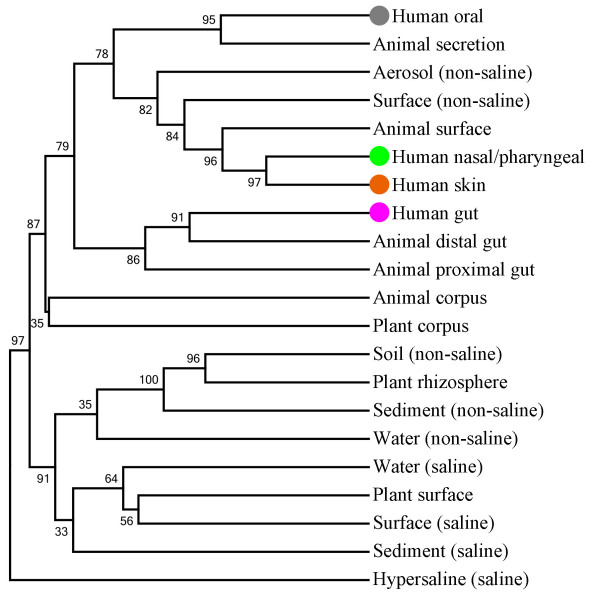

FIGURE 5 | Clustering analysis of microbial composition for human oral saliva, gut, nose/pharynx and skin samples and EMPO environmental labels. The tree was built by using the UPGMA method and Bray-Curtis similarity distance. Bootstrapping with 1,000 re-samplings was performed to determine the robustness of the clustering. The saliva, nose/pharynx, skin and gut are highlighted by gray, light green, orange-red and purple solid spheres, respectively.

gut, nasal/pharyngeal, and skin habitats, but only identified 4, 2, and 8 core OTUs, respectively. These core taxa also had a very low proportion of OTUs (0.17% in gut, 0.01% in nose/pharynx, and 0.05% in skin) but occupied a high sequence abundance (19.84% in gut, 21.50% in nose/pharynx, and 33.65% in skin). Core taxa had the greatest impact on the oral habitat among the four habitats, and most of these core taxa were unique to the oral habitat. The core taxa of the gut habitat were completely different from those of the other three habitats. Only one core taxon from *Streptococcus* was shared by oral, nasal/pharyngeal and skin habitats.

Association of Human Oral Saliva Microbes With Environmental Microbes

A number of microbes are exchanged with the external environment through the human oral cavity. Therefore, we tried to further analyze the association between oral microbes and environmental microbes. EMP classified the samples in different environments into the corresponding environmental labels. These environmental labels were first divided into two categories: Free-living and Host-associated, and further subdivided into 17 subcategories denominated as EMP Ontology (EMPO) level 3. We performed a cluster analysis to display the association of microbe compositions between human oral saliva, gut, nasal/pharyngeal and skin samples and EMPO environmental labels (**Figure 5**). The results showed that the closest EMPO environmental label to human oral samples was Animal secretion, the closest one to human nasal/pharyngeal and skin samples was the Animal surface, and the closest one to human gut samples was the Animal distal gut. Furthermore, the EMPO environmental labels that were close to human oral samples mostly belonged to the host-associated type but also included the two free-living environments of non-saline Aerosol and Surface. Aerosol and surface are the two types of environments where humans are most in close contact. Specifically, Aerosol is aerosolized dust or liquid. Surface is the biofilm from wet (<5 psu) or dry surface, wood, dust, and microbial mat.

For the 10 genera with more than 1% abundance in the oral microbiome, their abundances in all the EMPO environmental labels were obviously lower than those in the oral cavity. For example, the abundance of *Streptococcus* was 22.62% in the oral cavity, 5.07% in Aerosol (non-saline), 4.22% in Animal surface, 3.02% in Surface (non-saline), 1.35% in Animal proximal gut, and 0.58% in Animal distal gut. Interestingly, all 10 genera had the highest abundance in the Aerosol (non-saline) of the nine free-living environments, as well as the second highest abundance in the Surface (non-saline). Furthermore, we found that the abundance of these 10 genera in various environments had an obviously positive correlation. Therefore, the composition of the oral microbes represented by these 10 genera was specific and had a certain similarity with the microbial composition in the free-living Aerosol and Surface environments.

Aerosol (non-saline) and Surface (non-saline) are the two most closely related types of environments with humans, and the microorganisms in them can be expected to be the most easily transferred to the human body. For different parts of human body, the skin and nasal/pharyngeal microbiomes had the highest similarity with Aerosol (non-saline) and Surface (non-saline) in the microbiome compositions, followed by the oral cavity, and finally the gut. The microbiome compositions of the skin, nose/pharynx, oral cavity, and gut not only overlapped to a large extent but also had a clear gradient from *in vitro* to *in vivo*. On the one hand, the oral cavity communicates microorganisms with the environment in close contact, and on the other hand, oral microorganisms also have a great chance to enter and colonize the intestinal tract along with a large amount of swallowed saliva. These results point toward contact and shared environments being the driving factors of microbial transmission, which is consistent with previous research (Mukherjee et al., 2021). In conclusion, these results emphasize that the oral microbiome is an important link between the environmental microbiome and the human microbiome.

MATERIALS AND METHODS

Human Oral Sample Data Acquisition Based on Earth Microbiome Project Data

The EMP developed a unified standard workflow that leveraged existing sample and data reporting standards to allow biomass and metadata collection across diverse environments on Earth (Thompson et al., 2017). The samples submitted by the global community of microbial ecologists were used to perform the microbiome analysis. DNA extraction and 16S rRNA amplicon sequencing were performed using EMP

standard protocols (Caporaso et al., 2011b). The sequence data were error-filtered and trimmed to the length of the shortest sequencing run (90 bp) using Deblur software (Amir et al., 2017).

The EMP data contain a total of 97 studies and 27,742 samples, which are available at http://ftp.microbio.me/emp/release1. We acquired 404 human oral saliva samples from the EMP study to analyze their microbial diversity (**Supplementary Data**). These oral samples are from 5 independent studies and include the populations from Italy, Puerto Rico, and United States (Caporaso et al., 2011a,b; Piombino et al., 2014). We also selected 216 gut, 253 nasal/pharyngeal and 346 skin samples from the EMP study to proceed with the compared analysis (Caporaso et al., 2011a,b; Lax et al., 2014; Vitaglione et al., 2015).

Earth Microbiome Project Ontology Classification

The EMP classified the samples in different environments into the corresponding environmental labels (Thompson et al., 2017). The EMPO classified the microbial environments (level 3) as free-living or host-associated (level 1) and saline or non-saline (if free-living) or animal or plant (if host-associated) (level 2). A subset containing 10,000 samples was then generated that gave equal (as possible) representation across environments (EMPO level 3) and across studies within those environments. In this subset, each sample must have $\geq$5000 observations in the Deblur 90 bp observation table.

Comparison Against Reference Databases and Core Diversity Analyses

The representative sequences of OTUs were analyzed by the Ribosomal Database Project Classifier algorithm using a confidence threshold of 50% against the Silva 16S rRNA gene database (Quast et al., 2013; Cole et al., 2014).

The alpha diversity was computed with the input Deblur 90 bp BIOM table rarefied to 5000 observations for each sample. The alpha diversity included observed OTUs (number of unique tag sequences), Shannon index (Shannon diversity index), Chao1 index, and Faith's PD value (Shannon, 1948; Chao, 1984; Faith, 1992).

The clustering of samples was conducted due to storage conditions by PCoA based on Bray-Curtis similarity distance. The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) clustering was based on Bray-Curtis similarity distance. Bootstrapping with 1,000 resamplings was performed to determine the robustness of the clustering. All these analyses were performed with the statistical software PAST (Hammer et al., 2001).

Significant differences in all analyses were evaluated using ANOVA by the software package IBM SPSS Statistics.

Co-occurrence Network

We selected the 500 most abundant OTUs of the environment types and performed pairwise calculations of the Pearson's r and p-values associated with relative abundance using the 'psych' package in R. Values of |Pearson's r|>0.5 and $p<0.05$ were considered to indicate valid relationship. The network topological features were calculated using Gephi.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

JLW, JNW, and WC conceived and designed the analysis. JLW, JF, DL, and JNW collected the data and performed the experiments. JLW and JNW performed the analysis. JLW, YZ, JNW, and WC wrote the manuscript. All authors reviewed and agreed with the content of the manuscript.

FUNDING

This work was supported by the Medicine and Health Science Technology Development Project of Shandong Province (2019WS522) (JLW) and the China Postdoctoral Science Foundation (2021M701987) (DL).

ACKNOWLEDGMENTS

We thank Zheng Zhang (Shandong University) for meaningful discussion in a draft of this manuscript. Part of the results published in this manuscript are accessible online under the preprint publication (Wang et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.931065/full#supplementary-material

REFERENCES

Amir, A., Mcdonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech, et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2:e00191-16. doi: 10.1128/mSystems.00191-16

Atarashi, K., Suda, W., Luo, C., Kawaguchi, T., Motoo, I., Narushima, S., et al. (2017). Ectopic colonization of oral bacteria in the intestine drives TH1 cell induction and inflammation. *Science* 358, 359–365. doi: 10.1126/science.aan4526

Bacali, C., Vulturar, R., Buduru, S., Cozma, A., Fodor, A., Chis, A., et al. (2022). Oral microbiome: getting to know and befriend neighbors, a biological approach. *Biomedicines* 10:671. doi: 10.3390/biomedicines10030671

Blod, C., Schlichting, N., Schulin, S., Suttkus, A., Peukert, N., Stingu, C. S., et al. (2018). The oral microbiome-the relevant reservoir for acute pediatric

appendicitis? *Int. J. Colorectal Dis.* 33, 209–218. doi: 10.1007/s00384-017-2948-8

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., et al. (2011a). Moving pictures of the human microbiome. *Genome Biol.* 12:R50. doi: 10.1186/gb-2011-12-5-r50

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011b). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U S A.* 108(Suppl. 1), 4516–4522. doi: 10.1073/pnas.1000080107

Chao, A. (1984). Nonparametric-Estimation of the number of classes in a population. *Scand. J. Statistics* 11, 265–270.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., Mcgarrell, D. M., Sun, Y., et al. (2014). ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642. doi: 10.1093/nar/gkt1244

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science* 326, 1694–1697. doi: 10.1126/science.1177486

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., et al. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U S A.* 107, 11971–11975. doi: 10.1073/pnas.1002601107

Escapa, I. F., Chen, T., Huang, Y., Gajare, P., Dewhirst, F. E., and Lemon, K. P. (2018). New insights into human nostril microbiome from the expanded human oral microbiome database (eHOMD): a resource for the microbiome of the human aerodigestive tract. *mSystems* 3:e00187-18. doi: 10.1128/mSystems.00187-18

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* 61, 1–10.

Faust, K., and Raes, J. (2012). Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* 10, 538–550. doi: 10.1038/nrmicro2832

Gao, L., Xu, T., Huang, G., Jiang, S., Gu, Y., and Chen, F. (2018). Oral microbiomes: more and more importance in oral cavity and whole body. *Protein Cell* 9, 488–500. doi: 10.1007/s13238-018-0548-1

Gilbert, J. A., Jansson, J. K., and Knight, R. (2014). The Earth microbiome project: successes and aspirations. *BMC Biol.* 12:69. doi: 10.1186/s12915-014-0069-1

Gilbert, J. A., Meyer, F., Antonopoulos, D., Balaji, P., Brown, C. T., Brown, C. T., et al. (2010). Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. *Stand. Genomic Sci.* 3, 243–248. doi: 10.4056/sigs.1433550

Hammer, Ø, Harper, D. A., and Ryan, P. D. (2001). PAST: paleontological statistics software package for education and data analysis. *Palaeontol. Electron.* 4:9.

He, J., Li, Y., Cao, Y., Xue, J., and Zhou, X. (2015). The oral microbiome diversity and its relation to human diseases. *Folia Microbiol. (Praha)* 60, 69–80. doi: 10.1007/s12223-014-0342-2

Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., and Whiteley, M. (2014). Metatranscriptomics of the human oral microbiome during health and disease. *mBio* 5:e01012-14. doi: 10.1128/mBio.01012-14

Lax, S., Smith, D. P., Hampton-Marcell, J., Owens, S. M., Handley, K. M., Scott, N. M., et al. (2014). Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* 345, 1048–1052. doi: 10.1126/science.1254529

Lira-Junior, R., and Bostrom, E. A. (2018). Oral-gut connection: one step closer to an integrated view of the gastrointestinal tract? *Mucosal Immunol.* 11, 316–318. doi: 10.1038/mi.2017.116

Mark Welch, J. L., Dewhirst, F. E., and Borisy, G. G. (2019). Biogeography of the oral microbiome: the site-specialist hypothesis. *Annu. Rev. Microbiol.* 73, 335–358. doi: 10.1146/annurev-micro-090817-062503

Mark Welch, J. L., Ramirez-Puebla, S. T., and Borisy, G. G. (2020). Oral microbiome geography: micron-scale habitat and niche. *Cell Host Microbe* 28, 160–168. doi: 10.1016/j.chom.2020.07.009

Mukherjee, C., Moyer, C. O., Steinkamp, H. M., Hashmi, S. B., Beall, C. J., Guo, X., et al. (2021). Acquisition of oral microbiota is driven by environment, not host genetics. *Microbiome* 9:54. doi: 10.1186/s40168-020-00986-8

Peters, B. A., Wu, J., Pei, Z., Yang, L., Purdue, M. P., Freedman, N. D., et al. (2017). Oral microbiome composition reflects prospective risk for esophageal Cancers. *Cancer Res.* 77, 6777–6787. doi: 10.1158/0008-5472.CAN-17-1296

Philip, N., Suneja, B., and Walsh, L. J. (2018). Ecological approaches to dental caries prevention: paradigm shift or shibboleth? *Caries Res.* 52, 153–165. doi: 10.1159/000484985

Piombino, P., Genovese, A., Esposito, S., Moio, L., Cutolo, P. P., Chambery, A., et al. (2014). Saliva from obese individuals suppresses the release of aroma compounds from wine. *PLoS One* 9:e85611. doi: 10.1371/journal.pone.0085611

Plaza-Diaz, J., Ruiz-Ojeda, F. J., Gil-Campos, M., and Gil, A. (2018). Immune-Mediated mechanisms of action of probiotics and synbiotics in treating pediatric intestinal diseases. *Nutrients* 10:42. doi: 10.3390/nu10010042

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Reddy, R. M., Weir, W. B., Barnett, S., Heiden, B. T., Orringer, M. B., Lin, J., et al. (2018). Increased variance in oral and gastric microbiome correlates with esophagectomy anastomotic leak. *Ann. Thorac. Surg.* 105, 865–870. doi: 10.1016/j.athoracsur.2017.08.061

Saikaly, S. K., Saikaly, T. S., and Saikaly, L. E. (2018). Recurrent aphthous ulceration: a review of potential causes and novel treatments. *J. Dermatolog. Treat.* 29, 542–552. doi: 10.1080/09546634.2017.1422079

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* 27, 379–423.

Shaw, L., Ribeiro, A. L. R., Levine, A. P., Pontikos, N., Balloux, F., Segal, A. W., et al. (2017). The human salivary microbiome is shaped by shared environment rather than genetics: evidence from a large family of closely related individuals. *mBio* 8:e01237-17. doi: 10.1128/mBio.01237-17

Soffritti, I., D'accolti, M., Fabbri, C., Passaro, A., Manfredini, R., Zuliani, G., et al. (2021). Oral microbiome dysbiosis is associated with symptoms severity and local immune/inflammatory response in COVID-19 patients: a cross-sectional study. *Front. Microbiol.* 12:687513. doi: 10.3389/fmicb.2021.687513

Thompson, L. R., Sanders, J. G., Mcdonald, D., Amir, A., Ladau, J., Locey, K. J., et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551, 457–463. doi: 10.1038/nature24621

Vitaglione, P., Mennella, I., Ferracane, R., Rivellese, A. A., Giacco, R., Ercolini, D., et al. (2015). Whole-grain wheat consumption reduces inflammation in a randomized controlled trial on overweight and obese subjects with unhealthy dietary and lifestyle behaviors: role of polyphenols bound to cereal dietary fiber. *Am. J. Clin. Nutr.* 101, 251–261. doi: 10.3945/ajcn.114.088120

Wang, J., Li, D., Wang, J., and Zhang, Z. (2019). Human oral microbiome characterization and its association with environmental microbiome revealed by the Earth Microbiome Project. *bioRxiv [preprint]* doi: 10.1101/732123

Wasfi, R., Abd El-Rahman, O. A., Zafer, M. M., and Ashour, H. M. (2018). Probiotic *Lactobacillus sp*. inhibit growth, biofilm formation and gene expression of caries-inducing *Streptococcus mutans*. *J. Cell Mol. Med.* 22, 1972–1983. doi: 10.1111/jcmm.13496

Willis, J. R., Gonzalez-Torres, P., Pittis, A. A., Bejarano, L. A., Cozzuto, L., Andreu-Somavilla, N., et al. (2018). Citizen science charts two major "stomatotypes" in the oral microbiome of adolescents and reveals links with habits and drinking water composition. *Microbiome* 6:218. doi: 10.1186/s40168-018-0592-3

Zarco, M. F., Vess, T. J., and Ginsburg, G. S. (2012). The oral microbiome in health and disease and the potential impact on personalized dental medicine. *Oral. Dis.* 18, 109–120. doi: 10.1111/j.1601-0825.2011.01851.x

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

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 14 June 2022
doi: 10.3389/fmicb.2022.919169

The Species Identification and Genomic Analysis of *Haemobacillus shengwangii*: A Novel Pathogenic Bacterium Isolated From a Critically Ill Patient With Bloodstream Infection

Yingying Du[1†], Xuming Li[2†], Yuhao Liu[1†], Shikui Mu[1], Dandan Shen[3], Shu Fan[2], Zheng Lou[2], Shouqin Zhang[1], Han Xia[2], Yinghua Yuan[3*] and Sheng Wang[1*]*

[1] Department of Critical Care Medicine, School of Medicine, Shanghai Tenth People's Hospital, Tongji University, Shanghai, China, [2] Department of Scientific Affairs, Hugo Biotech Co., Ltd., Beijing, China, [3] Department of Clinical Microbiology, School of Medicine, Shanghai Tenth People's Hospital, Tongji University, Shanghai, China

OPEN ACCESS

Edited by:
Zheng Zhang,
Shandong University, China

Reviewed by:
Quan Shen,
Jiangsu University, China
Fen Qu,
Aviation General Hospital of China Medical University, China
Vibhor Tak,
AIIMS Jodhpur, India

***Correspondence:**
Han Xia
xiahan@hugobiotech.com
Yinghua Yuan
yuan_yinghua@126.com
Sheng Wang
wangsheng@tongji.edu.cn

[†] These authors have contributed equally to this work

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

Received: 13 April 2022
Accepted: 17 May 2022
Published: 14 June 2022

Citation:
Du Y, Li X, Liu Y, Mu S, Shen D, Fan S, Lou Z, Zhang S, Xia H, Yuan Y and Wang S (2022) The Species Identification and Genomic Analysis of Haemobacillus shengwangii: A Novel Pathogenic Bacterium Isolated From a Critically Ill Patient With Bloodstream Infection. Front. Microbiol. 13:919169. doi: 10.3389/fmicb.2022.919169

Since the first strain related to Thermicanaceae was reported in 1999, almost no literature on Thermicanaceae is available, particularly its genomics. We recently isolated a novel pathogenic bacterium, the $^{\Delta}$ strain DYY3, from the blood sample of a critically ill patient. The morphological, physiological, and biochemical characteristics of $^{\Delta}$ strain DYY3 were presented in this study, and the virulence factor genes and antibiotic resistance of DYY3 were also determined. Interestingly, the average nucleotide identity (ANI) and core-genes average amino acid identity (cAAI) analysis indicated that $^{\Delta}$ strain DYY3 was genus novel and species novel. Moreover, phylogenetic analysis based on both 16S rRNA gene and whole genomic core gene sequences suggested that $^{\Delta}$ strain DYY3 belonged to the family Thermicanaceae, and this novel taxon was thus named *Haemobacillus shengwangii* gen. nov., sp. nov. Besides, both the whole genome-based phylogenetic tree and amino acid identity analysis indicated that *Thermicanus aegyptius*, *Hydrogenibacillus schlegelii*, *Brockia lithotrophica*, and the newly discovered species *H. shengwangii* should belong to Thermicanaceae at the family level, and *T. aegyptius* was the closest species to *H. shengwangii*. We also constructed the first high-quality genome in the family Thermicanaceae using the next-generation sequencing (NGS) and single-molecule real-time (SMRT) sequencing technologies, which certainly contributed to further genomics studies and metagenomic-based pathogenic detection in the future.

Keywords: catheter-associated bloodstream infection, pathogenic bacterium, novel species identification, Thermicanaceae, genome *de novo* assembly, single-molecule real-time sequencing, comparative genomics

INTRODUCTION

Catheter-related bloodstream infection (CRBSI) is a frequent and life-threatening condition in the intensive care unit (ICU), which is associated with increased morbidity, mortality, and healthcare costs (Schwab et al., 2018; Gerver et al., 2020; Zeng et al., 2021). For example, according to a prospective multi-center study in China, the average incidence of CRBSI and the mortality due to CRBSI in ICU were 1.5/1,000 catheter days and 18.09%, respectively (Zeng et al., 2021). A similar

incidence of CRBSI was reported in studies from European countries (Schwab et al., 2018; Gerver et al., 2020), and even higher rates were found in developing countries, up to 5.3/1,000 catheter days with 28–30% of mortality (Rosenthal et al., 2021).

To improve the clinical outcomes of patients with CRBSI, a rapid and accurate diagnosis of the causative pathogen is a critical step (Zhong et al., 2021). The current guideline recommends diagnosing CRBSI by hemoculture in suspected patients (Mermel et al., 2009), but this approach takes 48–96 h to isolate, identify, and perform antibiotic susceptibility tests (Tabak et al., 2018). Furthermore, it is often challenging to culture many fastidious or uncultivatable pathogens in standard automated systems (Murray and Masur, 2012). Therefore, emerging technologies, especially high-throughput sequencing, were attempted to replace conventional culture-based methods and initiate timely targeted anti-infection therapy (Li et al., 2021). However, few clinical studies thus far identified and classified unknown species with extraordinary genetic distances from known species.

The only reported bacteria that might belong to the family Thermicanaceae were *Thermicanus aegyptius*, *Hydrogenibacillus schlegelii*, *Brockia lithotrophica*, and *Carbobacillus altaicus*. *T. aegyptius* was first identified from soil and described as a fermentative microaerophile in 1999 (Gossner et al., 1999), and the reference genome was *T. aegyptius* DSM 12793, available online in 2013. *H. schlegelii* was originally named *Bacillus schlegelii* in 1979 (Schenk and Aragno, 1979) and was transferred to be a novel genus due to its massive divergence from other species in the genus *Bacillus* in 2013 (Kampfer et al., 2013). *H. schlegelii* was known for its ability of hydrogen-oxidizing (Barbosa et al., 2020) and was classified into order Bacillales and family Bacillaceae with NCBI taxonomy ID of 1484. *B. lithotrophica* was isolated from a hot spring in Russia and reported as a new taxon in 2013 (Perevalova et al., 2013). Finally, *C. altaicus* was still a candidate taxon classified into order Bacillales and family Bacillales *incertae sedis* with NCBI taxonomy ID 2163959. However, whether the classification method of the above bacteria is correct needs to be verified by genomic analysis, as this approach is increasingly being accepted as reliable data for bacterial taxonomy and species identification (Hayashi Sant'Anna et al., 2019).

In this study, the $^{\Delta}$ strain DYY3, isolated from the blood sample of a critically ill patient diagnosed with CRBSI, cannot be identified by VITEK-MS automatic microbiological analyzer and the 16S rRNA sequence analysis. Thus, the high-quality genome of $^{\Delta}$ strain DYY3 was constructed by the next-generation sequencing (NGS) and single-molecule real-time (SMRT) sequencing technologies, and multiple comparative genomics analyses were applied to identify this new strain.

MATERIALS AND METHODS

Case Report

In January 2021, a 68-year-old female patient was admitted to the Shanghai Tenth People's Hospital ICU due to acute respiratory failure, aspiration pneumonia, and cerebral infarction. Invasive mechanical ventilation, femoral vein catheterization, and urinary catheterization were performed during the treatment, and pulmonary infection was verified by fever, cough, and chest CT scanning. The patient's infection was effectively controlled initially by the empirical use of ceftazidime. However, the patient's body temperature, leukocyte count, and C-reactive protein were raised again after 9 days of anti-infection treatment. Considering the possibility of CRBSI, the femoral vein catheter was removed immediately, blood samples and the terminal of the central venous catheter were collected for bacterial culture, and vancomycin hydrochloride was added empirically to strengthen the anti-infection treatment. Three days later, both blood culture and catheter culture suggested unrecognized Gram-positive bacterial infection, and the drug sensitivity test showed that vancomycin was sensitive. Thus, vancomycin continued to be used, and the patient's bloodstream infection was cured in 2 weeks. The strain designated DYY3 was isolated from the blood sample and preserved in a -80°C refrigerator to identify the unknown pathogenic bacterium further.

Strain Isolation

The $^{\Delta}$ strain DYY3 was isolated from a critically ill patient's blood sample with a catheter-associated bloodstream infection. Briefly, the blood specimen was inoculated in a blood culture bottle (BD BACTEC Plus aerobic/F Culture Vials, Becton, Dickinson and Company, United States) at 35°C until it showed a positive result. For the isolation of $^{\Delta}$ strain DYY3, blood agar plates (bioMérieux, Marcy l'Etoile, France) were used, and the plates were incubated

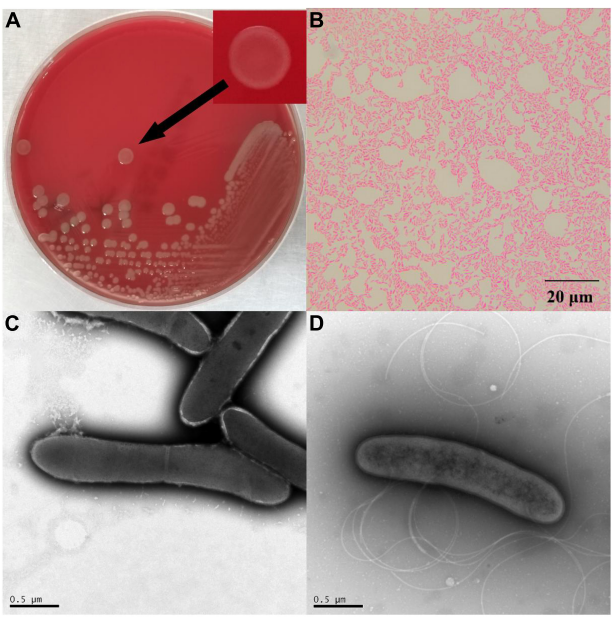

FIGURE 1 | The morphology of strain DYY3. **(A)** The cultured bacterial colonial morphology on blood agar plate of $^{\Delta}$ strain DYY3; **(B)** the Gram staining result shows that $^{\Delta}$ strain DYY3 is Gram weak positive; **(C)** the electron microscope photograph of the dividing cells of $^{\Delta}$ strain DYY3 after 24 h culture; **(D)** the electron microscope photograph shows the flagellum of $^{\Delta}$ strain DYY3.

TABLE 1 | Characteristics of *Haemobacillus shengwangii* and close species.

Characteristics	*Haemobacillus shengwangii* DYY3	*Thermicanus aegyptius* DSMZ 12793^T	*Brockia lithotrophica* Kam1851^T	*Hydrogenibacillus schlegelii* DSM 2000^T
Cell shape	Rod	Rod	Rod	Rod
Gram reaction	Week positive	Week positive	Positive	Positive
Motility	+	+	+	+
Spores	+	-	+	+
Flagellum	+	+	+	+
Colony color	Grayish white	Geige	NA*	Cream
Temperature for growth				
Range	20–45°C	37–65°C	46–78°C	37–80°C
Optimum	30–37°C	55–60°C	60–65°C	70–75°C
pH for growth				
Range	6–8	5.5–7.7	5.5–8.5	4.2–7.5
Optimum	7	6.5–7	6.5	6–7
NaCl concentration for growth (%, w/v)				
Range	0–2	NA	NA	3–5
Optimum	0	NA	NA	3
Oxygen requirement	Facultative anaerobic	Facultative anaerobic	Strictly aerobic	Strictly aerobic
DNA G + C content (mol%)	40.62	50.3	63	67–68
Major fatty acids	C15:0 iso, C15:0 anteiso, C17:0 iso, C16:0 iso	NA	C16:0, C16:iso, C18:0, C17:0	C16:0 iso
Polar lipids[#]	DPG, PG, PE	NA	NA	DPG, PG
Quinone	MK7	NA	NA	MK7

[#]*DPG, diphosphatidylglycerol; PG, phosphatidylglycerol; PE, phosphatidylethanolamine.*
**NA, not available.*

in a CO_2 incubator for 48 h. Later, dozens of single colonies were picked up from the blood agar plates. VITEK-MS automatic microbiological analyzer (bioMérieux, Marcy l'Etoile, France) was used to identify the taxonomic classification of the $^{\Delta}$ strain DYY3 according to the standard operation process using the VITEK MS IVD KB V3.2 database as the reference. The total length of the 16S rRNA sequence was amplified by PCR using the primers of 27F (5′-AGAGTTTGATCMTGGCTCAG-3′) and 1492R (5′-GGTTACCTTGTTACGACTT-3′), and the amplified fragments were sequenced using a 3730XL sequencer (Applied Biosystems, United States).

Morphology, Physiology, and Chemotaxonomy Analysis

Gram staining of $^{\Delta}$ strain DYY3 was performed, referring to the procedures described by Wagner et al. (2018). The fresh biomass of DYY3 was stained with 1% (w/v) uranyl acetate, and the electron micrograph was taken by a transmission electron microscopy system JEM-1010 (JEOL, Japan). The growth tests were performed at various temperatures, NaCl concentrations, and pH levels using R2A agar plates (Difco, United States) as the culture medium. The tested temperatures were 4, 10, 15, 20, 30, 37, 40, 45, 50, and 55°C. The tested pH ranged from 5.0 to 11.0 with a gradient value of 1, K_2HPO_4/KH_2PO_4 buffer was used for pH 5–8, and $NaHCO_3$/NaOH buffer was used for pH 9–11. The tested NaCl concentration ranged from 0 to 10% w/v using the interval of 1%. Acid production tests, enzyme activity tests, and additional phenotypic tests were performed using API 50CHB, API ZYM, and API 20NE galleries (bioMérieux, Marcy l'Etoile, France), respectively. The utilization of carbon sources was tested using Biolog GPIII Microplates (Biolog, United States), and quinones were extracted and identified using the HPLP LC-20AT system (Shimadzu, Japan). The fresh biomass of $^{\Delta}$ strain DYY3 was hydrolyzed at 120°C for about 12 h to determine the composition of saccharides on its cell wall using ribose, arabinose, glucose, rhamnose, xylose, mannose, and galactose as references. The Cell Fatty Acid-Fatty Acid Methyl Ester (CFA-FAME) components were assayed by Agilent 6890 gas chromatograph (Agilent, United States), and the data were collected by the Sherlock Microbial Identification System (version 6.0, MIDI). The polar lipid analysis of $^{\Delta}$ strain DYY3 (1 g of freeze-dried cells) was performed and examined by thin-layer chromatography (TCL) on cellulose sheets. The spots for polar lipids were identified by spraying with 10% phosphomolybdic acid in ethanol, a-naphthol, and ninhydrin, respectively.

Antibiotic Susceptibility Test

The minimum inhibitory concentrations (MICs) of the $^{\Delta}$ strain DYY3 to penicillin, ampicillin, vancomycin, gentamicin, erythromycin, ciprofloxacin, levofloxacin, clindamycin, trimethoprim/sulfisoxazole, rifampicin, and imipenem were determined by MicroScan Pos Combo Panel Type 33 (MicroScan, United States), with the interpretation of drug sensitivity results

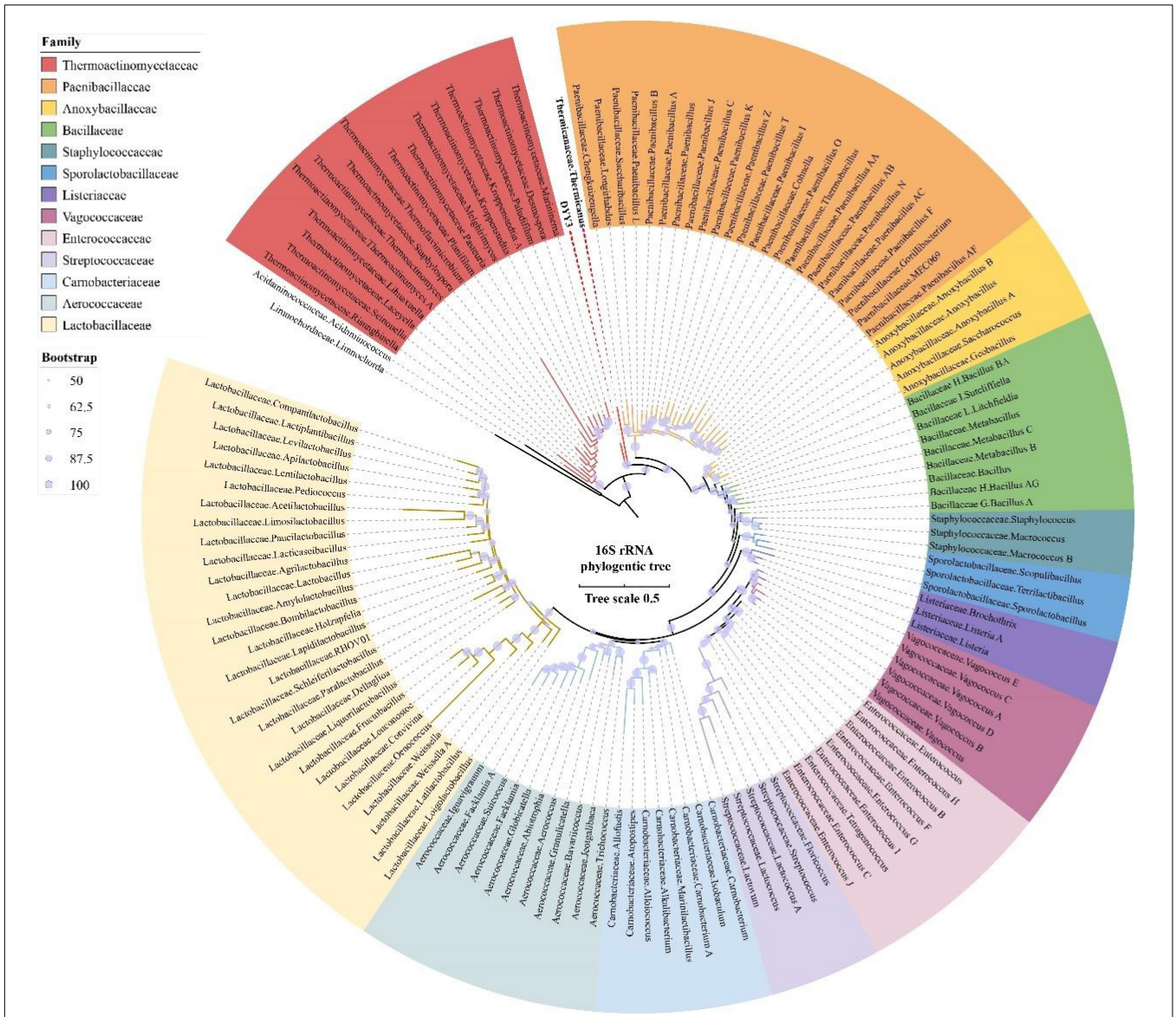

FIGURE 2 | 16S rRNA gene tree of strain DYY3. The rooted 16S rRNA gene ML tree was constructed by RAxML utilizing the typical species of Bacilli and also the matched species on NCBI with a bootstrap value of 1,000. Acidaminococcaceae, Acidaminococcus, Limnochordaceae, and Limnochorda were set as the out-group, and the other species were Bacilli in class.

referred to the Clinical and Laboratory Standards Institute (CLSI) M45 guidelines.

Genome Sequencing, Assembly, and Annotations

A 350-bp paired-end library was constructed and sequenced using the Illumina NovaSeq 6,000 sequencing platform (Illumina Inc., San Diego, CA, United States) with a PE150 layout. A 10-kb SMRT library was constructed and sequenced by the PacBio Sequel system (Pacific Biosciences, United States). The data were assembled by using Unicycler (version 0.4.7), and the genome was annotated by using Prokka (version 1.14.6) with default parameters. Prophages were annotated using phiSpy (version 4.2.12), and genomic islands were identified by using Islanpath-DIMOB (version 1.0.6). The BUSCO database (version 5.2.2) was used to evaluate the completeness of the genome sequence (Manni et al., 2021). BLASTp (version 2.10.1) was used to query the non-redundant (nr) protein sequence database and hit with the highest score, and the identity higher than 60 was recognized as a match for each gene. Eggnog-mapper (version 2.0.1) with the parameter of "seed_ortholog_evalue 1e-5-m diamond" was used to query the eggNOG database. HMMER (version 3.3.2) with the parameter of "-E 1e-5" was used for Pfam (version 33.1) database annotation, and Diamond (version 0.9.24.125) with the parameter of "-e 1e-5" was used for the Swiss-Port annotation. The genome atlas was plotted

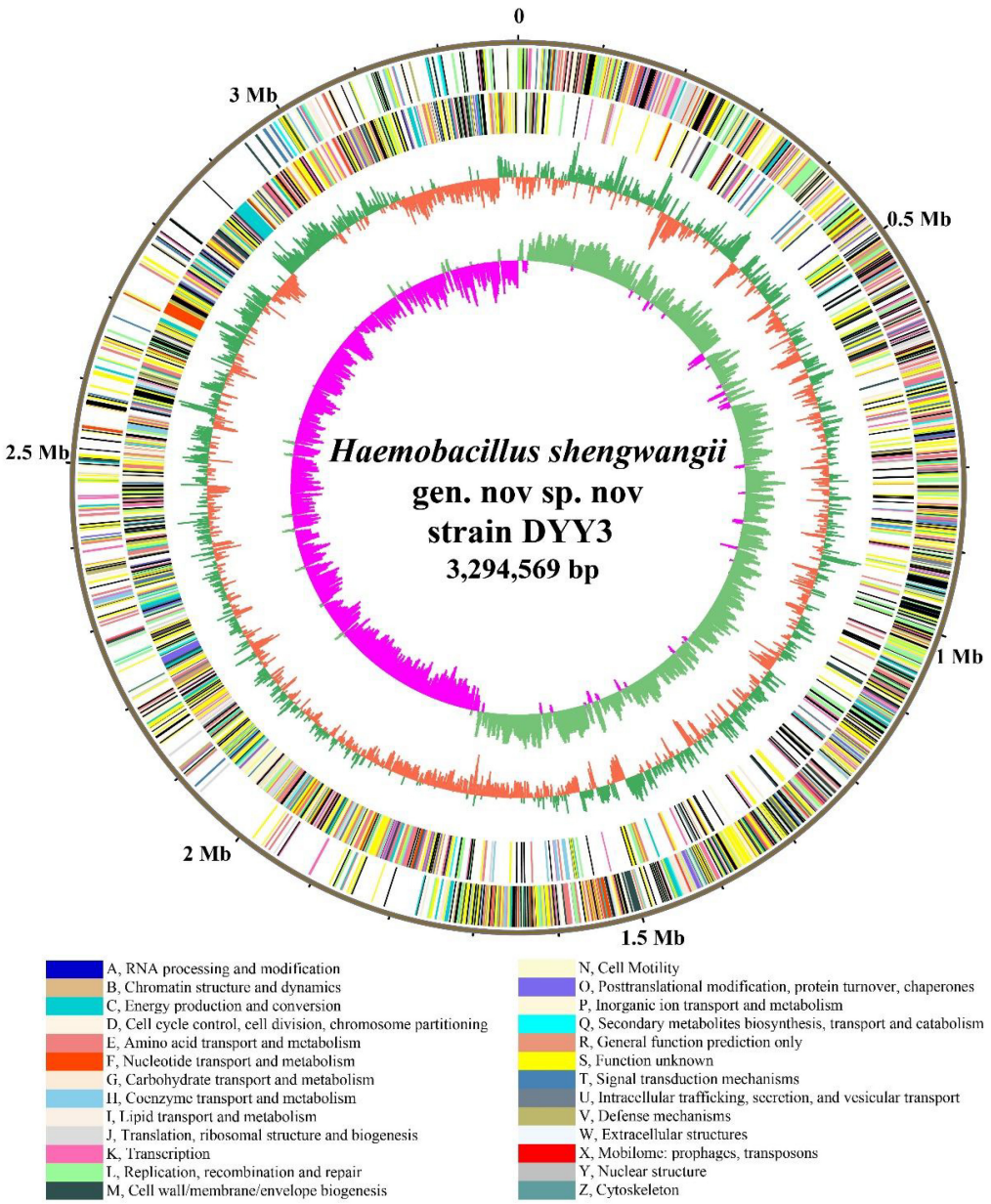

FIGURE 3 | Genome atlas of strain DYY3. The outer black circle shows the genome coordinates, and the next two circles represent forward and reverse strand CDSs with colors representing the functional classification of COG. The last two circles are GC content and GC skew using a 5-kb window overlapping at 1,000 bp. The COG functional classifications and colors are shown at the bottom of the figure. The green and orange colors of the fourth circle mean the GC content is higher and lower than the average GC content of the genome, respectively. The inner circle's purple and light blue colors show the GC-skew values lower and higher than 0, respectively.

using CIRCOS.[1] The virulence factor genes were predicted by querying VFDB using a web-based VFanalyzer. Antibiotic resistance genes were annotated by the CARD (version 3.1.4) database with BLASTp parameter of "-qcov_hsp_perc 80," and hits with an identity less than 80 were filtered. The pathogen–host interaction associated genes were identified by the PHI database (version 4.12) with Diamond (version 0.9.24.125) parameter of "-e 1e-5" and hits with an identity less than 50 were filtered.

Phylogenetic Relationship Analysis

The 16S rRNA maximum likelihood (ML) phylogenetic tree was constructed by RAxML[2] using the GTR substitution matrix model. The whole genomic tree was constructed

[1]http://circos.ca/

[2]https://raxml-ng.vital-it.ch/#/

by IQ-TREE (version 1.6.12) with a bootstrap value of 1,000, referring to the core genes of the Genome Taxonomy Database (GTDB) (release202) (Rinke et al., 2021). One genome was selected as a representative for each genus of the principal families in class Bacilli. Two species from the family Acidaminococcaceae (phylum Firmicutes and class Negativicutes) and family Limnochordaceae (phylum Firmicutes and class Limnochordia) were set as out-group. The phylogenetic trees were plotted using iTOL.[3] The average nucleotide identity (ANI) values and the values of the pairwise core-genes average amino acid identity (cAAI) were calculated using fastANI (Jain et al., 2018) and CompareM (version 0.1.2)[4] with default parameters, respectively. Genes presented over 90% of genomes were used for cAAI calculation. The digital DNA-DNA hybridization (dDDH) values were calculated using a genome-to-genome distance calculator (GGDC)[5].

RESULTS

Morphology, Physiology, and Chemotaxonomy

After 48 h of cultivation at 30°C, the colonies of $^{\Delta}$ strain DYY3 on blood agar were 1–2 mm in size, grayish-white, round, smooth, and moist. The cells were weakly Gram-positive, rod-shaped, about 1.4–2.2 μm in length, and 0.4–0.5 μm in width, with flagellum and spore, motile, and facultative anaerobic (**Figure 1**). The $^{\Delta}$ strain DYY3 grew well on R2A blood agar plates with a temperature range of 20–45°C (preferred 30–37°C). The strain can grow in NaCl concentration ranging from 0 to 2% w/v but not in NaCl concentration over 2%. In addition, the strain can also grow at a pH of 6–8 with an optimum of 7.

API galleries tests revealed inactive biochemical reactions of $^{\Delta}$ strain DYY3 (**Supplementary Table 1**). The API ZYM assay indicated that alkaline phosphatase, esterase (C4), lipid esterase (C8), leucine aromatase, pancreatic rennet, acid phosphatase, and naphthol-AS-Bl-phosphate hydrolase tests were positive. The API 20NE assay suggested a positive assimilation test for glucose, arabinose, mannitol, mannose, N-acetyl glucosamine, maltose, gluconate, capric acid, adipic acid, malic acid, citric acid, and phenylacetic acid. However, the $^{\Delta}$ strain DYY3 was only positive with 5-keto-gluconate in the API 50CHB assay.

Diphosphatidylglycerol (DPG), phosphatidylethanolamine (PE), and phosphatidylglycerol (PG) were identified to be the main polar lipids of $^{\Delta}$ strain DYY3, and the two-dimensional TCL of the polar lipids photographs was shown in **Supplementary Figure 1**. The $^{\Delta}$ strain DYY3 had no typical saccharides in the whole-cell hydrolysate experiment (**Supplementary Figure 2**). MK7 was the main methylanthraquinone in $^{\Delta}$ strain DYY3. The fatty acid analysis revealed that DYY3 synthesized mainly iso- and anteiso-branched saturated fatty acids, mainly including C15:0 iso (61.75%), C15:0 anteiso (13.29%), C17:0 iso (3.56%), and C16:0 iso (3.31%), and a spot of unsaturated fatty acids C17:1 iso w10c (3.54%). According to the RTSBA6 (version 6.21) database, *Bacillus* was the closest genus, but the similarities were not high (26.70%; refer to **Supplementary Tables 2A,B**). Carbon source utilization assays showed that $^{\Delta}$ strain DYY3 only used D-serine and glucuronamide as carbon sources (**Supplementary Table 3**). The VITEK-MS typing results cannot assign $^{\Delta}$ strain DYY3 to any species in the database, but later genetic analysis (see below) suggested that DYY3 belonged to the family Thermicanaceae. Hence, the morphological, physiological, and biochemical characteristics of the four species in the family Thermicanaceae are summarized in **Table 1**.

16S rRNA Phylogenetic Tree

To identify the phylogenetic relationship of $^{\Delta}$ strain DYY3, we first amplified and sequenced its full-length region of 16S rRNA genes. Then, we queried the 16S rRNA sequence at NCBI using online BLASTn. The result showed that the taxon of $^{\Delta}$ strain DYY3 was close to *Bacillus* at the genus level, and belonged to *Bacillus* at class level, and the best matching degree was 94.49%, and all the hit taxon belonged to Bacilli in class. Later, we extracted the entire length of the 16S rRNA gene sequence to construct a phylogenetic tree, utilizing the typical Bacilli species and the matched species on NCBI (**Figure 2**). The sequence information, blast results, and sequence similarity are shown in **Supplementary Table 4**. The phylogenetic tree showed that $^{\Delta}$ strain DYY3 was close to *Thermicanus* at the genus level and might belong to Thermicanaceae at the family level. However, this new taxon had six copies of the 16S rRNA gene, which were

TABLE 2 | Genome features of strain DYY3.

Genome features	Values
Chromosome	1 (circular)
Genome size	3,294,569
Genome coverage (NGS)	577
Genome coverage (TGS)	256
G + C content (%)	40.62%
BUSCO	100%
rRNAs (5, 16, 23S)	18 (6, 6, 6)
tRNAs	72 (35 families)
CDS genes	3,347
Genomic Island	13

TABLE 3 | Gene functional annotation of strain DYY3.

Database	Annotated gene number	Percentage
Nr	1,834	54.80%
Pfam	2,685	80.22%
Swiss-Prot	2,034	60.77%
EggNOG	2,935	87.69%
GO	697	20.82%
KEGG	1,990	59.46%
COG	2,741	81.89%
Total	3,003	89.72%

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

[4]http://github.com/dparks1134/CompareM

[5]http://ggdc.dsmz.de/

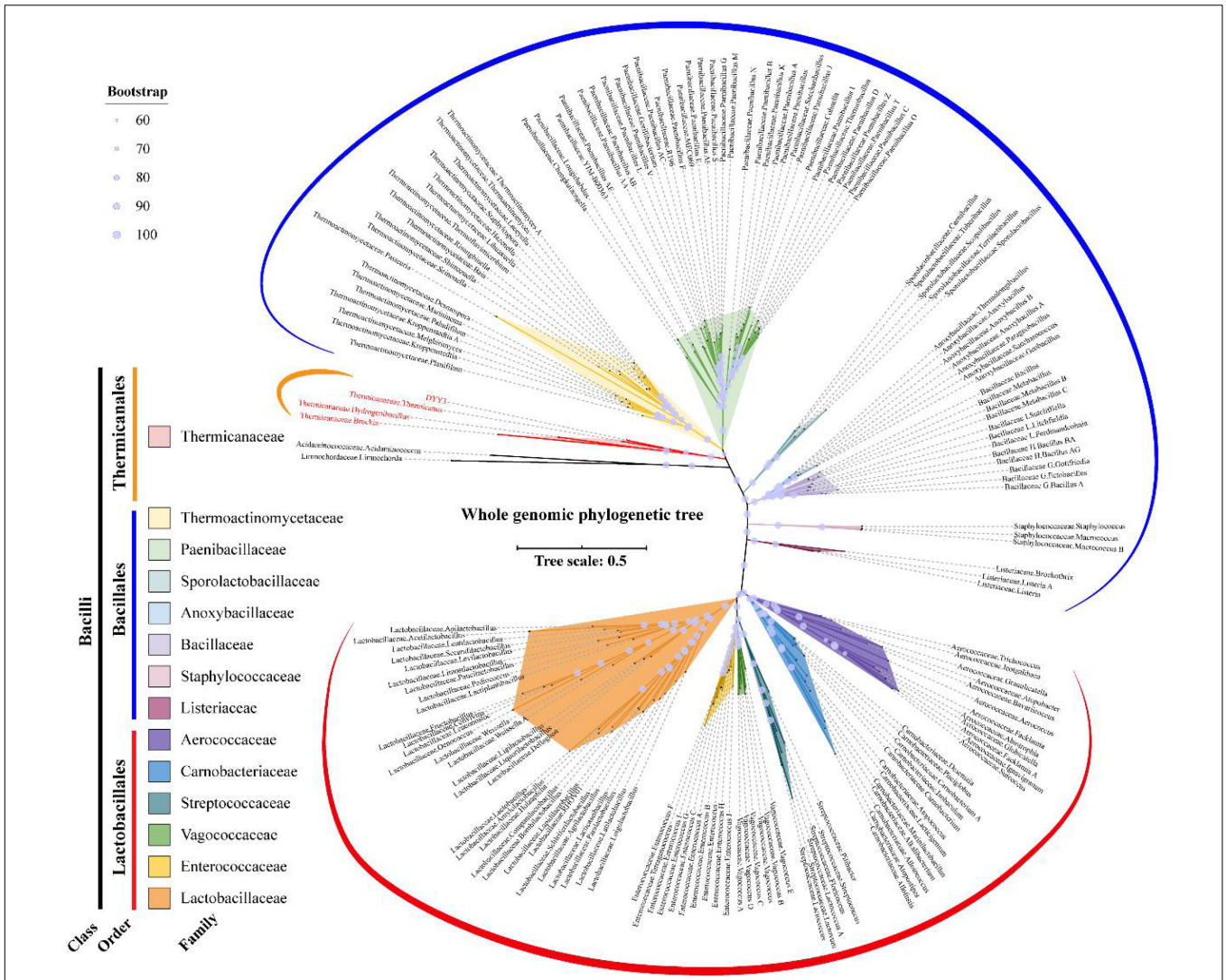

FIGURE 4 | Whole genome-based phylogenetic relationship of strain DYY3. The unrooted whole genomic maximum likelihood tree was constructed by IQ-TREE using GTDB-based 120 conserved core genes with a bootstrap value of 1,000. Acidaminococcaceae, Acidaminococcus, Limnochordaceae, and Limnochorda were set as the out-group, and the other species were Bacilli in class. The 14 families of the Bacilli class were differently colored, and the bootstrap values for each node were shown by the size of a light blue circle.

100% identical. Therefore, contradictory conclusions were drawn by the online NCBI blast results and sequence information of the 16S rRNA phylogenetic tree; thus, $^{\Delta}$ strain DYY3 cannot be assigned to any existing species or genus.

Genome Features

We combined NGS and SMRT technologies to construct the genome of $^{\Delta}$ strain DYY3 (**Supplementary Tables 5A,B**), and a circular genome with a total length of 3,294,569 bp and Guanine and Cytosine content of 40.62% was obtained (**Figure 3** and **Table 2**). All conserved BUSCO genes (100%) were identified within the genome (**Supplementary Table 6**), indicating the high quality of this constructed genome. A total of 3,264 CDSs with an average length of 861 bp, 18 rRNAs, and 72 tRNAs were predicted. Besides, thirteen genomic islands (**Supplementary Table 7**) and two partial prophage regions (**Supplementary Table 8**) were identified in the genome. A total of 3,003 (89.72%) coding seuqence could be assigned with annotations, and 366 (10.94%) CDSs, including 22 database-based annotations, were annotated as hypothetical proteins (**Table 3** and **Supplementary Tables 9A–D**).

Phylogenetic Relationship

As shown in **Figure 4**, the whole genome-based phylogenetic tree was constructed, employing the protein sequence of the 120 conserved genes from the GTDB database, and 165 strains, including all the possible and available genomes of the Thermicanaceae family, were selected. The genomes' source, information, and dDDH are listed in **Supplementary Table 10**. This whole genome-based phylogenetic tree showed high

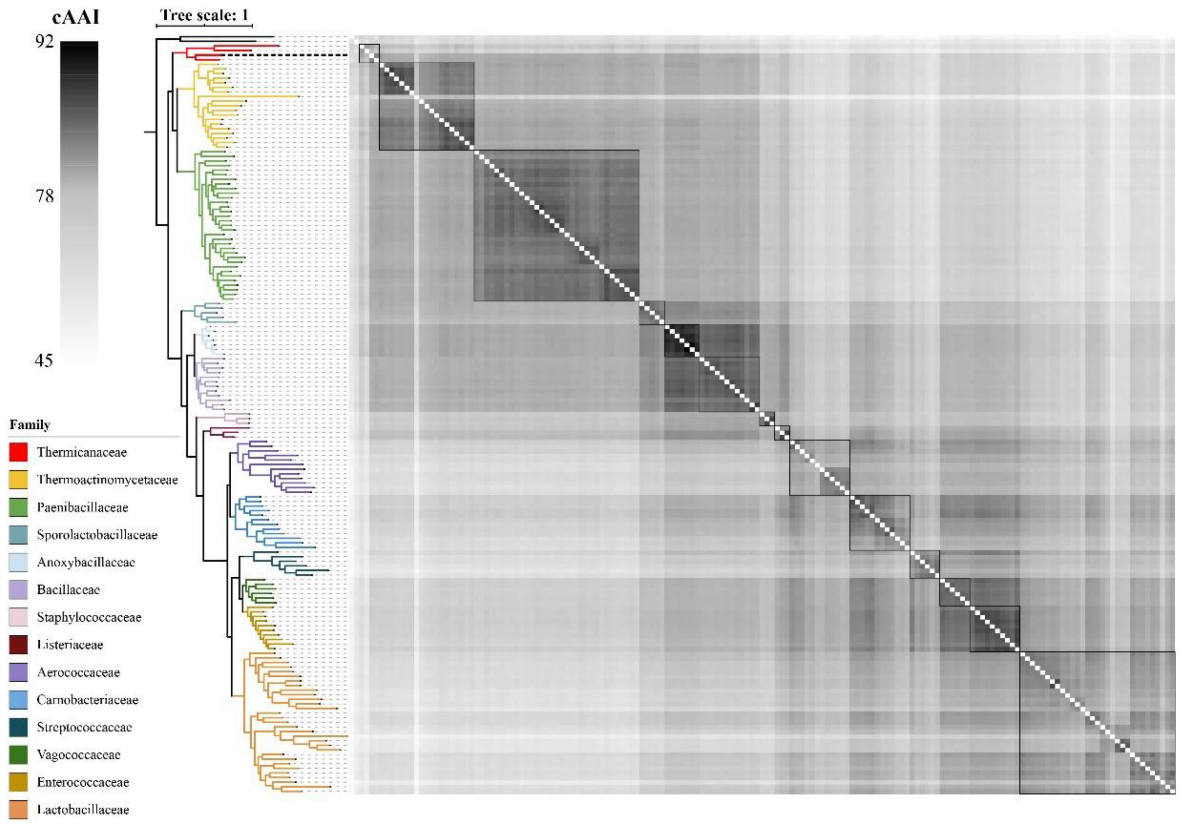

FIGURE 5 | Heatmap depicting the cAAI values of the 165 pairwise comparisons in class Bacilli. The left phylogenetic tree shows the phylogenetic relationships of different families, and the right heatmap shows the cross matrix of the core-genes average amino acid identity values between each genus. The order of each branch is identical to that of **Figure 4**. The black dash line shows $^{\Delta}$strain DYY3.

accordance with the 16S rRNA phylogenetic tree, demonstrating that DYY3 was close to the genus *Thermicanus* and belonged to the same taxon at the family level. Moreover, the sequences of majority groups were significantly distinct from calculating reliable values as the collinear regions were mainly less than 1% of the whole genomes. We then calculated cAAI to show the divergence between each genus, considering that the amino acid sequence is more conserved than the nucleotide sequence of gene coding regions. The plotted heatmap of the cAAI values agreed well with the core genes-based phylogenetic tree (**Figure 5** and **Supplementary Table 11**), suggesting that $^{\Delta}$ strain DYY3 belonged to Thermicanaceae at the family level. The average cAAI value between DYY3 and species from other families was 55.68, and the average cAAI value of each genus within the same existing family was 70.22 (**Table 4**). The cAAI value between *T. aegyptius* and DYY3 was 67.21, close to the average cAAI value for a single family, further demonstrating that $^{\Delta}$ strain DYY3 belonged to Thermicanaceae at the family level.

Finally, we compared the whole genomic similarity within the possible and available genomes of the Thermicanaceae family using BLAST-based ANI (ANIb) (**Table 5**). The ANIb between DYY3 and *T. aegyptius* was calculated to be 69.51%, indicating that they did not belong to the same species, as generally 95% ANI was found to recapitulate the majority species (Jain et al., 2018; Olm et al., 2020; Parks et al., 2020). More importantly, *T. aegyptius*, *B. lithotrophica*, and *H. schlegelii* represented three different genera, and the ANIbs within the four species (e.g., DYY3) ranged from 68.86 to 75.40%, suggesting that $^{\Delta}$ strain DYY3 did not belong to any existing genus. Therefore, $^{\Delta}$ strain DYY3, the novel taxon, was named *Haemobacillus shengwangii* gen. nov., sp. nov. *Haemobacillus* referred to the strain isolated from the blood and was spore-forming and rod-shaped, and *Shengwangii* was named for appreciating the outstanding effort of Doctor Wang to save patients' lives.

Virulence Factors

A total of 30 genes of 16 virulence factor classes were predicted to encode putative virulence factors in $^{\Delta}$ strain DYY3′s genome by the virulence factor database (VFDB). DYY3 owned eight unique predicted virulence factor genes associated with adherence, anti-phagocytosis, immune evasion, intracellular survival, iron uptake, and motility (**Supplementary Table 12**). In addition, twenty genes were predicted associated with hypervirulence in DYY3′s genome using the Pathogen Host Interactions database (PHI-base), with the highest number of the four species in the Thermicanaceae family (**Supplementary Tables 13A,B**).

TABLE 4 | Statistical result of cAAI values between different families in Bacilli class.

Family name	Average cAAI values between DYY3 and each family	Average cAAI values between each genus in a single family	Genomes in each family
Thermoactinomycetaceae	59.59	68.25	19
Paenibacillaceae	60.47	70.38	33
Sporolactobacillaceae	58.73	70.23	5
Anoxybacillaceae	60.67	83.38	7
Bacillaceae	59.76	72.48	4
Staphylococcaceae	54.64	73.17	3
Listeriaceae	56.11	70.51	3
Aerococcaceae	51.74	61.95	12
Carnobacteriaceae	53.04	66.4	12
Streptococcaceae	50.91	66.75	6
Vagococcaceae	53.64	72.11	6
Enterococcaceae	53.65	75.34	10
Lactobacillaceae	50.91	61.86	31
Average	**55.68**	**70.22**	**12**
Min	**50.91**	**61.86**	**3**
Max	**60.67**	**83.38**	**33**
***T. aegyptius* vs. strain DYY3**	**67.21**		

Antibiotic Resistance

Two genes, shared by $^{\Delta}$ strain DYY3 and *T. aegyptius* DSMZ 12793T, were predicted to be associated with antibiotic resistance (**Supplementary Table 14**). Of the two genes, one gene was related to ARO:3003438, annotated as AMR Gene Family associated with elfamycin. Another gene was related to ARO:3002838, annotated as LNU lincosamide nucleotidyltransferase associated with lincosamide. Besides, the MIC test revealed that $^{\Delta}$ strain DYY3 was sensitive to the most commonly used antibiotics, including penicillin, ampicillin, vancomycin, gentamicin, erythromycin, ciprofloxacin, levofloxacin, clindamycin, rifampicin, imipenem, and trimethoprim/sulfamethoxazole (**Supplementary Table 15**).

DISCUSSION

Traditional morphological, physiological, and biochemical studies and comparative genomic analysis demonstrated that the $^{\Delta}$ strain DYY3 was a novel bacterial pathogen belonging to class Bacilli, order Thermicanales, family Thermicanaceae, and genus *Haemobacillus*. We thus proposed *H. shengwangii* gen. nov., sp. nov. to be the name of the novel taxon. In addition, our data supported that *T. aegyptius*, *H. schlegelii*, *B. lithotrophica*, and *H. shengwangii* should belong to Thermicanaceae at a family level. Among the genome published in the family Thermicanaceae, *H. shengwangii* was the first genome constructed with high quality, which undoubtedly contributed to the future research of this family.

Based on 16S rRNA gene sequencing, querying, and phylogenetic tree analysis, $^{\Delta}$ strain DYY3 was identified as a species belonging to class Bacilli, and *T. aegyptius* was the most adjacent taxon. However, when we investigated the taxonomy background of the genus *Thermicanus*, the taxonomic status of this genus was inconsistent at the family level in different authoritative taxonomic databases, including the NCBI taxonomy database, GTDB, SILVA, and GBIF database. We then constructed the core genes-based phylogenetic tree using 165 representative species, including the typical genus of the representative families in class Bacilli and the genus associated with the family Thermicanaceae. The phylogenetic tree indicated that *T. aegyptius*, *H. schlegelii*, *B. lithotrophica*, and $^{\Delta}$ strain DYY3 were in the same tree clade, and the cAAI heatmap further verified their close genetic relationships. Hence, these four species should be classified into the same family. *T. aegyptius* was the closest species to $^{\Delta}$ strain DYY3, supported by the 16S rRNA gene tree and the whole genomic phylogenetic tree. The calculated ANIb (69.51%) between $^{\Delta}$ strain DYY3 and *T. aegyptius* str. DSM 12793 was far from the standard (95%) belonging to the same species, which indicated that $^{\Delta}$ strain DYY3 was a novel species. Since three of the four species in the family Thermicanaceae were known as different genera, the narrow range of the ANIb values (68.86–75.40%) suggested that DYY3 also did not belong to any existing genus. Therefore, $^{\Delta}$ strain DYY3 was verified belonging to a genus novel and species novel taxon in Thermicanaceae. This new taxon was proposed to be named *Haemobacillus shengwangii* gen. nov., sp. nov.

The MIC test showed that $^{\Delta}$ strain DYY3 was sensitive to all commonly used antibiotics. Ceftazidime was initially used in the patient's antibiotic treatment, which turned ineffective. Subsequently, vancomycin was empirically added to the anti-infection therapy, and the symptoms of infection were improved rapidly. Since *H. shengwangii* is a Gram-positive, penicillins, glycopeptides, and cephalosporins should be recommended. In addition, the MIC test also indicated that $^{\Delta}$ strain DYY3 did not have antibiotic resistance, although the CARD database predicted two antibiotic resistance genes associated with elfamycin and

TABLE 5 | The ANIb values and alignment length of strain DYY3 and species in family Thermicanaceae.

Species		Alignment length			
		***H. shengwangii* DYY3**	***T. aegyptius* DSM 12793**	***B. lithotrophica a* DSM 22653**	***H. schlegelii* MA48**
ANIb	**DYY3**		304,571	36,373	65,834
	T. aegyptius	69.51%		77,647	176,981
	B. lithotrophica	69.27%	68.86%		301,196
	H. schlegelii	68.82%	69.75%	75.40%	

lincosamide, respectively. To further clarify the virulence of $^{\Delta}$ strain DYY3, its virulence factor genes were predicted based on sequence similarity comparison against the VFDB. The $^{\Delta}$ strain DYY3 owned eight unique predicted virulence factor genes associated with adherence, anti-phagocytosis, immune evasion, intracellular survival, iron uptake, and motility. Besides, twenty genes were predicted to be associated with hypervirulence in the $^{\Delta}$ strain DYY3 genome using PHI-base, with the highest number of the four species in the Thermicanaceae family. Given that *Bacillus* was the closest genus that can be queried in the database, the phylogenetic relationships of $^{\Delta}$ strain DYY3 and species in genus *Bacillus* were quite distant, and the sequence similarities of $^{\Delta}$ strain DYY3 and species in genus *Bacillus* were relatively low; we cannot exclude the possibility of $^{\Delta}$ strain DYY3 owning more genes associated with virulence factors.

Although blood culture is the golden standard for pathogenic diagnosis, it cannot fully support that $^{\Delta}$ strain DYY3 was the only pathogenetic bacterium responsible for this patient's infection, as only a tiny proportion of pathogens are identifiable by the culture-based methods (Li et al., 2021). mNGS was more suitable for pathogenic detection when the pathogen was unculturable, novel, or variant species, such as DYY3, as all the nucleic acids can be sequenced and analyzed indiscriminately (Wilson et al., 2019; Price et al., 2021), and NGS has been widely used to perform comprehensive and precise diagnosis of pathogens with various sample types (Miller et al., 2019; Chen et al., 2021; Gu et al., 2021). Nonetheless, without a reference genome, even if the genome of a pathogen was sequenced, the species information cannot be disclosed by mNGS. Therefore, the availability of the *H. shengwangii* genome sequence could provide a valuable source for further comparative genomics analysis in the family Thermicanaceae and facilitate the family's detection rate when conducting an mNGS-based pathogenic detection or study.

According to the conventional taxonomic features, *T. aegyptius*, *H. schlegelii*, *B. lithotrophica*, and *H. shengwangii* shared the identical phenotypes of rod-shaped, Gram-positive, spore-forming, and motor ability, indicating that these were the standard features of family Thermicanaceae. However, a clear description of the Thermicanaceae needs further studies, as few reference species and literature on Thermicanaceae are currently available. In addition, we did not trace the source of $^{\Delta}$ strain DYY3, such as the patient's skin, living environment, and so on, and only one strain of this novel species was isolated so far. Finally, the pathogenic mechanism was not thoroughly investigated, although virulence factor genes and antibiotic resistance genes were predicted in this study.

CONCLUSION

This study identified a novel pathogenic bacterium, *H. shengwangii* gen. nov., sp. nov., isolated from a critically ill patient with CRBSI. In addition to the traditional methods of species identification, we used multiple comparative genomics analyses to prove that $^{\Delta}$ strain DYY3 was genus novel and species novel in the family of Thermicanaceae. Moreover, the constructed high-quality *H. shengwangii* genome will contribute to further genomics research and NGS-based pathogenic detection or study in the family Thermicanaceae.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

SW and YY conceived and designed the project. YD, YL, DS, and SM participated in the strains collection. YD, YL, and SZ performed the morphology physiology and chemotaxonomy identification test. XL, SF, and HX conducted the bioinformatic analysis. YD and XL wrote the manuscript. SW and ZL revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Clinical Research Project of SHDC (No. SHDC2020CR6030-003) and the Three-Year Action Plan for Constructing the Shanghai Public Health System (GWV-3.1).

ACKNOWLEDGMENTS

We thank Yu Kong and Xue Xu (Electron Microscopy Facilities of Center for Excellence in Brain Science and Technology, Chinese Academy of Sciences) for their assistance with EM sample preparation and image analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.919169/full#supplementary-material

Supplementary Figure 1 | Two-dimensional TLC of polar lipids photographs from strain DYY3. Figure **(A–E)** shows the dyeing results using ninhydrin, molybdenum blue 1-methylnaphthol, d reagent, and molybdophosphate. DPG, diphosphatidylglycerol; PG, phosphatidylglycerol; PE, phosphatidylethanolamine; PL1–5, unidentified phospholipids; GL, unidentified glycolipid; L1–2, unidentified lipids; AGPL, unidentified amino-glycerophospholipid.

Supplementary Figure 2 | Whole-cell saccharide hydrolysate experiment. The results indicated that $^{\Delta}$ strain DYY3 did not have any characteristic sugars.

REFERENCES

Barbosa, R. G., van Veelen, H. P. J., Pinheiro, V., Sleutels, T., Verstraete, W., and Boon, N. (2020). Enrichment of hydrogen oxidizing bacteria from high temperature and salinity environments. *Appl. Environ. Microbiol.* 87:e02439-20. doi: 10.1128/AEM.02439-20

Chen, Y., Feng, W., Ye, K., Guo, L., Xia, H., Guan, Y., et al. (2021). Application of metagenomic next-generation sequencing in the diagnosis of pulmonary infectious pathogens from bronchoalveolar lavage samples. *Front. Cell Infect. Microbiol.* 11:541092. doi: 10.3389/fcimb.2021.541092

Gerver, S. M., Mihalkova, M., Bion, J. F., Wilson, A. P. R., Chudasama, D., Johnson, A. P., et al. (2020). Surveillance of bloodstream infections in intensive care units in England, May 2016-April 2017: epidemiology and ecology. *J. Hosp. Infect.* 106, 1–9. doi: 10.1016/j.jhin.2020.05.010

Gossner, A. S., Devereux, R., Ohnemuller, N., Acker, G., Stackebrandt, E., and Drake, H. L. (1999). Thermicanus aegyptius gen. nov., sp. nov., isolated from oxic soil, a fermentative microaerophile that grows commensally with the thermophilic acetogen Moorella thermoacetica. *Appl. Environ. Microbiol.* 65, 5124–5133. doi: 10.1128/AEM.65.11.5124-5133.1999

Gu, W., Deng, X., Lee, M., Sucu, Y. D., Arevalo, S., Stryke, D., et al. (2021). Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nat. Med.* 27, 115–124. doi: 10.1038/s41591-020-1105-z

Hayashi Sant'Anna, F., Bach, E., Porto, R. Z., Guella, F., Hayashi Sant'Anna, E., and Passaglia, L. M. P. (2019). Genomic metrics made easy: what to do and where to go in the new era of bacterial taxonomy. *Crit. Rev. Microbiol.* 45, 182–200. doi: 10.1080/1040841X.2019.1569587

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

Kampfer, P., Glaeser, S. P., and Busse, H. J. (2013). Transfer of Bacillus schlegelii to a novel genus and proposal of Hydrogenibacillus schlegelii gen. nov., comb. nov. *Int. J. Syst. Evol. Microbiol.* 63, 1723–1727. doi: 10.1099/ijs.0.045146-0

Li, N., Cai, Q., Miao, Q., Song, Z., Fang, Y., and Hu, B. (2021). High-Throughput metagenomics for identification of pathogens in the clinical settings. *Small Methods* 5:2000792. doi: 10.1002/smtd.202000792

Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A., and Zdobnov, E. M. (2021). BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* 38, 4647–4654. doi: 10.1093/molbev/msab199

Mermel, L. A., Allon, M., Bouza, E., Craven, D. E., Flynn, P., O'Grady, N. P., et al. (2009). Clinical practice guidelines for the diagnosis and management of intravascular catheter-related infection: 2009 Update by the Infectious Diseases Society of America. *Clin. Infect. Dis.* 49, 1–45. doi: 10.1086/599376

Miller, S., Naccache, S. N., Samayoa, E., Messacar, K., Arevalo, S., Federman, S., et al. (2019). Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res.* 29, 831–842. doi: 10.1101/gr.238170.118

Murray, P. R., and Masur, H. (2012). Current approaches to the diagnosis of bacterial and fungal bloodstream infections in the intensive care unit. *Crit. Care Med.* 40, 3277–3282. doi: 10.1097/CCM.0b013e318270e771

Olm, M. R., Crits-Christoph, A., Diamond, S., Lavy, A., Matheus Carnevali, P. B., and Banfield, J. F. (2020). Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. *mSystems* [Preprint]. doi: 10.1128/msystems.00731-19

Parks, D. H., Chuvochina, M., Chaumeil, P. A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* 38, 1079–1086.

Perevalova, A. A., Kublanov, I. V., Baslerov, R. V., Zhang, G., and Bonch-Osmolovskaya, E. A. (2013). Brockia lithotrophica gen. nov., sp. nov., an anaerobic thermophilic bacterium from a terrestrial hot spring. *Int. J. Syst. Evol. Microbiol.* 63, 479–483. doi: 10.1099/ijs.0.041285-0

Price, T. K., Realegeno, S., Mirasol, R., Tsan, A., Chandrasekaran, S., Garner, O. B., et al. (2021). Validation, Implementation, and Clinical Utility of Whole Genome Sequence-Based Bacterial Identification in the Clinical Microbiology Laboratory. *J. Mol. Diagn.* 23, 1468–1477. doi: 10.1016/j.jmoldx.2021.07.020

Rinke, C., Chuvochina, M., Mussig, A. J., Chaumeil, P. A., Davin, A. A., Waite, D. W., et al. (2021). A standardized archaeal taxonomy for the genome taxonomy database. *Nat. Microbiol.* 6, 946–959. doi: 10.1038/s41564-021-00918-8

Rosenthal, V. D., Duszynska, W., Ider, B. E., Gurskis, V., Al-Ruzzieh, M. A., Myatra, S. N., et al. (2021). International Nosocomial Infection Control Consortium (INICC) report, data summary of 45 countries for 2013-2018, Adult and Pediatric Units, Device-associated Module. *Am. J. Infect. Control* 49, 1267–1274. doi: 10.1016/j.ajic.2021.04.077

Schenk, A., and Aragno, M. (1979). Bacillus schlegelii, a New Species of Thermophilic, Facultatively Chemolithoautotrophic Bacterium Oxidizing Molecular Hydrogen. *J. Gen. Microbiol.* 115, 333–341.

Schwab, F., Geffers, C., Behnke, M., and Gastmeier, P. (2018). ICU mortality following ICU-acquired primary bloodstream infections according to the type of pathogen: A prospective cohort study in 937 Germany ICUs (2006-2015). *PLoS One* 13:e0194210. doi: 10.1371/journal.pone.0194210

Tabak, Y. P., Vankeepuram, L., Ye, G., Jeffers, K., Gupta, V., and Murray, P. R. (2018). Blood Culture Turnaround Time in U.S. Acute Care Hospitals and Implications for Laboratory Process Optimization. *J. Clin. Microbiol.* 56:e00500-18. doi: 10.1128/JCM.00500-18

Wagner, K., Springer, B., Pires, V. P., and Keller, P. M. (2018). Pathogen identification by multiplex lightmix real-time PCR assay in patients with meningitis and culture-negative cerebrospinal fluid specimens. *J. Clin. Microbiol.* 56:e01492-17. doi: 10.1128/JCM.01492-17

Wilson, M. R., Sample, H. A., Zorn, K. C., Arevalo, S., Yu, G., Neuhaus, J., et al. (2019). Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. *N. Engl. J. Med.* 380, 2327–2340. doi: 10.1056/NEJMoa1803396

Zeng, C., Wu, A., Li, L., and Jia, H. (2021). Multi-center prospective study on central line-associated bloodstream infections in 79 ICUs of China. *BMC Infect. Dis.* 21:1208. doi: 10.1186/s12879-021-06871-5

Zhong, Y., Zhou, L., Liu, X., Deng, L., Wu, R., Xia, Z., et al. (2021). Incidence, Risk factors, and attributable mortality of catheter-related bloodstream infections in the intensive care unit after suspected catheters infection: a retrospective 10-year cohort study. *Infect. Dis. Ther.* 10, 985–999. doi: 10.1007/s40121-021-00429-3

Conflict of Interest: XL, SF, HX, and ZL were employed by Hugo Biotechnologies Co., Ltd.

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.

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 19 July 2022
doi: 10.3389/fmicb.2022.886201

Prediction of Smoking Habits From Class-Imbalanced Saliva Microbiome Data Using Data Augmentation and Machine Learning

*Celia Díez López†, Diego Montiel González†, Athina Vidaki and Manfred Kayser**

Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, Rotterdam, Netherlands

Human microbiome research is moving from characterization and association studies to translational applications in medical research, clinical diagnostics, and others. One of these applications is the prediction of human traits, where machine learning (ML) methods are often employed, but face practical challenges. Class imbalance in available microbiome data is one of the major problems, which, if unaccounted for, leads to spurious prediction accuracies and limits the classifier's generalization. Here, we investigated the predictability of smoking habits from class-imbalanced saliva microbiome data by combining data augmentation techniques to account for class imbalance with ML methods for prediction. We collected publicly available saliva 16S rRNA gene sequencing data and smoking habit metadata demonstrating a serious class imbalance problem, i.e., 175 current vs. 1,070 non-current smokers. Three data augmentation techniques (synthetic minority over-sampling technique, adaptive synthetic, and tree-based associative data augmentation) were applied together with seven ML methods: logistic regression, k-nearest neighbors, support vector machine with linear and radial kernels, decision trees, random forest, and extreme gradient boosting. K-fold nested cross-validation was used with the different augmented data types and baseline non-augmented data to validate the prediction outcome. Combining data augmentation with ML generally outperformed baseline methods in our dataset. The final prediction model combined tree-based associative data augmentation and support vector machine with linear kernel, and achieved a classification performance expressed as Matthews correlation coefficient of 0.36 and AUC of 0.81. Our method successfully addresses the problem of class imbalance in microbiome data for reliable prediction of smoking habits.

Keywords: human microbiome, trait prediction, smoking status, prediction modeling, class imbalance, data augmentation, machine learning, saliva microbiome

OPEN ACCESS

Edited by:
Zheng Zhang,
Shandong University, China

Reviewed by:
Hu T. Huang,
Syapse, Inc., United States
Feargal J. Ryan,
South Australian Health and Medical Research Institute (SAHMRI), Australia

***Correspondence:**
Manfred Kayser
m.kayser@erasmusmc.nl

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

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

Received: *28 February 2022*
Accepted: *21 June 2022*
Published: *19 July 2022*

Citation:
Díez López C, Montiel González D, Vidaki A and Kayser M (2022) Prediction of Smoking Habits From Class-Imbalanced Saliva Microbiome Data Using Data Augmentation and Machine Learning. Front. Microbiol. 13:886201. doi: 10.3389/fmicb.2022.886201

INTRODUCTION

In recent years, human microbiome research has elucidated the importance of microbes in the host's wellbeing and their interplay with different phenotypes (Cho and Blaser, 2012; Gilbert et al., 2018). Human microbiome research is currently moving from characterization and association studies toward translational applications. These include diagnosis of metabolic diseases, such as

type 2 diabetes (Duvallet et al., 2017; He et al., 2018; Reitmeier et al., 2020), chronic inflammation disorders (Duvallet et al., 2017; Zhou et al., 2018; Su et al., 2020), and cancer (Duvallet et al., 2017; Poore et al., 2020; Su et al., 2020; Zheng et al., 2020) among others, as well as the prediction of the likely outcomes in personalized interventions, such as therapeutic response (Ananthakrishnan et al., 2017; Zhou et al., 2018) and nutrition (Zeevi et al., 2015; Asnicar et al., 2021). In more specialized applications, such as forensics, novel uses of the human microbiome have been reported to help in reconstructing the crime scene (Díez López et al., 2019, 2020), estimating the post-mortem interval of corpses (PMI) (Belk et al., 2018), or identifying the potential perpetrator(s) of crime (Woerner et al., 2019; Yang et al., 2019a). This current trend is possible due to advances in high-throughput sequencing technologies and bioinformatics analysis methods, together with the large amount of microbiome data that has become available from public repositories. Often, machine learning (ML) methods are preferred for data analysis, with the random forest model standing out as the most often used method so far.

Despite the great promises of ML methods in microbiome research and their application in trait prediction (Reitmeier et al., 2020; Su et al., 2020; Zheng et al., 2020), they also face practical challenges. Heterogeneity in methods, such as nucleic acid isolation or target region of the marker gene, is often encountered in cumulative microbiome datasets and is an obstacle for cross-study applications due to introduced study-specific technical variation (Debelius et al., 2016). Avoiding the pooling of data from different studies can bypass the study-specific effect issue, though greatly reduces the statistical power with negative effects on the reliability of the outcome. Additionally, microbiome data commonly suffer from imbalanced sample distribution (Khan and Kelly, 2020; Poore et al., 2020; Anyaso-Samuel et al., 2021). Particularly in (binary) classification applications, it is commonly the case that one class is overrepresented (majority class) while the other is underrepresented (minority class). This class imbalance leads to spurious high classification accuracy favoring the majority class, while research often focuses on the minority class, and limits the classifier's generalization (Japkowicz and Stephen, 2002; Abd Elrahman and Abraham, 2013; Ali et al., 2013; Thabtah et al., 2020). Some microbiome studies have reported problems in their classifiers due to the class imbalance issue in their datasets. These problems include the different classification performances over different datasets (Wang and Liu, 2020), the inability to perform accurate classifications (Bokulich et al., 2022), or even the classification of every sample to the same class (LaPierre et al., 2019). Therefore, the class imbalance should be considered in the data analysis approach. However, collecting data from more samples is often not viable, and therefore many public datasets come with serious class imbalance problems. Thus, researchers must explore novel methods for solving the class imbalance at the data and/or algorithm level (Japkowicz and Stephen, 2002; Abd Elrahman and Abraham, 2013; Ali et al., 2013).

At the data level, synthetic sampling methods have been suggested for microbiome research (Knights et al., 2011), though studies applying them are scarce. With these methods, to balance the classes, new samples are synthesized *in silico* based on existing minority class samples and added to the training set, an approach referred to as data augmentation. For example, the synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002) is one of the most widely used methods to deal with the class imbalance problem in real-life applications, and has been employed in some microbiome studies (Brooks et al., 2018; Wingfield et al., 2018; Chen et al., 2020; Gomez-Alvarez and Revetta, 2020; Mehta et al., 2020). An alternative is the adaptive synthetic sampling approach for imbalanced learning (ADASYN) (He et al., 2008). More recently, the tree-based associative data augmentation (TADA) method (Sayyari et al., 2019) has been proposed as a microbiome-specific data augmentation method, since it takes into account the phylogenetic relationship between the microbial taxa, but has not been widely applied by the microbiome community as of yet.

In this study, we investigated the predictability of individuals' smoking habits from saliva using publicly available microbiome data that unavoidably are class-imbalanced. Smoking is prevalent in the general population; therefore, smoking prediction from human biological materials, such as saliva, is useful in epidemiology and public health research, can be relevant for medical interventions, and may be of interest to other applied fields, such as forensics. Typically, in epidemiology, public health, and medical studies, smoking habit phenotypes are collected *via* self-reported questionnaires, which, however, are known to be unreliable (Rebagliato, 2002). Alternatively, they are collected *via* laboratory tests, such as cotinine measurements, a metabolite of nicotine, in biological samples like the serum, saliva, or urine. However, cotinine levels are not always available, or collecting them is not always affordable in clinical settings, and smoking classification heavily depends on a suitable threshold. More recently, approaches based on host epigenetics have been introduced *via* the detection of smoking-associated DNA methylation signatures, but issues arise regarding tissue specificity of epigenetic models, as well as model accuracy and suitable laboratory test development, given the large number of epigenetic biomarkers required for accurate predictions (Maas et al., 2019). Hence, microbiome-based prediction of smoking habits from saliva may provide a suitable alternative solution.

Previous studies have established the association between some saliva microbes and the host's tobacco smoking habit (Kato et al., 2016; Takeshita et al., 2016; Wu et al., 2016; Rodriguez-Rabassa et al., 2018; Beghini et al., 2019; Sato et al., 2020). More specifically, these association studies found that the abundance of some bacteria, such as those from the *Proteobacteria* phylum, is decreased in the saliva of smokers, while that of other bacteria, such as from the *Actinobacteria* phylum, is increased. However, at the lower taxonomic levels, there are some discrepancies between studies and study-specific associations. Notably, the largest available studies suffer from class imbalance with a ratio of about 1:5 between the minority class of current smokers and the majority class of non-smokers (Takeshita et al., 2016; Wu et al., 2016). Such class imbalance in the available microbiome data causes a typical and serious problem that needs to be solved by developing and applying suitable data augmentation methods to avoid a negative impact on the final prediction outcome.

In the present study, we deal with class-imbalanced saliva microbiome data for the purpose of predicting individuals' smoking habits. Our strategy consists of (i) optimization and validation of different data augmentation techniques and ML methods using nested cross-validation, (ii) identifying the best-performing approach for predicting smoking habits by taking class imbalance in the underlying microbiome data into account, and (iii) applying the best-performing approach for prediction modeling of human smoking habits from saliva microbiome data despite the underlying class-imbalanced data. The data and the code used are made publicly available.

METHODS

Datasets

Publicly available 16S rRNA gene amplicon sequencing data and associated metadata from two different studies were obtained from the European Bioinformatics Institute (EMBL-EBI). The first study (Wu et al., 2016) (referred to as dataset S1 in this study) included two cohorts: the American Cancer Society (ACS) Cancer Prevention Study II (CPS-II) Nutrition cohort ($N = 543$) (Wu et al., 2016) and the National Cancer Institute (NCI) Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial cohort ($N = 661$) (Wu et al., 2016). The second study (Beghini et al., 2019) (referred to as dataset S2 in this study) included a single cohort from the New York City Health and Nutrition Examination Survey (NYC HANES) ($N = 297$) (Beghini et al., 2019). Dataset S1 comprised 454 pyrosequencing data, whereas dataset S2 comprised Illumina MiSeq data. We discarded samples based on the following criteria: (i) samples lacking metadata information for age, sex, and/or ethnicity, (ii) samples from donors <15 years old based on microbial community differences between youth and adults (Burcham et al., 2020), (iii) duplicate samples from dataset S1 to avoid data redundancy, and (iv) samples from non-smokers with second-hand exposure and "alternative" smokers from the dataset S2. The selected characteristics of the two analyzed datasets are described in **Table 1**. The setup of the experimental studies is described in further detail in **Supplementary Table 1**.

TABLE 1 | Characteristics of the two saliva microbiome datasets used in this study.

	Dataset S1 ($N = 1,088$)	Dataset S2 ($N = 157$)
Smoking status, *N* (%)		
Never smoker	473 (43.5)	39 (24.8)
Former smoker	519 (47.7)	39 (24.8)
Current smoker	96 (8.8)	79 (50.4)
Sex, *N* (%)		
Female	429 (39.4)	88 (56.1)
Male	659 (60.6)	69 (43.9)
Age group, *N* (%)		
20–29	–	20 (12.7)
30–39	–	31 (19.8)
40–49	–	40 (25.5)
50–59	147 (13.5)	29 (18.5)
60–69	505 (46.4)	21 (13.4)
70–79	377 (34.7)	9 (5.7)
80–89	59 (5.4)	6 (3.8)
≥90	–	1 (0.6)
Ethnicity, *N* (%)		
European	1,028 (94.5)	59 (37.6)
Non-European	60 (5.5)	98 (62.4)

Processing of 16S rRNA Gene Amplicon Sequencing Data

The data from the two selected studies were processed separately. Primer sequences were obtained from the original studies and were removed from the raw sequencing reads using cutadapt (v.1.15) (Martin, 2011) by setting the minimum length to >100 bp. The resulting FASTQ files were quality-filtered and de-noised using DADA2 (v.1.12.1) (Callahan et al., 2016). We used recommended parameters that we only modified when needed for our own data. Briefly, in both studies, parameters maxNN and maxEE were set to 0 to avoid unambiguous nucleotides and "expected errors" in the sequencing reads, respectively. Additionally, in dataset S1 (single-end), parameter maxLen was set to 500, and in dataset S2 (paired-end), parameter truncLen was set to 200–150 based on the read quality profiles, making sure to maintain overlap between forward and reverse reads to merge them later. After sample inference of true sequence variants, an amplicon sequence variants (ASV) table was constructed for each study, and chimeric sequences were removed using the command *removeBimeraDenovo()* with default parameters. Subsequently, the naïve Bayesian classifier method was employed for taxonomy assignation using the expanded Human Oral Microbiome Database (eHOMD) (v.15.2) (Escapa et al., 2018) as reference. At this point, only high-coverage samples (>1,000 reads) were kept for downstream analysis, and species with mean relative abundance < 1E−04 across samples were discarded. Taxa counts were normalized using total-sum scaling (TSS) for relative abundance (Paulson et al., 2013) (dataset S1 vs. dataset S2; PERMANOVA Bray-Curtis $R^2 = 0.20$, $q < 0.05$; PERMANOVA Jaccard $R^2 = 0.13$, $q < 0.05$). Moreover, microbiome datasets are normally sparse and characterized by a zero-inflated distribution, where most taxa are not shared among the majority of the samples. This is magnified in cross-study applications with study-specific taxa, which can limit the generalizability of the applications. Based on this finding, we merged the two ASV tables from the two analyzed studies and filtered out study-specific taxa, keeping 124 species from 30 families that were common between the two datasets for downstream analyses (**Supplementary Table 2**) (dataset S1 vs. dataset S2; PERMANOVA Bray-Curtis $R^2 = 0.14$, $q < 0.05$; PERMANOVA Jaccard $R^2 = 0.09$, $q < 0.05$).

Statistical Analyses

The overall differences in the saliva microbial communities between the smoking classes were calculated in QIIME

2 (v.2019.10) (Bolyen et al., 2019): current vs. never vs. former, and current vs. non-current (combined never and former). For this, the weighted UniFrac distance matrix was analyzed by analysis of similarities (ANOSIM) and permutation multivariate analysis of variance (PERMANOVA) where q values ($q < 0.05$ for significance) were obtained with default 999 permutations.

Data Augmentation Techniques

For the prediction of an individual's current smoking habit (smoker vs. non-smoker), we aimed to employ a binary machine learning (ML) classifier. For that, data imbalance was a marked issue in our dataset with a ratio of about 1:6 between the minority class ($N = 175$ smokers) and the majority class ($N = 1{,}070$ non-smokers) (**Table 1**). The problem stems from the ML algorithms that assume an equal number of samples for each class, which would lead to spurious high classification accuracy, favoring the majority class and limiting the classifier's generalization. Therefore, we applied different data augmentation techniques to overcome the data imbalance issue at the data level in our dataset. We used two techniques commonly employed in different fields to handle data imbalance, namely, the synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002) and the adaptive synthetic sampling approach (ADASYN) (He et al., 2008), as well as a recently introduced technique specific for microbiome data named tree-based associative data augmentation (TADA) (Sayyari et al., 2019).

The general approaches to deal with data imbalance are over-sampling (increase the minority class), under-sampling (decrease the majority class), or a combination of the two. Particularly, SMOTE and ADASYN techniques differ in the generation of synthetic samples in the minority class (over-sampling). For that, SMOTE over-sampling pinpoints the samples belonging to the minority class in a Euclidean space, and a random sample is first chosen for which k of its nearest neighbors are found. A line is drawn between the original sample and one randomly chosen neighbor, where a new synthetic sample is generated at a random point along the line (linear combination of samples). The process is repeated generating the same number of synthetic samples for each original minority sample until a specific ratio between the minority and majority classes is reached or to equal the majority class. On the other hand, ADASYN adds random small values to the neighbor samples; hence, they are not linearly correlated to the original sample. By this, ADASYN considers a density distribution between the original sample and its neighborhood, which acts as the criterion to set the number of synthetic samples to be generated from each original sample. On another point, with the under-sampling approach, random majority class samples are dropped out until a specific ratio between the classes is reached. Both SMOTE and ADASYN techniques were implemented using the imbalanced-learn Python toolbox (v.0.6.1) (Lemaître et al., 2017) with default parameters. We employed a combination of over- and under-sampling methods, indicated as SMOTE-1 and ADASYN-1 in this study. In order to set the final ratio between the minority and majority classes, we used the following equation:

$$t = |C_{min} - C_{max}|,\ over - sampling = \frac{|t - C_{min}|}{C_{max}},$$
$$under - sampling = \frac{C_{max} - t}{C_{min}}$$

where C_{min} is the number of the minority class samples, C_{max} is the number of the majority class samples, and t is the absolute value of the difference between C_{min} and C_{max}.

We also used the over-sampling approach alone, indicated as SMOTE-2 and ADASYN-2 in this study, by which the number of the samples in the minority class was equaled to the majority class.

The microbiome-specific TADA technique generates minority class synthetic samples based on a statistical generative model that takes into account the phylogenetic relationships between microbial taxa. We implemented TADA with default parameters, which equals the number of samples in the minority class with the majority class. For the rooted phylogenetic tree required as input, we used the merged ASV table of the two studies to obtain a single consensus sequence for all those sequences assigned to the same taxa at the species level. For that, we used the *ConsensusSequence* function in DECIPHER (v.2.12.0) (Wright, 2016) and subsequently performed multiple sequence alignment of the consensus sequences using MAFFT with auto parameter (v.7.310) (Katoh et al., 2002). A rooted phylogenetic tree was obtained using FastTree (v.2.1.11) (Price et al., 2010) with a generalized time-reversible (GTR) model.

Machine Learning Methods

We evaluated seven different ML methods for binary classification (0: smoker, 1: non-smoker) included in the scikit-learn (v.0.23.2) Python package (Pedregosa et al., 2011): logistic regression (LR), k-nearest neighbors (KNN), support vector machine with linear (SVML) and radial (SVMR) kernels, decision trees (DT), random forest (RF), and extreme gradient boosting (XGBoost). LR is a parametric ML model that assumes a linear dependency between the input features (taxa) and the categorical outcome. The output of the logistic regression linear function is a probability x between 0 and 1, where if $x < 0.5$, the categorical outcome is 0 (smoker), otherwise it is 1 (non-smoker). KNN is a non-parametric model and as such supports non-linear solutions. It finds the Euclidean distances between a query sample and a k number of its closest samples (nearest neighbors) in the feature space and identifies their most frequent class label. SVM models take the data points and find a separating hyperplane between the two classes. SVML is a linear method that looks for linear dependencies among the input features to separate classes. SVMR is a non-linear method that adds an extra dimension to the data (kernel), so they become linearly separable and then project back the decision boundary to the original dimension using the dot product of two vectors in the feature space known as the kernel function. DT is a tree-based ML algorithm that mimics a decision diagram. Each input feature constitutes a node in the tree, where based

upon a certain condition or rule splits into sub-nodes and extends until the leaf node that represents the classification decision (0 for smoker or 1 for non-smoker). Finally, RF and XGBoost are tree-based ensemble models that combine several models to improve their outcome predictions. RF generates a large number of decision trees on different subsamples and combines their outputs using averages at the end of the learning process. On the contrary, XGBoost combines the decision trees during the learning process for which it uses a gradient descent algorithm. By this, the mistakes done in a previous model are learned and improved in the subsequent model until no further improvement can be achieved. Hyperparameter optimization for all the ML models was performed using nested k-fold cross-validation (**Figure 1E**).

Nested K-Fold Cross-Validation

Nested cross-validation (nCV) is a resampling procedure that enables both model optimization and evaluation (Krstajic et al., 2014). The difference between non-nested and nested CV approaches is that the former use the same cross-validation set for hyperparameter optimization and model evaluation, which biases the model to the dataset and leads to optimistically biased classifier's performance; in other words, non-nested approach leads to over-fitting in model selection. The nCV approach overcomes this by evaluating the ML algorithm and the model hyperparameters separately in multiple randomized partitions of the data (Cawley and Talbot, 2010), though it requires more computational time. In nCV, apart from splitting the original training set into *k*-folds of training and test sets (outer folds) (**Figure 1D**), each *k* training fold is at the same time split into *n*-folds of training and validation sets (inner folds) for model hyperparameter tuning (**Figure 1E**). The optimized model is then validated in the corresponding *k* test fold (**Figure 1F**). We employed a 5×2 ($k \times n$) nCV where each of the first splits is named outer-fold (k) and each of the inside splits for hyperparameter tuning as inner-fold (n). Hyperparameter optimization for the seven ML models was performed using the *RandomizedSearchCV()* function in scikit-learn.

Validation of Data Types With Machine Learning Methods

Since most ML algorithms operate trying to maximize the classification accuracy, spurious high classification occurs in imbalanced datasets by correctly classifying all or almost all the samples from the majority class at the cost of misclassifying many samples from the minority class. Hence, performance metrics, such as accuracy, or F1 score alone can lead to misleading results in imbalanced datasets (Chicco and Jurman, 2020). In contrast, the Matthews correlation coefficient (MCC) offers a balanced metric by considering the four confusion matrix categories: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), according to the following equation:

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

For a high MCC score, the classifier has to correctly predict a high percentage of the samples in both the majority and the minority classes, independent of their ratios in the overall dataset, and hence is independent of data imbalance. MCC gives a score ranging from [−1 to +1], where 1 means perfect prediction, 0 random prediction, and −1 perfect inverse prediction. Based on this finding, in order to compare all the possible combinations of input data types, including original non-augmented data (**Figure 1**, $d = 6$) and the ML method (**Figure 1**, $m = 7$), we used the MCC metric. However, we also reported the AUC metric to show potential misleading results in those models with baseline non-augmented data (highly imbalanced). Comparisons among the different data types with a given ML method were performed in the R environment (v.3.6.1) using the Kruskal–Wallis and Wilcoxon tests. Significant *p*-values were determined with a cut-off value of 0.05 following Benjamini-Hochberg (BH) correction.

Approach Setup

The original dataset ($N = 175$ smokers and $N = 1{,}070$ non-smokers) was split into a training set (80%) and a holdout test set (20%) maintaining the sample ratio between the classes (**Figures 1A–C**). Data augmentation techniques were applied to the training split: ADASYN-1 (over- and under-sampling), ADASYN-2 (over-sampling alone), SMOTE-1 (over- and under-sampling), SMOTE-2 (over-sampling alone), and TADA. We evaluated a total of six training data types ($d = 6$), including the original non-augmented and the five augmented data types (**Figure 1B**). Considering each data type separately, we optimized and evaluated seven ML methods (LR, KNN, SVML, SVMR, DT, RF, and XGBoost) ($m = 7$) using an nCV approach as explained before (**Figures 1D–F**). This entire process was repeated 10 times ($i = 10$) (**Figure 1H**), aiming to avoid introduced variation by the original data partitions. The performance metrics (MCC and AUC) resulted from the validation of each optimized model in the five outer test folds (k) over 10 times ($i = 10$) (total of 50 (5 * 10) resulting values for each metric) (**Figure 1G**). The best-performing data type with the ML method was based on the highest resulting MCC value (**Figure 1I**), and the final classifier trained in the final 80% training set (**Figure 1I**) was validated in the final 20% holdout test set (**Figure 1J**).

RESULTS

Saliva Microbiome Data

The data comprised saliva 16S rRNA gene amplicon sequencing data and associated metadata from two different studies referred to here as dataset S1 (Wu et al., 2016) and dataset S2 (Beghini et al., 2019) (see the *Datasets* in the METHODS section for more details). Filtering samples for quality-controlled metadata, de-noising of sequencing reads, and sequencing depth filtering resulted in a total of 1,245 samples ($N = 1{,}088$ from dataset S1 and $N = 157$ from dataset S2). In the whole dataset, class imbalance in smoking habits was large with 512 (44.1%) never smokers, 558 (44.8%) former smokers, and only 175 (14.1%) current smokers. Female samples accounted for 41.5% of the total sample, and the average age (±standard deviation) was 65.2 (±11.0) years. European ancestry of the saliva sample

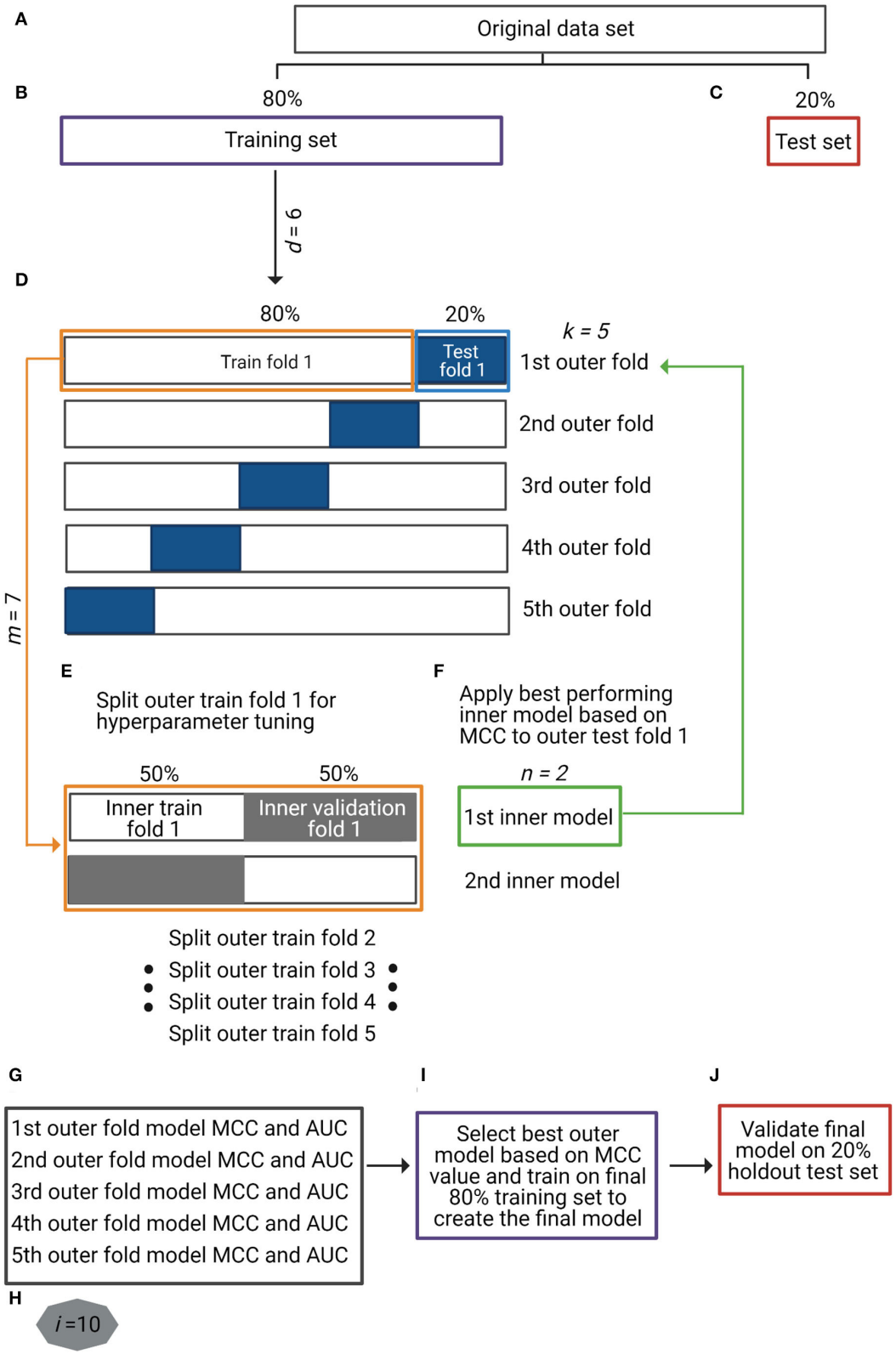

FIGURE 1 | Overview of the study's analytical strategy. **(A–C)** The original dataset was split into a training set (80%) (purple box in **B**) and a holdout test set (20%) (red box in **C**) by maintaining the original ratio between classes in the partitions. Data augmentation techniques were applied to the training set, making a total of six different input data types ($d = 6$), including baseline non-augmented and differently augmented data types. **(D)** For the nested cross-validation (nCV) approach, the

(Continued)

FIGURE 1 | training set was split into five outer *k*-folds of training (80%) (orange box in **D**) and test (20%) (blue box in **D**) sets each. **(E)** Each outer *k*-fold was split into two inner *n*-folds of training (50%) and validation (50%) sets (orange box in **E**) in which seven different machine learning (ML) models ($m = 7$) were optimized and validated (inner models). **(F)** The best-performing *n*-fold inner model (green box in **F**) was applied to the corresponding *k*-fold test set (green arrow to blue box in **F**). **(G)** For each *k*-fold test set, two performance metrics were obtained: Matthews correlation coefficient (MCC) and area under the receiver operating characteristic curve (AUC). Repetition of steps **(D)** to **(G)** for all the input data types ($d = 6$) with ML method ($m = 7$) (total of 42 different approaches). **(H)** Repetition of steps **(A)** to **(G)** 10 times ($i = 10$) to control for introduced variation by data partitions. **(I)** Selection of the best-performing data type with ML method based on MCC metric and training on full final 80% training set to create the final prediction model. **(J)** Validation of final prediction model on final 20% holdout test set.

donors was overrepresented (87.3%), as typically encountered in human microbiome data publicly available thus far. The selected characteristics of the two datasets are described in further detail in **Table 1**.

Microbial taxonomy assignment using the expanded Human Oral Microbiome Database (eHOMD) (v.15.2) (Escapa et al., 2018) as reference (see *Processing of 16S rRNA gene amplicon sequencing data* in the Methods section for more details) and abundance filtering resulted in 200 species from 33 families in dataset S1 and 168 species from 35 families in dataset S2. Both datasets were dominated by a few species that comprised more than 75% of the total microbial composition (21 species in S1 and 15 species in S2). These species belonged to different genera, including *Streptococcus*, *Rothia*, *Haemophilus*, *Prevotella*, *Veillonella,* and *Actinomyces*. Dataset S1 was dominated by *Streptococcus oralis* (0.26 of total relative abundance), followed by *S. salivarius* (0.09), *Rothia mucilaginosa* (0.06), *S. parasanguinis* (0.05), and *Haemophilus parainfluenzae* (0.05). Dataset S2 was also dominated by *S. oralis* (0.24), followed by *S. parasanguinis* (0.06), *S. salivarius* (0.06), *Prevotella melaninogenica* (0.06), and *R. mucilaginosa* (0.06) (**Supplementary Figure 1**). These top abundant species were prevalent in both datasets, appearing in more than 87% of individuals. Our observations are consistent with the reported composition of the saliva microbiome (Segata et al., 2012). For downstream analyses, we selected 124 species from 30 families that were common between the two datasets, to ensure that our proposed strategy was generalizable for the prediction of samples from both datasets (**Supplementary Table 2**). These common species accounted for 86% of the sequencing reads in dataset S1 and 61% in dataset S2.

Classification of Smoking Habits

The overall saliva microbial communities differed with statistical significance between current and never smokers (ANOSIM $R = 0.04$, $q = 0.03$; PERMANOVA pseudo-$F = 11.37$, $q = 0.002$), and current and former smokers (ANOSIM $R = 0.04$, $q = 0.03$; PERMANOVA pseudo-$F = 11.91$, $q = 0.002$), but not between never and former smokers (ANOSIM $R = 0$, $q = 0.51$; PERMANOVA pseudo-$F = 0.64$, $q = 0.63$). Therefore, we grouped the never and former smokers into a single category of non-current smokers, which when compared with the current smokers showed statistically significant differences in the overall microbial communities (ANOSIM $R = 0.04$, $q = 0.02$; PERMANOVA pseudo-$F = 13.26$, $q = 0.001$). Based on these results, we used two classes of non-current and current smokers in all downstream analyses.

Validation of Data Types and Machine Learning Models for Smoking Habit Prediction

A step-by-step overview of our analytical setup can be found in **Figure 1**. For each input data type ($d = 6$), including augmented data and baseline non-augmented data, and each ML model ($m = 7$), the resulting classifiers' performance metrics are expressed as Matthews correlation coefficient (MCC) and area under the receiver operating characteristic curve (AUC), which are summarized in **Figure 2** and **Supplementary Table 3**. Overall, data augmentation techniques combined with ML methods outperformed baseline methods based on the MCC values, except for the KNN method. Briefly, the MCC values resulting from the baseline non-augmented methods increased on average when applying data augmentation techniques with percentages of increase as follows: XGBoost (99.8%), SVMR (92.7%), DT (48.9%), RF (30.6%), and LR (8.8%). The highest increase was observed with SVML where the baseline non-augmented method resulted in random prediction (MCC equal or close to zero), which was highly improved with data augmentation techniques (MCC values 0.31–0.33). Notably, the AUC baseline values did not change so drastically when applying data augmentation techniques [percentage increase or decrease (–)]: XGBoost (15.8%), SVML (8%), SVMR (null increase/decrease), RF (−1.0%), KNN (−4.6%), DT (−6.1%), and LR (−10.4%).

The SVML method performed the best in predicting smoking habits from microbiome data based on the MCC metric. As the reference metric for comparison purposes, we chose the MCC, since it is independent of data imbalance, which is not the case for the AUC metric. MCC values were significantly higher with each of the five augmented data types compared to non-augmented data (Wilcoxon test, BH-adjusted $p = 9.93\text{E}{-20}$) (**Supplementary Table 4**). However, there were no statistically significant differences in the MCC metric between the augmented data pair comparisons (Wilcoxon test, BH-adjusted p-values between $p = 0.392$ and $p = 0.882$) (**Supplementary Table 4**). From these results, we concluded that SVML with augmented data performed better than with imbalanced non-augmented data.

For the training (**Figure 1I**) and the validation in the holdout test set (**Figure 1J**) of the final smoking prediction model, we chose SVML combined with TADA. We based our decision on the following: (i) the SVML method performed the best in predicting smoking habits from microbiome data based on the MCC metric (no statistical difference), and (ii) we selected TADA as the preferred data augmentation technique since it takes into account the phylogenetic relationship between

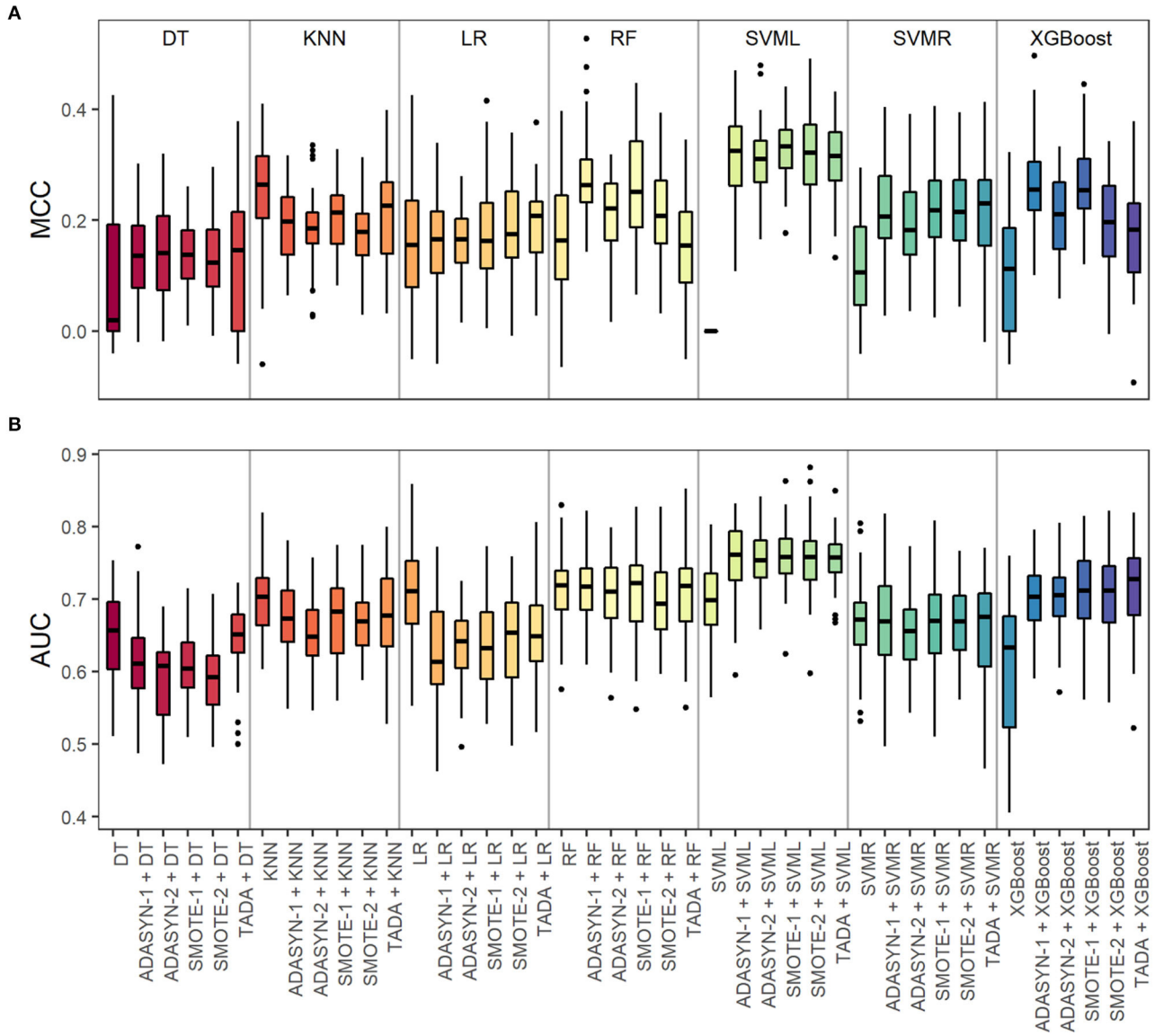

FIGURE 2 | Validation of data types with machine learning (ML) methods for microbiome-based prediction of smoking habits based on the S1 and S2 datasets together. For each ML method, we evaluated six types of input data: baseline non-augmented and five augmented datasets based on different methods (ADASYN-1, ADASYN-2, SMOTE-1, SMOTE-2, and TADA). **(A)** Matthews correlation coefficient (MCC) and **(B)** area under the receiver operating characteristic curve (AUC) values from the 5-fold nested crossed-validation were repeated for 10 times (5 * 10). For MCC, +1 represents a perfect prediction, 0 random prediction, and −1 perfect inverse prediction. For AUC, 1 indicates perfectly accurate prediction and 0.5 indicates random prediction. ML method abbreviations: DT, decision trees; KNN, k-nearest neighbors; LR, logistic regression; RF, random forest; SVML, support vector machine with linear kernel; SVMR, support vector machine with radial kernel; XGBoost, extreme gradient boosting.

the microbial taxa. The average model performance (standard deviation) metrics were MCC of 0.32 (0.07) and AUC of 0.74 (0.05) in the training set, and MCC of 0.36 (0.06) and AUC of 0.81 (0.04) in the holdout test set.

DISCUSSION

In this study, coming with the available 16S rRNA gene amplicon microbiome sequencing data, we deal with the common issue of data imbalance in human microbiome binary classification, with the aim of unlocking the prediction of human host's traits from saliva microbiome. As a data source, we focused on studies targeting the saliva microbiome and did not use data from studies targeting other niches in the oral cavity due to known diverse microbial assemblies on different oral sites (Aas et al., 2005; Zaura et al., 2009; Segata et al., 2012). We selected publicly available saliva microbiome data from two studies that might differ in their experimental setup (**Supplementary Table 1**) but

both have large sample sizes, while discarding other studies with very small sample sizes that could be a source of variation rather than useful information for the prediction. The lack of widespread consensus on microbiome analysis methods, together with the variation introduced at each step of the microbiome pipeline, constitutes hurdles for cross-study applications. This lack can sometimes outweigh the factor(s) of interest and limit the statistical power and generalization of the application (Brooks et al., 2015; Sinha et al., 2015, 2017; Wang and LêCao, 2019). Though we could not control for any potential variation introduced during the experimental analysis, we aimed to apply the same or the most similar bioinformatics analysis to the raw sequencing data to avoid study-specific computational variation, from which quality control choices are amongst the largest sources of variation (Sinha et al., 2015, 2017). Moreover, we only selected the species common between the datasets from the two studies for downstream analyses. On the one hand, we are aware that this might have reduced the power of our prediction analysis by discarding informative species in each of the two datasets separately. On the other hand, this procedure ensured that the approach is suitable for the prediction of samples from both datasets.

Our observations in the overall microbial composition of saliva were in agreement with the two original studies (Wu et al., 2016; Beghini et al., 2019), where microbiome variation did not significantly differ between never and former smokers (ANOSIM $R = 0$, $q = 0.51$; PERMANOVA pseudo-$F = 0.64$, $q = 0.63$), but significantly differed between never and current (ANOSIM $R = 0.04$, $q = 0.03$; PERMANOVA pseudo-$F = 11.37$, $q = 0.002$), and between former and current smokers (ANOSIM $R = 0.04$, $q = 0.03$; PERMANOVA pseudo-$F = 11.91$, $q = 0.002$). One of the two studies (Wu et al., 2016) also reported significant differences between current smokers and non-current smokers (combined never and former) as we observed in this study (ANOSIM $R = 0.04$, $q = 0.02$; PERMANOVA pseudo-$F = 13.26$, $q = 0.001$).

The available saliva microbiome dataset presents the problem of data imbalance, which is commonly encountered in microbiome datasets and in many other real-life applications, with a ratio of about 1:6 between the minority class of current smokers and the majority class of non-smokers (**Table 1**). Using class-imbalanced data in prediction modeling can lead to spurious high accuracy based on the correct classification of most of the samples from the majority class at the cost of misclassifying many or even most of the samples from the minority class (Japkowicz and Stephen, 2002; Abd Elrahman and Abraham, 2013; Ali et al., 2013; Thabtah et al., 2020). Regarding our study purpose, this would translate in the classifier's inability to correctly predict the positive observations for current smoking habits (minority class). This was seen in the baseline non-augmented data with the SVML method (**Figure 2** and **Supplementary Table 3**), where we obtained a low MCC of zero but a medium AUC of 0.7. Besides needing to address the class imbalance, this also highlights the necessity of not relying only on a single prediction accuracy score for model validation when dealing with imbalanced data (Chicco and Jurman, 2020).

The MCC performance metric allowed us for fair comparisons of the validated ML methods for both non-augmented and augmented data, since MCC is independent of data imbalance (Boughorbel et al., 2017; Ballabio et al., 2018). For the great majority of the ML methods, augmented data resulted in higher MCC scores compared to imbalanced non-augmented data, thus facilitating improved classification performance. This demonstrates that microbiome-based classification problems can benefit from data augmentation techniques, in line with previous suggestions (Knights et al., 2011). In our dataset, the combined over- and under-sampling approaches generally performed slightly better (though not statistically significantly) than the over-sampling approach alone (**Supplementary Tables 2, 3**).

The variation in the performance metric values for each input data type and ML method (**Figure 2**) highlights the variation introduced in the optimization and validation procedures (**Figure 1**). This underlines the necessity for an nCV approach for overall model validation and selection that is independent of the different data partitions (Cawley and Talbot, 2010; Krstajic et al., 2014). We avoided over-fitting in model validation and classifier selection as demonstrated by the very similar performance metrics between the final training (MCC: 0.32 ± 0.07, AUC: 0.74 ± 0.05) and test (MCC: 0.36 ± 0.06, AUC: 0.81 ± 0.04) datasets, which were very similar to those of the folds in the nCV (MCC: 0.31 ± 0.06, AUC: 0.75 ± 0.05). As it has been suggested before (Topçuoglu et al., 2020), with our strategy, we report the variation in the predictive performance on the different folds of nCV, as well as on both the final training and test sets, which unfortunately is not a very common practice in microbiome-based trait prediction.

With the best data augmentation and ML approach chosen, we predicted individuals' smoking habits from saliva 16S rRNA gene microbiome data in the final holdout test set with MCC of 0.36 and AUC of 0.81. Previously, Sato et al. (2020) predicted smoking habits from class-imbalanced tongue metagenomics data ($N = 234$ never, $N = 52$ current smokers) using an RF approach and conventional non-nested k-fold CV and obtained an AUC of 0.75 from the test set. This prediction was improved to AUC = 0.80–0.93 when using single-nucleotide variants of single species as input data instead of relative abundances of all species. More recently, Carrieri et al. (2021) predicted smoking habits from leg skin 16S rRNA gene amplicon sequencing data based on a less class-imbalanced but very small dataset ($N = 43$ never, $N = 19$ current smokers) using the XGboost method and conventional non-nested k-fold CV and reported the F1 performance metric in the CV folds ($F1 = 0.72 \pm 0.12$), training set ($F1 = 0.98$), and test set ($F1 = 0.85$). The noted differences in the F1 scores might be an indication of introduced variation by the different data partitions and bias toward model selection, which can be overcome using an nCV approach as proposed by us and others. Notably, the methods applied in both of these previous studies did not take the class imbalance problem in the used data into account. Therefore, and because of the small sample size in one of these studies at least, the previously reported prediction accuracies are not expected to be reliable, in contrast to the results from our study.

However, in our dataset, we acknowledge some metadata-related characteristics that might limit the prediction of microbiome-based smoking habits, even when the data imbalance issue was accounted for by our approach. Precise

phenotype descriptions were available in only one of the two studies (**Supplementary Table 1**), which is a commonly encountered problem in cross-study applications and can limit the interpretation of results (Huttenhower et al., 2014). Also, the dataset is overrepresented by the age range of 50–79 years old and European ancestry of the sample donors (**Table 1**), which might result in different prediction performances in other age groups (Lira-Junior et al., 2018; Liu et al., 2020) and ethnicities (Mason et al., 2013; Yang et al., 2019b). To add, one limitation of the data augmentation techniques is that synthesized metadata associated with the synthetically produced data is not reliable. This limits the possibility of statistically adjusting for covariates (i.e., age, sex, and ethnicity) in the ML methods, which can ultimately improve the prediction performance. Hence, the ideal scenario would be to start from a sample that is a good representation of the general population, though this is challenging in real-life applications.

To conclude, by testing different data augmentation techniques and ML methods on class-imbalanced microbiome data, we established a best-practice approach for reliable prediction of individuals' smoking habits from the saliva microbiome that takes the underlying data imbalance into account. We found that combining data augmentation with ML generally outperformed baseline methods in our dataset for our purpose, as other researchers have also suggested before (Knights et al., 2011). The prediction accuracies, expressed as MCC of 0.36 and AUC of 0.81, we achieved for our best model in the final test set implies that predicting human smoking habits from microbiome data needs further improvement before it can be considered for practical applications.

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. The source code and data used for the analyses performed in this study can be accessed at https://github.com/dmontielg/smokingmicrobiome.

AUTHOR CONTRIBUTIONS

CD and AV conceptualized the idea. CD, DM, and AV designed the study with contributions by MK. DM implemented and performed the data analysis. CD contributed to initial data collection and curation, performed some of the statistical analyses, and prepared the display items. MK provided resources. CD, AV, and MK wrote the manuscript with contributions by DM. All authors approved the final manuscript.

FUNDING

The work of all authors was supported by Erasmus MC University Medical Center Rotterdam. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ACKNOWLEDGMENTS

We thank the colleagues for making their microbiome and smoking data publicly available. We also thank Liese Boonstra for contributing to initial data collection and exploration as part of her bachelor's education at Leiden University of Applied Sciences carried out during her internship at Erasmus MC.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.886201/full#supplementary-material

REFERENCES

Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I., and Dewhirst, F. E. (2005). Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* 43, 5721–5732. doi: 10.1128/jcm.43.11.5721-5732.2005

Abd Elrahman, S. M., and Abraham, A. (2013). A review of class imbalance problem. *J. Netw.* 1, 332–340.

Ali, A., Shamsuddin, S. M., and Ralescu, A. L. (2013). Classification with class imbalance problem. A review. *Int. J. Advance Soft. Compu. Appl.* 7, 176–204.

Ananthakrishnan, A. N., Luo, C., Yajnik, V., Khalili, H., Garber, J. J., Stevens, B. W., et al. (2017). Gut microbiome function predicts response to anti-integrin biologic therapy in inflammatory bowel diseases. *Cell Host Microbe* 21 603–610.e603. doi: 10.1016/j.chom.2017.04.010

Anyaso-Samuel, S., Sachdeva, A., Guha, S., and Datta, S. (2021). Metagenomic geolocation prediction using an adaptive ensemble classifier. *Front. Genet.* 12, 642282. doi: 10.3389/fgene.2021.642282

Asnicar, F., Berry, S. E., Valdes, A. M., Nguyen, L. H., Piccinno, G., Drew, D. A., et al. (2021). Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* 27, 321–332. doi: 10.1038/s41591-020-01183-8

Ballabio, D., Grisoni, F., and Todeschini, R. (2018). Multivariate comparison of classification performance measures. *Chemom. Intell. Lab. Syst.* 174, 33–44. doi: 10.1016/j.chemolab.2017.12.004

Beghini, F., Renson, A., Zolnik, C., Geistlinger, L., Usyk, M., Moody, T. U., et al. (2019). Tobacco exposure associated with oral microbiota oxygen utilization in the New York City Health and Nutrition Examination Study. *Ann. Epidemiol.* 34, 18–25. doi: 10.1101/470286

Belk, A., Xu, Z. Z., Carter, D. O., Lynne, A., Bucheli, S., Knight, R., et al. (2018). Microbiome data accurately predicts the postmortem interval using random forest regression models. *Genes* 9, 104. doi: 10.3390/genes9020104

Bokulich, N. A., Łaniewski, P., Adamov, A., Chase, D. M., Caporaso, J. G., and Herbst-Kralovetz, M. M. (2022). Multi-omics data integration reveals metabolome as the top predictor of the cervicovaginal microenvironment. *PLoS Comput. Biol.* 18, e1009876. doi: 10.1371/journal.pcbi.1009876

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

Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE* 12, e0177678. doi: 10.1371/journal.pone.0177678

Brooks, A. W., Priya, S., Blekhman, R., and Bordenstein, S. R. (2018). Gut microbiota diversity across ethnicities in the United States. *PLoS Biol.* 16, e2006842. doi: 10.1371/journal.pbio.2006842

Brooks, J. P., Edwards, D. J., Harwich, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G., et al. (2015). The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol.* 15, 66. doi: 10.1186/s12866-015-0351-6

Burcham, Z. M., Garneau, N. L., Comstock, S. S., Tucker, R. M., Knight, R., Metcalf, J. L., et al. (2020). Patterns of oral microbiota diversity in

adults and children: a crowdsourced population study. *Sci. Rep.* 10, 2133. doi: 10.1038/s41598-020-59016-0

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

Carrieri, A. P., Haiminen, N., Maudsley-Barton, S., Gardiner, L.-J., Murphy, B., Mayes, A. E., et al. (2021). Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. *Sci. Rep.* 11, 1–18. doi: 10.1038/s41598-021-83922-6

Cawley, G. C., and Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* 11, 2079–2107.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357.

Chen, L., Li, Z., Zeng, T., Zhang, Y.-H., Liu, D., Li, H., et al. (2020). Identifying robust microbiota signatures and interpretable rules to distinguish cancer subtypes. *Front. Mol. Biosci.* 7, 604794. doi: 10.3389/fmolb.2020.604794

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* 21, 6. doi: 10.1186/s12864-019-6413-7

Cho, I., and Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* 13, 260–270. doi: 10.1038/nrg3182

Debelius, J., Song, S. J., Vazquez-Baeza, Y., Xu, Z. Z., Gonzalez, A., and Knight, R. (2016). Tiny microbes, enormous impacts: what matters in gut microbiome studies? *Genome Biol.* 17, 217–217. doi: 10.1186/s13059-016-1086-x

Díez López, C., Montiel González, D., Haas, C., Vidaki, A., and Kayser, M. (2020). Microbiome-based body site of origin classification of forensically relevant blood traces. *Forensic Sci. Int. Genet.* 47, 102280. doi: 10.1016/j.fsigen.2020.102280

Díez López, C., Vidaki, A., Ralf, A., Montiel González, D., Radjabzadeh, D., Kraaij, R., et al. (2019). Novel taxonomy-independent deep learning microbiome approach allows for accurate classification of different forensically relevant human epithelial materials. *Forensic Sci. Int. Genet.* 41, 72–82. doi: 10.1016/j.fsigen.2019.03.015

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* 8, 1784. doi: 10.1038/s41467-017-01973-8

Escapa, I. F., Chen, T., Huang, Y., Gajare, P., Dewhirst, F. E., and Lemon, K. P. (2018). New insights into human nostril microbiome from the expanded Human Oral Microbiome Database (eHOMD): a resource for the microbiome of the human aerodigestive tract. *mSystems* 3, e00187–e00118. doi: 10.1128/mSystems.00187-18

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. *Nat. Med.* 24, 392–400. doi: 10.1038/nm.4517

Gomez-Alvarez, V., and Revetta, R. P. (2020). Monitoring of nitrification in chloraminated drinking water distribution systems with microbiome bioindicators using supervised machine learning. *Front. Microbiol.* 11, 571009. doi: 10.3389/fmicb.2020.571009

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). "ADASYN: adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328.

He, Y., Wu, W., Zheng, H.-M., Li, P., McDonald, D., Sheng, H.-F., et al. (2018). Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* 24, 1532–1535. doi: 10.1038/s41591-018-0219-z

Huttenhower, C., Knight, R., Brown, C. T., Caporaso, J. G., Clemente, J. C., Gevers, D., et al. (2014). Advancing the microbiome research community. *Cell* 159, 227–230. doi: 10.1016/j.cell.2014.09.022

Japkowicz, N., and Stephen, S. (2002). The class imbalance problem: a systematic study. *Intel. Data Anal.* 6, 429–449. doi: 10.1016/j.neunet.2018.07.011

Kato, I., Vasquez, A. A., Moyerbrailean, G., Land, S., Sun, J., Lin, H.-S., et al. (2016). Oral microbiome and history of smoking and colorectal cancer. *J. Epidemiol. Res.* 2, 92–101. doi: 10.5430/jer.v2n2p92

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Khan, S., and Kelly, L. (2020). Multiclass disease classification from microbial whole-community metagenomes. *Pac. Symp. Ciocomput.* 25, 55–66.

Knights, D., Costello, E. K., and Knight, R. (2011). Supervised classification of human microbiota. *FEMS Microbiol. Rev.* 35, 343–359. doi: 10.1111/j.1574-6976.2010.00251.x

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.* 6, 10. doi: 10.1186/1758-2946-6-10

LaPierre, N., Ju, C. J. T., 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

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* 18, 559–563.

Lira-Junior, R., Åkerman, S., Klinge, B., Boström, E. A., and Gustafsson, A. (2018). Salivary microbial profiles in relation to age, periodontal, and systemic diseases. *PLoS ONE* 13, e0189374. doi: 10.1371/journal.pone.0189374

Liu, S., Wang, Y., Zhao, L., Sun, X., and Feng, Q. (2020). Microbiome succession with increasing age in three oral sites. *Aging* 12, 7874. doi: 10.18632/aging.103108

Maas, S. C. E., Vidaki, A., Wilson, R., Teumer, A., Liu, F., van Meurs, J. B. J., et al. (2019). Validated inference of smoking habits from blood with a finite DNA methylation marker set. *Eur. J. Epidemiol.* 34, 1055–1074. doi: 10.1007/s10654-019-00555-w

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* 17, 3. doi: 10.14806/ej.17.1.200

Mason, M. R., Nagaraja, H. N., Camerlengo, T., Joshi, V., and Kumar, P. S. (2013). Deep sequencing identifies ethnicity-specific bacterial signatures in the oral microbiome. *PLoS ONE* 8, e77287. doi: 10.1371/journal.pone.0077287

Mehta, S. D., Zhao, D., Green, S. J., Agingu, W., Otieno, F., Bhaumik, R., et al. (2020). The microbiome composition of a man's penis predicts incident bacterial vaginosis in his female sex partner with high accuracy. *Front. Cell. Infect. Microbiol.* 10, 433. doi: 10.3389/fcimb.2020.00433

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10, 1200–1202. doi: 10.1038/nmeth.2658

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Poore, G. D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., et al. (2020). Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* 579, 567–574. doi: 10.1038/s41586-020-2095-1

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5, e9490. doi: 10.1371/journal.pone.0009490

Rebagliato, M. (2002). Validation of self reported smoking. *J. Epidemiol. Community Health* 56, 163–164. doi: 10.1136/jech.56.3.163

Reitmeier, S., Kiessling, S., Clavel, T., List, M., Almeida, E. L., Ghosh, T. S., et al. (2020). Arrhythmic gut microbiome signatures predict risk of type 2 diabetes. *Cell Host Microbe* 28, 258–272.e256. doi: 10.1016/j.chom.2020.06.004

Rodriguez-Rabassa, M., Lopez, P., Rodriguez-Santiago, R. E., Cases, A., Felici, M., Sanchez, R., et al. (2018). Cigarette smoking modulation of saliva microbial composition and cytokine levels. *Int. J. Environ. Res. Public Health* 15, 2479. doi: 10.3390/ijerph15112479

Sato, N., Kakuta, M., Hasegawa, T., Yamaguchi, R., Uchino, E., Kobayashi, W., et al. (2020). Metagenomic analysis of bacterial species in tongue microbiome of current and never smokers. *NPJ Biofilms Microbiomes* 6, 11. doi: 10.1038/s41522-020-0121-6

Sayyari, E., Kawas, B., and Mirarab, S. (2019). TADA: phylogenetic augmentation of microbiome samples enhances phenotype classification. *Bioinformatics* 35, i31–i40. doi: 10.1093/bioinformatics/btz394

Segata, N., Haake, S. K., Mannon, P., Lemon, K. P., Waldron, L., Gevers, D., et al. (2012). Composition of the adult digestive tract bacterial microbiome based on

seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* 13, R42. doi: 10.1186/gb-2012-13-6-r42

Sinha, R., Abnet, C. C., White, O., Knight, R., and Huttenhower, C. (2015). The microbiome quality control project: baseline study design and future directions. *Genome Biol.* 16, 276. doi: 10.1186/s13059-015-0841-8

Sinha, R., Abu-Ali, G., Vogtmann, E., Fodor, A. A., Ren, B., Amir, A., et al. (2017). Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nat. Biotechnol.* 35, 1077–1086. doi: 10.1038/nbt.3981

Su, X., Jing, G., Sun, Z., Liu, L., Xu, Z., McDonald, D., et al. (2020). Multiple-disease detection and classification across cohorts via microbiome search. *mSystems* 5, e00150-20. doi: 10.1128/mSystems.00150-20

Takeshita, T., Kageyama, S., Furuta, M., Tsuboi, H., Takeuchi, K., Shibata, Y., et al. (2016). Bacterial diversity in saliva and oral health-related conditions: the Hisayama Study. *Sci. Rep.* 6, 22164. doi: 10.1038/srep22164

Thabtah, F., Hammoud, S., Kamalov, F., and Gonsalves, A. (2020). Data imbalance in classification: experimental evaluation. *Inf. Sci.* 513, 429–441. doi: 10.1016/j.ins.2019.11.004

Topçuoglu, B. D., Lesniak, N. A., Ruffin, M. T., Wiens, J., and Schloss, P. D. (2020). A framework for effective application of machine learning to microbiome-based classification problems. *mBio* 11, e00434-20. doi: 10.1128/mBio.00434-20

Wang, X.-W., and Liu, Y.-Y. (2020). Comparative study of classifiers for human microbiome data. *Med. Microecol.* 4, 100013. doi: 10.1016/j.medmic.2020.100013

Wang, Y., and LêCao, K.-A. (2019). Managing batch effects in microbiome data. *Brief Bioinform.* 21, 1954–1970. doi: 10.1093/bib/bbz105

Wingfield, B., Coleman, S., McGinnity, T. M., and Bjourson, A. J. (2018). Robust microbial markers for non-invasive inflammatory bowel disease identification. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 2078–2088. doi: 10.1109/TCBB.2018.2831212

Woerner, A. E., Novroski, N. M. M., Wendt, F. R., Ambers, A., Wiley, R., Schmedes, S. E., et al. (2019). Forensic human identification with targeted microbiome markers using nearest neighbor classification. *Forensic Sci. Int. Genet.* 38, 130–139. doi: 10.1016/j.fsigen.2018.10.003

Wright, E. S. (2016). Using DECIPHER v2. 0 to analyze big biological sequence data in R. *R J.* 8, 352–359.

Wu, J., Peters, B. A., Dominianni, C., Zhang, Y. L., Pei, Z. H., Yang, L. Y., et al. (2016). Data from: Cigarette smoking and the oral microbiome in a large study of American adults. European Nucleotide Archive. *ISME J.* 10, 2435–2446. doi: 10.1038/ismej.2016.37

Yang, J., Tsukimi, T., Yoshikawa, M., Suzuki, K., Takeda, T., Tomita, M., et al. (2019a). *Cutibacterium acnes* (Propionibacterium acnes) 16S rRNA genotyping of microbial samples from possessions contributes to owner identification. *mSystems* 4, e00594-19. doi: 10.1128/mSystems.00594-19

Yang, Y., Zheng, W., Cai, Q., Shrubsole, M. J., Pei, Z., Brucker, R., et al. (2019b). Racial differences in the oral microbiome: data from low-income populations of African ancestry and European ancestry. *mSystems* 4, e00639-19. doi: 10.1128/mSystems.00639-19

Zaura, E., Keijser, B. J. F., Huse, S. M., and Crielaard, W. (2009). Defining the healthy "core microbiome" of oral microbial communities. *BMC Microbiol.* 9, 259. doi: 10.1186/1471-2180-9-259

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., et al. (2015). Personalized nutrition by prediction of glycemic responses. *Cell* 163, 1079–1094. doi: 10.1016/j.cell.2015.11.001

Zheng, Y., Fang, Z., Xue, Y., Zhang, J., Zhu, J., Gao, R., et al. (2020). Specific gut microbiome signature predicts the early-stage lung cancer. *Gut Microbes* 11, 1030–1042. doi: 10.1080/19490976.2020.1737487

Zhou, Y., Xu, Z. Z., He, Y., Yang, Y., Liu, L., Lin, Q., et al. (2018). Gut microbiota offers universal biomarkers across ethnicity in inflammatory bowel disease diagnosis and infliximab response prediction. *mSystems* 3, e00188-17. doi: 10.1128/mSystems.00188-17

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

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

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 26 July 2022
DOI 10.3389/fmicb.2022.933189

OPEN ACCESS

EDITED BY
Zheng Zhang,
Shandong University, China

REVIEWED BY
Wenjie Fang,
Shanghai Changzheng Hospital, China
Sam Ma,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Jingjing Xia
xiajingjing@fudan.edu.cn
Jiucun Wang
jcwang@fudan.edu.cn
Jean Krutmann
Jean.Krutmann@iuf-duesseldorf.de

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

RECEIVED 30 April 2022
ACCEPTED 04 July 2022
PUBLISHED 26 July 2022

CITATION
Wei Q, Li Z, Gu Z, Liu X, Krutmann J, Wang J and Xia J (2022) Shotgun metagenomic sequencing reveals skin microbial variability from different facial sites.
Front. Microbiol. 13:933189.
doi: 10.3389/fmicb.2022.933189

Shotgun metagenomic sequencing reveals skin microbial variability from different facial sites

Qingzhen Wei[1,2], Zhiming Li[3], Zhenglong Gu[4], Xiao Liu[5], Jean Krutmann[6]*, Jiucun Wang[1,7,8]* and Jingjing Xia[4,6]*

[1]Human Phenome Institute, School of Life Sciences, Fudan University, Shanghai, China, [2]Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, China, [3]BGI-Shenzhen, Shenzhen, China, [4]Greater Bay Area Institute of Precision Medicine (Guangzhou), School of Life Sciences, Fudan University, Guangzhou, China, [5]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, [6]IUF-Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany, [7]Department of Dermatology, Huashan Hospital, Fudan University, Shanghai, China, [8]Research Unit of Dissecting the Population Genetics and Developing New Technologies for Treatment and Prevention of Skin Phenotypes and Dermatological Diseases (2019RU058), Chinese Academy of Medical Sciences, Beijing, China

Biogeography (body site) is known to be one of the main factors influencing the composition of the skin microbial community. However, site-associated microbial variability at a fine-scale level was not well-characterized since there was a lack of high-resolution recognition of facial microbiota across kingdoms by shotgun metagenomic sequencing. To investigate the explicit microbial variance in the human face, 822 shotgun metagenomic sequencing data from Han Chinese recently published by our group, in combination with 97 North American samples from NIH Human Microbiome Project (HMP), were reassessed. Metagenomic profiling of bacteria, fungi, and bacteriophages, as well as enriched function modules from three facial sites (forehead, cheek, and the back of the nose), was analyzed. The results revealed that skin microbial features were more alike in the forehead and cheek while varied from the back of the nose in terms of taxonomy and functionality. Analysis based on biogeographic theories suggested that neutral drift with niche selection from the host could possibly give rise to the variations. Of note, the abundance of porphyrin-producing species, i.e., *Cutibacterium acnes*, *Cutibacterium avidum*, *Cutibacterium granulosum*, and *Cutibacterium namnetense*, was all the highest in the back of the nose compared with the forehead/cheek, which was consistent with the highest porphyrin level on the nose in our population. Sequentially, the site-associated microbiome variance was confirmed in American populations; however, it was not entirely consistent. Furthermore, our data revealed correlation patterns between *Propionibacterium acnes* bacteriophages with genus *Cutibacterium* at different facial sites in both populations; however, *C. acnes* exhibited a distinct correlation with *P. acnes* bacteriophages in Americans/Chinese. Taken together, in this study, we

explored the fine-scale facial site-associated changes in the skin microbiome and provided insight into the ecological processes underlying facial microbial variations.

KEYWORDS

shotgun metagenomic sequencing, facial skin microbiome, Chinese, *Cutibacterium acnes* (*C. acnes*), *Propionibacterium acnes* bacteriophage, fine-scale, biogeography

Introduction

The human skin is considered a complex ecosystem colonized with various microorganisms, including bacteria, fungi, and viruses, collectively termed "skin microbiota" (Grice and Segre, 2011). Balanced microbial community composition is essential for maintaining skin health (Byrd et al., 2018). However, this ecosystem turned out to be highly variable between individuals (Schommer and Gallo, 2013) and the factors responsible for the unique variability included endogenous host factors (host genetics, gender, and age) and exogenous environmental factors (lifestyle, hygiene routine, cosmetics, climate, and seasonality) (Grice and Segre, 2011; Boxberger et al., 2021).

Biogeography (body site) has been suggested as a major factor influencing the composition of the skin microbial community (Grice et al., 2009; Perez Perez et al., 2016; Wang et al., 2021). Characterization of spatiotemporal patterns in species distribution is a key task in biogeography and is also fundamental to explore the ecological and evolutionary processes shaping communities (Bahram et al., 2015). For skin microbiome, many studies favored to divide skin into four microenvironments (i.e., sebaceous, moist, dry, and foot) according to the physical and chemical properties of the anatomical sites (Oh et al., 2014). Although this classification was not delicate enough, some prominent features of microbial distribution pattern were well-characterized, for example, genus *Cutibacterium* and *Malassezia* favored oily (sebaceous) areas; genus *Staphylococcus* and *Corynebacterium* were predominant in moist areas while Gram-negative microorganisms favored dry areas (Grice et al., 2009; Chen and Tsao, 2013; Oh et al., 2014, 2016). However, microbial variance from anatomic sites at a more fine-scale level, for example, different sites from one's face, was only partially understood (Lee et al., 2021). This is not trivial. Many facial conditions, exerting substantial adverse psychological and social influences, exhibited a clear and consistent site preference on the face, such as acne vulgaris and seborrheic dermatitis, prone to occur in oily areas with a rich supply of sebaceous glands (Williams et al., 2012; Tan and Bhate, 2015; Sparber et al., 2019), and rosacea often occurs in the central face such as the nose (Van Zuuren et al., 2011; Yigider et al., 2016). Therefore, it is valuable to learn about the microbial variance caused by this delicate body location, which may underlie the predisposition of skin dysbiosis conditions with site preference (Flowers and Grice, 2020).

Due to low microbial biomass from the skin (Chen et al., 2018), most studies deployed 16S rRNA sequencing and assessed only the bacterial community, leaving the fungal and viral communities largely unknown, particularly in the facial sites. To address this issue, we leveraged our shotgun metagenomic sequencing dataset generated from 822 Chinese samples (Li et al., 2021) and reassessed the data intensively, which allowed for more precise recognition of facial skin microbiota (forehead, cheek, and the back of the nose) across all kingdoms (bacteria, fungi, and viruses), in terms of microbial taxonomy and functionality. Sequentially, we reassessed 97 North-American metagenomic sequencing data from the Human Microbiome Project (HMP) (Oh et al., 2014) and compared the main features of the two populations. In particular, a series of *Propionibacterium acnes* bacteriophages, viral members which were considered important in regulating the balance of the microbiome, were assessed and highlighted.

Materials and methods

Study population

Ninety-seven North American samples from HMP (Oh et al., 2014) and 822 Han Chinese samples (Li et al., 2021) were selected. The datasets were downloaded from the integrated Human Skin Microbial Gene Catalog (iHSMGC). Detailed information about sampling, DNA preparation, and shotgun metagenomic sequencing can be obtained according to our previous study (Li et al., 2021).

Statistical analysis

The Shannon index was used to represent the alpha diversity of the microbiome. Kruskal-Wallis test and Wilcoxon rank-sum test were used to assess the significance of the difference in three anatomical sites. Probability (P) values < 0.05 were considered to indicate statistically significant differences. P-values were adjusted using the false discovery rate (FDR) correction.

Beta diversity (principal coordinate analysis (PCoA) based on Bray-Curtis distances) was to characterize the microbial profile in different sites. The permutational multivariate analysis of variance (PERMANOVA) was used to assess the effect of different anatomical sites. We performed the analysis using the

method implemented in the R package (vegan) and 1,000 times permutations to obtain the permuted *P*-value.

Linear discriminant analysis (LDA) effect size (LEfSe) was used to identify taxonomic differences between different anatomical sites. The threshold on the LDA score was set to 3.0.

Spearman correlation was carried out to investigate the existence of a correlation between *P. acnes* bacteriophages and four species that belong to the genus *Cutibacterium*, and the significance levels are $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. *P*-values were adjusted using the FDR correction.

The LEfSe was completed using the Wekemo Bioincloud[1]. Another analysis was conducted using R (version 4.1.2).

Neutral community model analysis was used to explore ecological processes underlying microbial variations. Bray-Curtis distance of each site (FH forehead, CK cheek, NS nose) from the center was assessed using a classic model inferring genetic distance in molecular evolution (Li, 1997). Specifically, the distance to FH = [Distance (FH-NS) + Distance (FH-CK) – Distance (CK-NS)]/2; the distance to CK = [Distance (CK-NS) + Distance (FH-CK) – Distance (FH-NS)]/2; and the distance to NS = [Distance (FH-NS) + Distance (CK-NS) – Distance (FH-CK)]/2).

$$d_{FH} = \frac{D_{NS-FH} + D_{CK-FH} - D_{CK-NS}}{2}$$

$$d_{CK} = \frac{D_{CK-NS} + D_{FH-CK} - D_{FH-NS}}{2}$$

$$d_{NS} = \frac{D_{FH-NS} + D_{CK-NS} - D_{FH-CK}}{2}$$

Results

The back of the nose exhibited distinct microbial community composition from the forehead and cheek in the Chinese

We first investigated skin microbiome in three facial sites (forehead, cheek, and the back of the nose) from our population, in terms of the alpha diversity, microbial composition, and potential functionality.

The overall alpha diversity, indicated by the Shannon index, was higher in the forehead and the cheek than in the nose, while the difference was not significant between the forehead and the cheek (**Figure 1A**). Furthermore, the Shannon index of each kingdom (bacteria, fungi, and viruses) from the three sites was also assessed. The results demonstrated that the back of the nose presented different microbial diversities from the other two sites, in regard to all kingdoms. However, in contrast to lower diversity in the bacterial community, the nose exhibited higher diversity in the fungal and viral community than that of the forehead and the cheek (**Figure 1B**). PCoA based on Bray-Curtis distance also confirmed a shift of nose microbiome from the other two sites, while the microbiome from the forehead and the cheek was more similar (PERMANOVA test, $R^2 = 0.04$, $P < 0.001$) (**Figure 1C**).

To specify the differential species, LEfSe analysis was further carried out. The result revealed 17 site-associated dominant species across kingdoms in different facial sites: *Cutibacterium acnes*, *Cutibacterium granulosum*, *Staphylococcus epidermidis,* and *Propionibacterium* phage PHL132N00 were more abundant in the back of the nose; *Mycobacterium sp. QIA-37*, *Ralstonia solanacearum, Mycobacteroides chelonae, Propionibacterium sp. oral taxon 193, Malassezia globosa, Komagataella phaffii*, and *Acinetobacter junii* were more enriched in the forehead; and *Moraxella osloensis*, *Streptococcus pneumoniae*, *Acinetobacter guillouiae*, *Streptococcus oralis*, *Neisseria sicca,* and *Acinetobacter haemolyticus* were more abundant in the cheek (**Figure 1D**). The relative abundance of these 17 differential species varied significantly, especially between the nose and the other two sites (**Figure 1D**). Of note, the nose harbored clear higher amount of *C. acnes* and lower amount of *M. osloensis*. These two species were proven to be distinctive in nutrient demand: whereas *C. acnes* was high nutrient demanding and prone to the sebum-rich area, and *M. osloensis* was a non-fastidious bacterium that was able to grow in a mineral medium supplemented with a single organic carbon source (Juni, 1974, 2015). Correlation analysis further confirms this negative association between the two species. In addition, we found that a series of site-differential species were internally positive-correlated, whereas mostly negatively correlated with other site-prone species (**Figure 1E**). To further explore the possible mechanisms shaping the microbial biogeography, we conducted an analysis based on a neutral community model (Li, 1997), which is commonly applied to predict the assembly pattern of the communities and is favorable for the relative simplicity. By measuring the Bray-Curtis distance from the center of three sites, we found that the nose is much further from the center than the other two sites, whereas the distances for the other two are only marginally different (**Figure 1F**). A strict neutral drift would predict similar distances among all three lineages, indicating that selective forces (e.g., host selection) may exist in shaping the microbial variability, especially in the nose area.

Shotgun metagenomic sequencing revealed that certain functionality underlies the site-associated microbiome variance in the Chinese population

As shotgun metagenomic sequencing provided gene abundance information, we further assessed the functionality potentials of the microbiota located in these three anatomical

1 https://www.bioincloud.tech

10.3389/fmicb.2022.933189

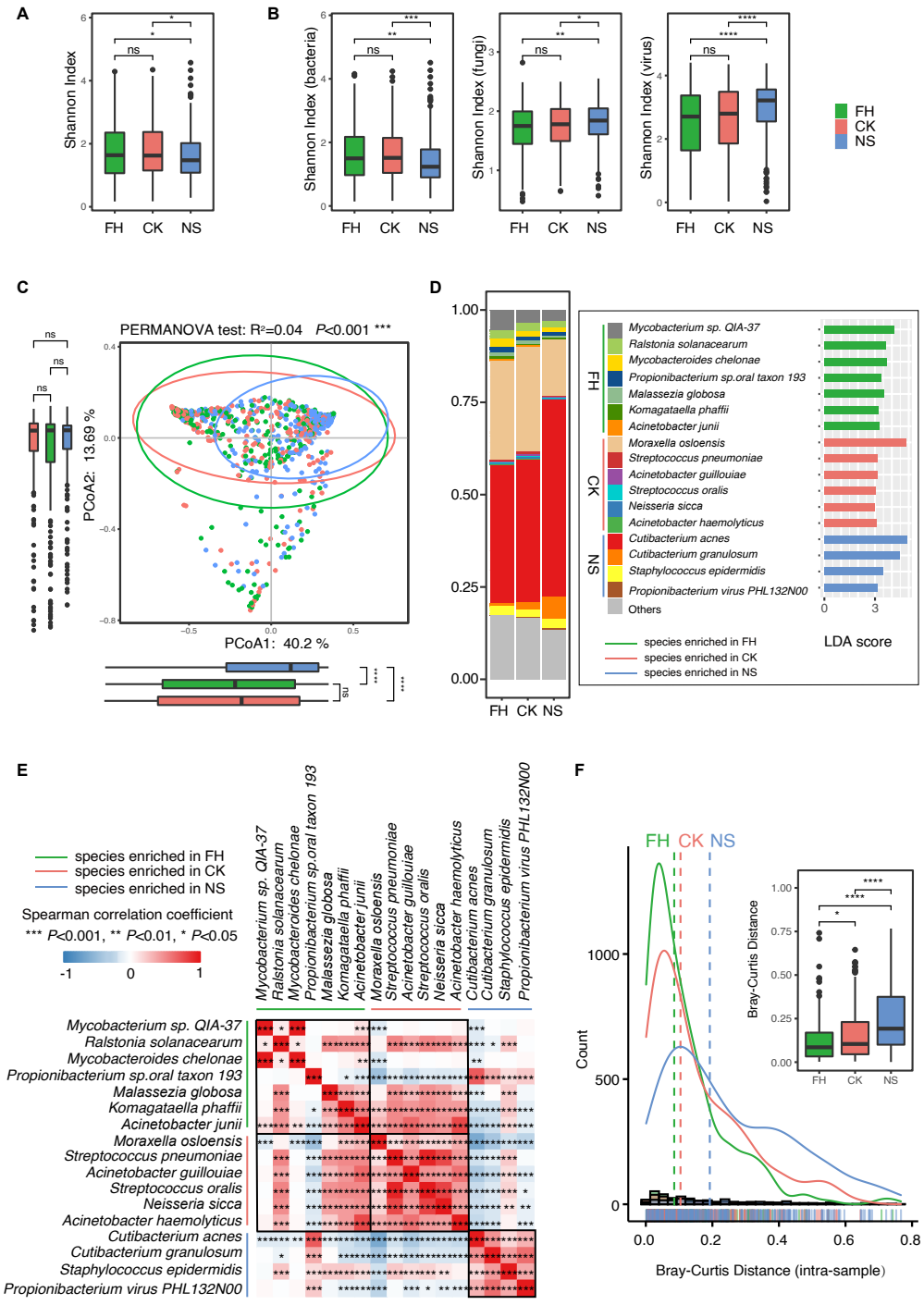

FIGURE 1

Comparative characterization of the skin microbiome present in three anatomical sites in Chinese samples. Boxplots comparing Shannon index of **(A)** overall skin microbiome and **(B)** bacteria, fungi, and virus microbiome of three anatomical sites in Chinese samples. **(C)** Principal coordinate analysis (PCoA) plot illustrating the comparison of the overall composition of skin microbiome between three sites in the Chinese population. The PERMANOVA test is used to determine significance. Boxplots indicate the distribution of samples along the PC1 and PC2. **(D)** Stack plot of the 17 differential species ranked by relative abundance. Linear discriminant analysis (LDA) effect size (LEfSe) histograms on the right showed the microbial comparisons of three anatomical sites, with an LDA threshold of 3.0. **(E)** Heat map of the Spearman's correlation between differential species. **(F)** The distance distribution lines on the left showed the Bray-Curtis distance from the center of three facial sites to the forehead, to the cheek, and to the back of the nose, respectively. The boxplot on the right quantified and compared the difference. FH, the forehead; CK, the cheek; NS, the back of the nose. *P*-values were adjusted using the false discovery rate (FDR) correction. The significance levels are: ns, not significant, $P > 0.05$; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

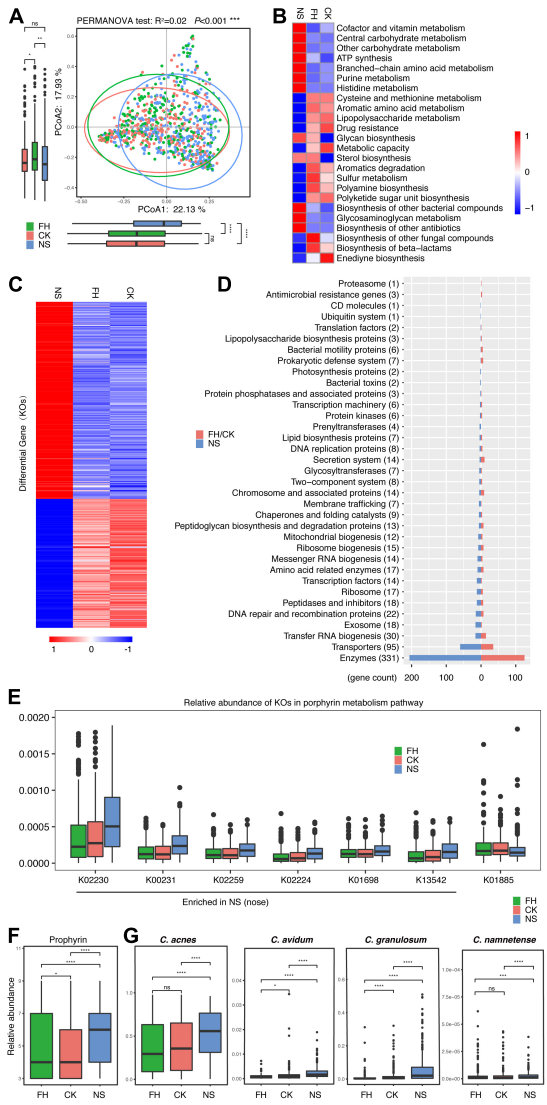

FIGURE 2
Comparative characterization of the skin microbiome present in three anatomical sites in American samples. **(A)** PCoA plot illustrating the comparison of the overall composition of skin microbiome between three sites in the Chinese population. The PERMANOVA test is used to determine significance. **(B)** Functional differences of the skin microbiome in three sites of Chinese samples. The heat map represents the relative abundance of function in three sites. **(C)** The heat map represents the relative abundance of 554 differential genes in three sites in Chinese samples. **(D)** Different categories that 554 differential genes belong to. **(E)** Boxplots comparing seven differential genes related to the porphyrin metabolism pathway in three anatomical sites (K02230: cobN, cobaltochelatase CobN; K00231: PPOX, protoporphyrinogen/coproporphyrinogen III oxidase; K02259: COX15, heme a synthase; K02224: cobB-cbiA; cobyrinic acid a,c-diamide synthase; K01698: hemB, porphobilinogen synthase; K13542: cobA-hemD, uroporphyrinogen III methyltransferase/synthase; K01885: EARS, glutamyl-tRNA synthetase). The relative abundance of facial porphyrin **(F)** and *Cutibacterium acnes*, *Cutibacterium avidum*, *Cutibacterium granulosum*, and *Cutibacterium namnetense* **(G)** in three sites. FH, the forehead; CK, the cheek; NS, the back of the nose. The significance levels are: ns, not significant, $P > 0.05$; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

sites. Overall, PCoA confirmed the variance in terms of gene features at the three sites (PERMANOVA test, $R^2 = 0.02$, $P < 0.001$) (Figure 2A). The PC1 indicator of the PCoA showed a minor but significant difference between the back of the nose and the forehead/cheek. Furthermore, the heat map showed the relative abundance of 24 functional modules (Kyoto Encyclopedia of Genes and Genomes (KEGG) level C) enriched in the forehead and cheek while different from the back of the nose (Kruskal Wallis test, P-adjust < 0.05) (Figure 2B and Supplementary Table 1). Specifically, seven microbial functions of high gene abundance were all enriched in the nose, i.e., cofactor and vitamin metabolism, central carbohydrate metabolism, other carbohydrate metabolism, ATP synthesis, branched-chain amino acid metabolism, purine metabolism, and histidine metabolism. Other functions, many of which also related to metabolism, were more enriched in the forehead/cheek, such as serine and threonine metabolism, aromatic amino acid metabolism, lipopolysaccharide metabolism, and drug resistance.

More intensively, we identified 641 differential genes (out of 863 genes with relative abundance > 0.1%) (Kruskal Wallis test, P-adjust < 0.05). Notably, 554 of them (86.4%, Supplementary Table 2) showed a clear difference between the back of the nose and the forehead/cheek (Figure 2C), including 331 enzymes, 95 transporters, and other genes (Figure 2D and Supplementary Table 3). While 219 genes were more enriched in the forehead/cheek, 335 genes were more enriched in the nose. Interestingly, we found that there were seven differential genes, essential for the porphyrin metabolism, and six genes were more enriched in the back of the nose (Figure 2E). In fact, we observed that porphyrin levels, assessed with VISIA-CR pictures (Canfield Scientific Inc., Fairfield, NJ, USA), were the highest in the back of nose compared with the other two sites in our cohort (Figure 2F). Furthermore, it is known that several skin commensals were able to produce porphyrin, and while predominant from *C. acnes* (Shu et al., 2013; Spittaels et al., 2021), other *Propionibacterium* strains, such as *C. granulosum*, *Cutibacterium avidum*, and *Cutibacterium modestum* (previously, "*Propionibacterium humerusii*") were also able to produce certain levels of porphyrin (Barnard et al., 2020). In consistent, our data revealed that the relative abundance of these porphyrin-producing species, i.e., *C. acnes*, *C. avidum*, *C. granulosum*, and *Cutibacterium namnetense*, were all the highest in the back of the nose compared with the forehead/cheek (Figure 2G).

Facial site-associated microbiome variation is different between the Chinese and North American populations

Sequentially, we assessed the site-associated microbiome variance in the North American population. Overall, the

10.3389/fmicb.2022.933189

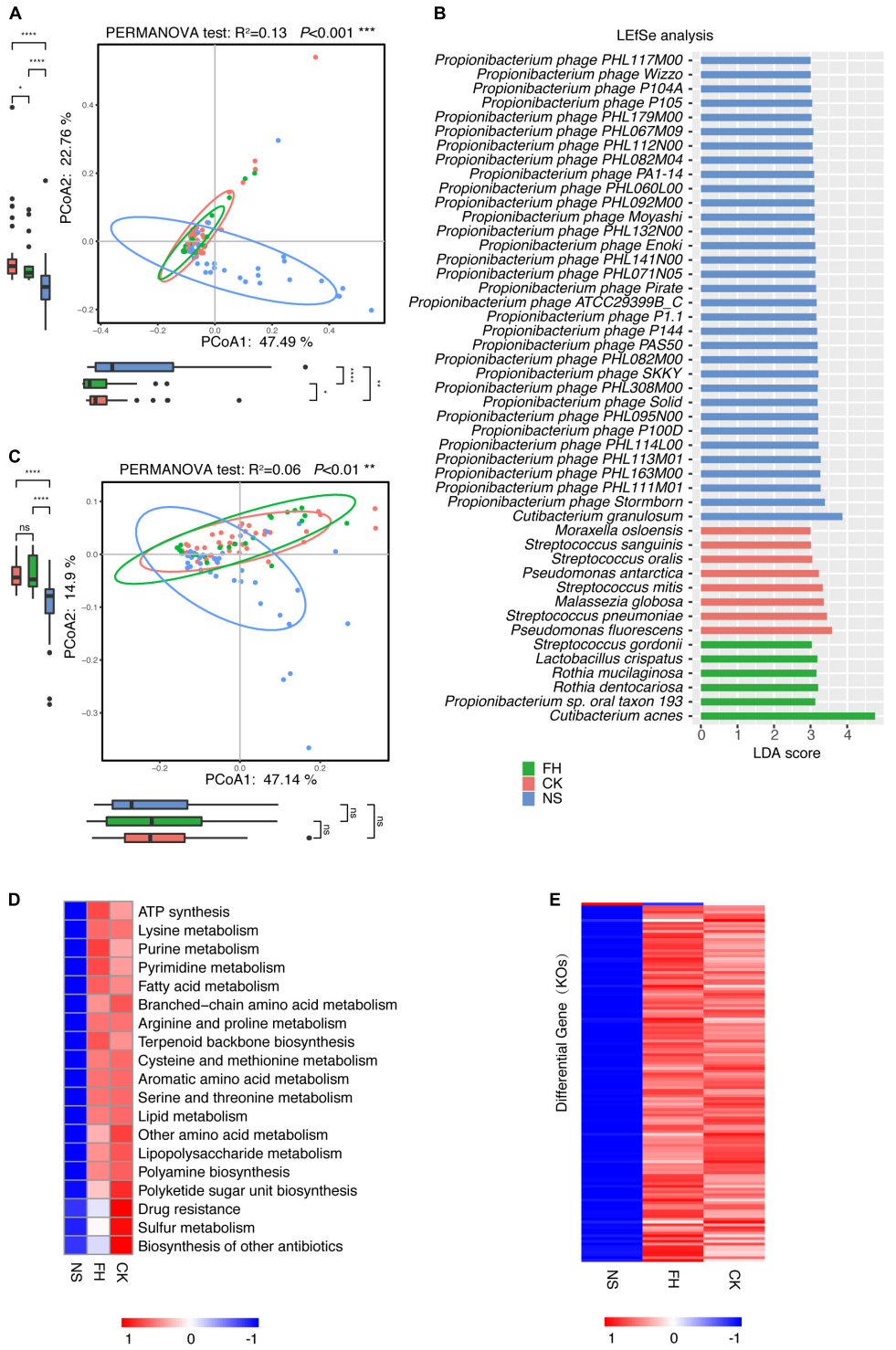

FIGURE 3
Comparative characterization of the skin microbiome present in three anatomical sites in American samples. **(A)** PCoA plot illustrating the comparison of the overall composition of skin microbiome between three sites in the American population. The PERMANOVA test is used to determine significance. Boxplots indicate the distribution of samples along the PC1 and PC2. **(B)** LEfSe histograms for the microbial comparisons of three anatomical sites, with an LDA threshold of 3.0. **(C)** PCoA plot illustrating the gene composition of skin microbiome between three sites in the American population. The PERMANOVA test is used to determine significance. **(D)** Functional differences of the skin microbiome in three sites of American samples. The heat map represents the relative abundance of function in three sites. **(E)** The heat map represents the relative abundance of 131 differential genes in three sites. FH, the forehead; CK, the cheek; NS, the back of the nose. *P*-values were adjusted using the FDR correction. The significance levels are: ns, not significant, $P > 0.05$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

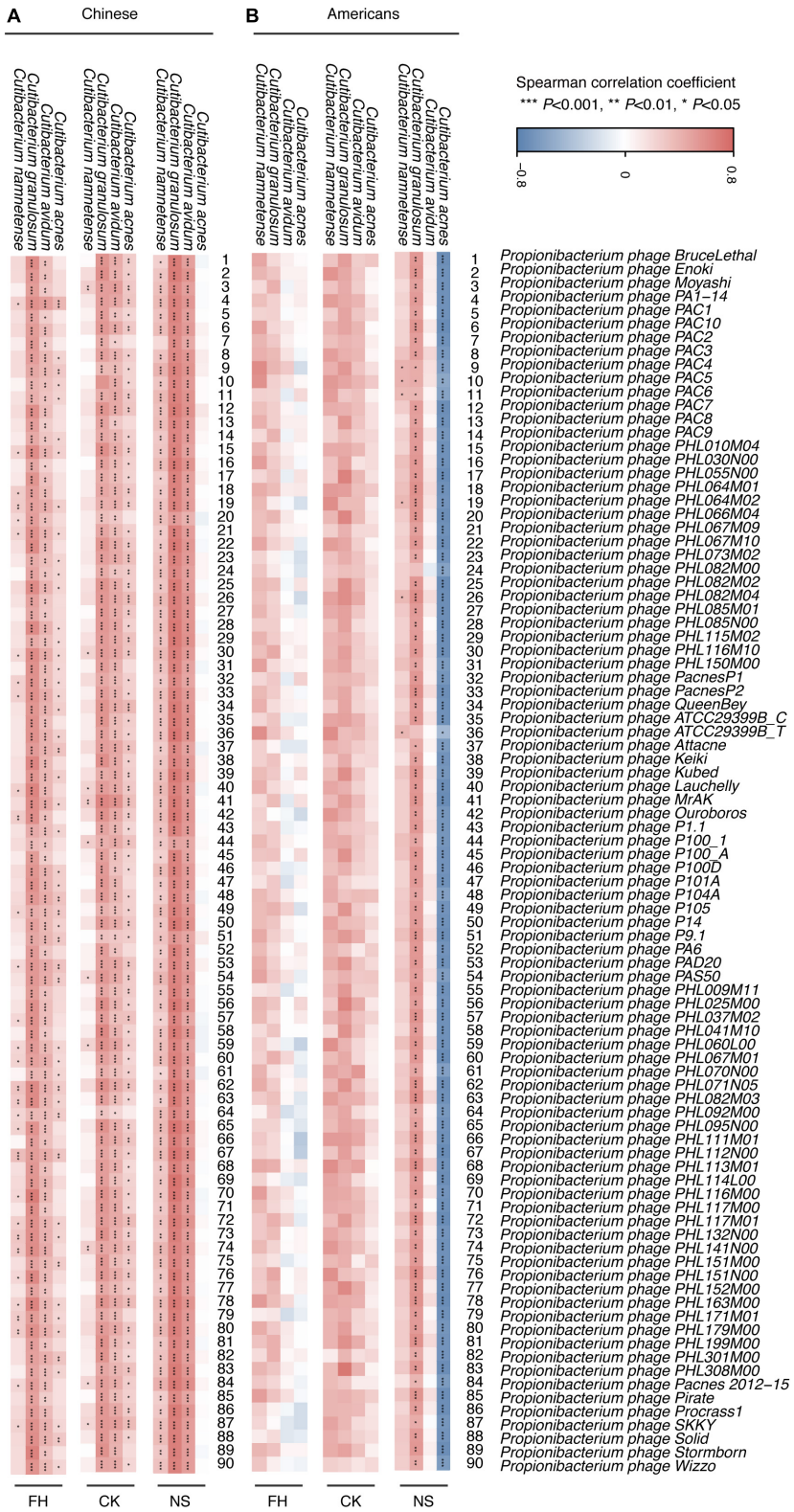

FIGURE 4
The correlation between *Cutibacterium* and *Propionibacterium acnes* bacteriophages in Chinese and American samples. Heat map of the results of the Spearman correlation between four species in genus *Cutibacterium* and *P. acnes* bacteriophages in three sites and in Chinese **(A)** and American **(B)** samples. FH, the forehead; CK, the cheek; NS, the back of the nose. *P*-values were adjusted using the FDR correction. The significance levels in the Spearman correlation are: $^{*}P < 0.05$; $^{**}P < 0.01$; $^{****}P < 0.001$.

PCoA suggested a microbiome variance existed among three facial sites (PERMANOVA test, $R^2 = 0.13$, $P < 0.001$) and the back of the nose was different from the forehead/cheek (**Figure 3A**), consistent with the conclusion drawn from the Chinese population. Furthermore, LEfSe analysis revealed site-associated dominant species across the kingdoms.

In consistent, *M. osloensis*, *S. pneumoniae*, *S. oralis*, *Propionibacterium sp. oral taxon 193*, and *M. globosa* were more enriched in the forehead/cheek, whereas *C. granulosum* and a large series of *P. acnes* bacteriophages were enriched in the back of the nose in both populations (**Figure 3B**). Of note, *C. acnes* was more enriched in the forehead in Americans, which contrasted with the highest abundance in the back of the nose in Chinese. *S. epidermidis* showed more enrichment in the back of the nose in Chinese, but no site difference in Americans.

Based on the gene abundance, the PCoA also showed that the forehead and the cheek were much more similar but both different from the back of the nose (PERMANOVA test, $R^2 = 0.06$, $P < 0.001$) (**Figure 3C**). Of note, 19 microbial functions (KEGG level C) were found to be significantly different in three sites (Kruskal Wallis test, P-adjust < 0.05) (**Figure 3D** and **Supplementary Table 4**). The heat map showed the relative abundance of differential functions from the three facial sites. Specifically, several microbial functions, such as cysteine and methionine metabolism, aromatic amino acid metabolism, lipopolysaccharide metabolism, drug resistance, sulfur metabolism, polyamine biosynthesis, and polyketide sugar unit biosynthesis were also higher in the forehead/cheek in the Chinese samples.

In Americans, there were 145 site-associated differential genes (Kruskal Wallis test, P-adjust < 0.05), and 131 of them (90.3%) showed similar abundance between the forehead and cheek but significantly different from the back of the nose. Among these 131 genes, only K17316 (glucose/mannose transport system permease protein) was more enriched in the back of the nose, while the rests were more enriched in the forehead/cheek (**Figure 3E** and **Supplementary Table 5**).

A distinct correlation between *Propionibacterium acnes* bacteriophages and *Cutibacterium acnes* was observed in the two populations

Propionibacterium acnes bacteriophages, members of the viral community, are dominant bacteriophages that existed in the skin microbiota, especially in the pilosebaceous unit (Liu et al., 2015). These bacteriophages were able to play an important role in maintaining the balance of the microbial community (Liu et al., 2015). However, the association between these bacteriophages with other skin microbiota was rarely studied.

In this study, we assessed the correlation between *P. acnes* bacteriophages and all detectable species from the genus *Cutibacterium* (Dekio et al., 2021), in three sites of two populations. In general, the Chinese showed more correlations in three sites compared with the Americans (**Figure 4**). For Chinese populations, the forehead and the cheek presented mostly consistent positive correlation between genus *Cutibacterium,* particular *C. acnes*, *C. granulosum*, and *C. avidum* with most detected *P. acnes* bacteriophages; while in the back of the nose, *C. acnes* and *P. acnes* bacteriophages showed no significant correlation, but the correlations between *C. namnetense* and bacteriophages were significant (**Figure 4A**). In contrast, most correlations in the forehead/cheek showed similar trends but not significant in Americans. In particular, *P. acnes* bacteriophages exhibited a consistent positive correlation with *C. granulosum*, but a significant negative correlation with *C. acnes* in the back of the nose in Americans (**Figure 4B**).

Discussion

In this study, we centered on addressing site-associated microbiome variance in Chinese facial skin, by intensively reassessing our shotgun metagenomic dataset generated from 822 Chinese samples (Li et al., 2021). Overall, our data revealed that microbial features in the back of the nose were distinctive from the forehead and cheek in Chinese. Furthermore, we confirmed a similar site-associated microbial pattern in the North American population, although varied in detail.

It is long known that biogeography (body site) is a major factor influencing the composition of the skin microbial community (Grice et al., 2009; Perez Perez et al., 2016; Wang et al., 2021). However, there was very limited understanding of the mechanisms shaping microbial biogeography as it is often rather difficult to determine the relative importance of drift, dispersal, speciation, and selection, the four processes (mechanisms) determining the patterns of biogeography and community dynamics (Ma, 2021). Nevertheless, there were several studies worked on the relative significance of stochastic neutral forces and deterministic niche selection and brought us new insights into the mechanisms, shaping the biogeography of the human microbiome (Ma et al., 2018; Tong et al., 2019; Ma, 2021). For example, an analysis of a multi-site microbiome, covering five major habitats (i.e., airway, oral, gut, skin, and urogenital) suggested the relative significance of stochastic neutral forces and deterministic niche selection in shaping the biogeography of the human microbiome (Ma et al., 2018). Another study also suggested that while skin mycobiome assembly is a predominantly neutral process, taxa that could be under the influence of selective forces (e.g., host selection) are potentially key to the structure of a community network (Tong et al., 2019). In this study, we observed a similar pattern that

fit to a "nutrient-drive" model by explaining the site-associated microbial disparity.

In contrast, addressing site-associated microbial variation at a fine-scale level is important for digging the niche selection pressure for the skin microbiome. Multiple variables, such as hygiene routine, cosmetics, climate, and seasonality, which were known to impact microbial niche conditions (Grice and Segre, 2011; Harris-Tryon and Grice, 2022), were well-controlled in this adjacent subsite area and thereby substantially facilitate decoding the microbial variation. Learned from classical ecology, the selection pressures for the ecosystem include resource availability (presence of nutrients), environmental conditions (temperature, geographical access), and biological factors (predators and pathogens) (Williams, 1996). In this study, we revealed a series of site-prone species, many of which were previously proven to be distinct in nutrient requirements, i.e., *C. acnes as* high nutrient-demand, and *M. osloensis* as low nutrient-demand species able to grow in a mineral medium supplemented with a single organic carbon source (Juni, 1974; Juni and Bøvre, 2015). Furthermore, we revealed that site-associated species correlated with each in pattern, suggesting that specific interactions between species underlie the formation of networks to compete in the niche occupation. In turn, the colonization of microbiota in different sites may also reflect niche conditions. In consistent, *C. acnes* tends to colonize in oily areas, and the abundance increases with the sebum level (Mukherjee et al., 2016). A study in Korean women revealed higher sebum secretion in the nose than in the forehead and cheek (Youn et al., 2005), consistent with the higher abundance of *C. acnes* in the nose than in the forehead/cheek in our study. However, the relative abundance of *C. acnes* was demonstrated the highest in the forehead in Americans, which may be due to the ethnical differences in regard to delicate anatomic structures, such as the count and size of sebaceous glands and physiological phenotypes (Rawlings, 2006; Voegeli et al., 2019).

In addition, our data revealed site-associated microbial features not only in taxonomical composition but also in functionality. In Chinese, the carbohydrate metabolism of microbiota was more enriched in the back of the nose, which is consistent with the fact that *C. acnes* utilized carbohydrates as the main carbon source (Li et al., 2021). In contrast, *M. osloensis* was incapable of utilizing any carbohydrates or possessing any saccharolytic activity but strictly depend on other carbon sources such as acetic or lactic acid (Baumann et al., 1968; Juni, 1974; Moss et al., 1988; Juni and Bøvre, 2015).

In our study, *M. osloensis* was the most abundant differential species in the cheek, which exhibited the lowest hydration level (Lee et al., 2013; Machkova et al., 2018) as well as the sebum level (Youn et al., 2005). In both populations, some function modules were more enriched in the forehead/cheek compared with the back of the nose, including sulfur metabolism, cysteine and methionine metabolism, aromatic amino acid metabolism, polyketide sugar unit biosynthesis, and drug resistance, which may imply a more challenged or competitive environment for microbes to adapt in the forehead/cheek than the nose. For example, sulfur is an essential nutrient and can be metabolized into the sulfur-containing amino acids (cysteine and methionine) in microorganisms, protecting against oxidative and environmental stresses such as dryness (Ernst, 1998; Yi et al., 2010; Chan et al., 2019).

In this study, some microbial composition-associated skin feature was validated, i.e., the enrichment of genes in porphyrin metabolism in the nose was demonstrated to link to the abundance of porphyrin-producing species, which were further proven to be positively associated with high porphyrin level on the nose. It is known that bacterial porphyrins are considered to be pro-inflammatory and linked to inflammatory skin diseases (Schaller et al., 2005). Our findings may underlie this site preference for specific inflammatory skin conditions, such as acne vulgaris or rosacea.

Furthermore, one of the highlights of this study was that we were able to explore the composition of other communities, in addition to bacteria, in these facial sites. Bacteriophages, viruses that infect corresponding host bacteria, may play an important regulatory role in human skin health (Liu et al., 2015). However, the interaction between bacteriophage with other skin microbiota is rarely known. In this study, we found that *C. granulosum* and various *P. acnes* bacteriophages were enriched in the nose in both populations. Furthermore, there was an intriguing correlation pattern between *P. acnes* bacteriophages with genus *Cutibacterium* at different facial sites in both populations. Of note, *C. acnes* demonstrated a distinct correlation with *P. acnes* bacteriophages in American/Chinese. It is known that the distribution of *P. acnes* bacteriophages depends on their specific host species (Jonczyk-Matysiak et al., 2017) and recent studies revealed the complexity of different lineages of *C. acnes* on the skin (Dekio et al., 2021; Conwill et al., 2022). These all implied that the significance of more deep sequencing was needed in the future to address complicated correlations.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee, School of Life Sciences,

Fudan University, China. The patients/participants provided their written informed consent to participate in this study.

Author contributions

QZW and ZML: formal analysis and visualization. QZW and JJX: writing – original draft preparation. QZW, JJX, ZLG, and JK: writing – review and editing. JJX, JCW, ZLG, JK, and XL: scientific supervision. JJX and JCW: funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

This study was supported by National Natural Science Foundation of China (81703097), the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), the CAMS Innovation Fund for Medical Sciences (2019-I2M-5-066), the 111 Project (B13016), and a startup grant from the Greater Bay Area Institute of Precision Medicine (Guangzhou), Fudan University to JJX.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.933189/full#supplementary-material

References

Bahram, M., Peay, K. G., and Tedersoo, L. (2015). Local-scale biogeography and spatiotemporal variability in communities of mycorrhizal fungi. *New Phytol.* 205, 1454–1463. doi: 10.1111/nph.13206

Barnard, E., Johnson, T., Ngo, T., Arora, U., Leuterio, G., Mcdowell, A., et al. (2020). Porphyrin Production and Regulation in Cutaneous Propionibacteria. *MSphere* 2020:5. doi: 10.1128/mSphere.00793-19

Baumann, P., Doudoroff, M., and Stanier, R. Y. (1968). Study of the *Moraxella* group. I. Genus *Moraxella* and the Neisseria catarrhalis group. *J. Bacteriol.* 95, 58–73. doi: 10.1128/jb.95.1.58-73.1968

Boxberger, M., Cenizo, V., Cassir, N., and La Scola, B. (2021). Challenges in exploring and manipulating the human skin microbiome. *Microbiome* 9:125. doi: 10.1186/s40168-021-01062-5

Byrd, A. L., Belkaid, Y., and Segre, J. A. (2018). The human skin microbiome. *Nat. Rev. Microbiol.* 16, 143–155. doi: 10.1038/nrmicro.2017.157

Chan, K. X., Phua, S. Y., and Van Breusegem, F. (2019). Secondary sulfur metabolism in cellular signalling and oxidative stress responses. *J. Exp. Bot.* 70, 4237–4250. doi: 10.1093/jxb/erz119

Chen, Y. E., Fischbach, M. A., and Belkaid, Y. (2018). Skin microbiota-host interactions. *Nature* 553, 427–436. doi: 10.1038/nature25177

Chen, Y. E., and Tsao, H. (2013). The skin microbiome: current perspectives and future challenges. *J. Am. Acad. Dermatol.* 69, 143–155. doi:

Conwill, A., Kuan, A. C., Damerla, R., Poret, A. J., Baker, J. S., Tripp, A. D., et al. (2022). Anatomy promotes neutral coexistence of strains in the human skin microbiome. *Cell Host Microbe* 30, 171–182e177. doi: 10.1016/j.chom.2021.12.007

Dekio, I., Asahina, A., and Shah, H. N. (2021). Unravelling the eco-specificity and pathophysiological properties of *Cutibacterium* species in the light of recent taxonomic changes. *Anaerobe* 71:102411. doi: 10.1016/j.anaerobe.2021.102411

Ernst, W. H. (1998). Sulfur metabolism in higher plants: potential for phytoremediation. *Biodegradation* 9, 311–318. doi: 10.1023/A:1008250827209

Flowers, L., and Grice, E. A. (2020). The Skin Microbiota: Balancing Risk and Reward. *Cell Host Microbe* 28, 190–200. doi: 10.1016/j.chom.2020.06.017

Grice, E. A., Kong, H. H., Conlan, S., Deming, C. B., Davis, J., Young, A. C., et al. (2009). Topographical and temporal diversity of the human skin microbiome. *Science* 324, 1190–1192. doi: 10.1126/science.1171700

Grice, E. A., and Segre, J. A. (2011). The skin microbiome. *Nat. Rev. Microbiol.* 9, 244–253. doi: 10.1038/nrmicro2537

Harris-Tryon, T. A., and Grice, E. A. (2022). Microbiota and maintenance of skin barrier function. *Science* 376, 940–945. doi: 10.1126/science.abo0693

Jonczyk-Matysiak, E., Weber-Dabrowska, B., Zaczek, M., Miedzybrodzki, R., Letkiewicz, S., Lusiak-Szelchowska, M., et al. (2017). Prospects of Phage Application in the Treatment of Acne Caused by *Propionibacterium acnes*. *Front. Microbiol.* 8:164. doi: 10.3389/fmicb.2017.00164

Juni, E. (1974). Simple genetic transformation assay for rapid diagnosis of *Moraxella osloensis*. *Appl. Microbiol.* 27, 16–24. doi: 10.1128/am.27.1.16-24.1974

Juni, E., and Bøvre, K. (2015). "Moraxella," in *Bergey's Manual of Systematics of Archaea and Bacteria*, (Hoboken, NJ: Wiley), 1–17. doi: 10.1002/9781118960608.gbm01204

Juni, E. B. K. (2015). *Bergey's Manual of Systematics of Archaea and Bacteria*. Hoboken, NJ: John Wiley & Sons, Inc.

Lee, H., Jeong, J., Oh, Y., Lee, C. J., Mun, S., Lee, D. G., et al. (2021). Comparative analysis of human facial skin microbiome between topical sites compared to entire face. *Genes Gen.* 43, 1483–1495. doi: 10.1007/s13258-021-01180-2

Lee, M. R., Nam, G. W., Jung, Y. C., Park, S. Y., Han, J. Y., Cho, J. C., et al. (2013). Comparison of the skin biophysical parameters of Southeast Asia females: forehead-cheek and ethnic groups. *J. Eur. Acad. Dermatol. Venereol.* 27, 1521–1526. doi: 10.1111/jdv.12042

Li, W.-H. (1997). *Molecular Evolution*. Sunderland, MA: Sinauer Associates.

Li, Z., Xia, J., Jiang, L., Tan, Y., An, Y., Zhu, X., et al. (2021). Characterization of the human skin resistome and identification of two microbiota cutotypes. *Microbiome* 9:47. doi: 10.1186/s40168-020-00995-7

Liu, J., Yan, R., Zhong, Q., Ngo, S., Bangayan, N. J., Nguyen, L., et al. (2015). The diversity and host interactions of *propionibacterium acnes* bacteriophages on human skin. *ISME J* 9, 2078–2093. doi: 10.1038/ismej.2015.47

Ma, Z., Li, L., and Li, W. (2018). Assessing and interpreting the within-body biogeography of human microbiome diversity. *Front. Microbiol.* 9:1619. doi: 10.3389/fmicb.2018.01619

Ma, Z. S. (2021). Niche-neutral theoretic approach to mechanisms underlying the biodiversity and biogeography of human microbiomes. *Evol. Appl.* 14, 322–334. doi: 10.1111/eva.13116

Machkova, L., Svadlak, D., and Doleckova, I. (2018). A comprehensive *in vivo* study of Caucasian facial skin parameters on 442 women. *Arch. Dermatol. Res.* 310, 691–699. doi: 10.1007/s00403-018-1860-6

Moss, C. W., Wallace, P. L., Hollis, D. G., and Weaver, R. E. (1988). Cultural and chemical characterization of CDC groups EO-2, M-5, and M-6, *Moraxella* (*Moraxella*) species, Oligella urethralis, Acinetobacter species, and Psychrobacter immobilis. *J. Clin. Microbiol.* 26, 484–492. doi: 10.1128/jcm.26.3.484-492.1988

Mukherjee, S., Mitra, R., Maitra, A., Gupta, S., Kumaran, S., Chakrabortty, A., et al. (2016). Sebum and hydration levels in specific regions of human face significantly predict the nature and diversity of facial skin microbiome. *Sci. Rep.* 6:36062. doi: 10.1038/srep36062

Oh, J., Byrd, A. L., Deming, C., Conlan, S., Kong, H. H., and Segre, J. A. (2014). Biogeography and individuality shape function in the human skin metagenome. *Nature* 514, 59–64. doi: 10.1038/nature13786

Oh, J., Byrd, A. L., Park, M., Program, N. C. S., Kong, H. H., and Segre, J. A. (2016). Temporal stability of the human skin microbiome. *Cell* 165, 854–866.

Perez Perez, G. I., Gao, Z., Jourdain, R., Ramirez, J., Gany, F., Clavaud, C., et al. (2016). Body site is a more determinant factor than human population diversity in the healthy skin microbiome. *PLoS One* 11:e0151990. doi: 10.1371/journal.pone.0151990

Rawlings, A. V. (2006). Ethnic skin types: are there differences in skin structure and function? *Int. J. Cosmet. Sci.* 28, 79–93. doi: 10.1111/j.1467-2494.2006.00302.x

Schaller, M., Loewenstein, M., Borelli, C., Jacob, K., Vogeser, M., Burgdorf, W. H. C., et al. (2005). Induction of a chemoattractive proinflammatory cytokine response after stimulation of keratinocytes with *propionibacterium acnes* and coproporphyrin III. *Br. J. Dermat.* 153, 66–71. doi: 10.1111/j.1365-2133.2005.06530.x

Schommer, N. N., and Gallo, R. L. (2013). Structure and function of the human skin microbiome. *Trends Microbiol.* 21, 660–668. doi: 10.1016/j.tim.2013.10.001

Shu, M., Kuo, S., Wang, Y., Jiang, Y., Liu, Y. T., Gallo, R. L., et al. (2013). Porphyrin metabolisms in human skin commensal *propionibacterium acnes* bacteria: potential application to monitor human radiation risk. *Curr. Med. Chem.* 20, 562–568. doi: 10.2174/0929867311320040007

Sparber, F., De Gregorio, C., Steckholzer, S., Ferreira, F. M., Dolowschiak, T., Ruchti, F., et al. (2019). The skin commensal yeast malassezia triggers a type 17 response that coordinates anti-fungal immunity and exacerbates skin inflammation. *Cell Host Microbe* 2019:25. doi: 10.1016/j.chom.2019.02.002

Spittaels, K. J., Van Uytfanghe, K., Zouboulis, C. C., Stove, C., Crabbe, A., and Coenye, T. (2021). Porphyrins produced by acneic *cutibacterium acnes* strains activate the inflammasome by inducing K(+) leakage. *iScience* 24:102575. doi: 10.1016/j.isci.2021.102575

Tan, J. K. L., and Bhate, K. (2015). A global perspective on the epidemiology of acne. *Br. J. Dermat.* 172(Suppl.):1. doi: 10.1111/bjd.13462

Tong, X., Leung, M. H. Y., Wilkins, D., Cheung, H. H. L., and Lee, P. K. H. (2019). Neutral processes drive seasonal assembly of the skin mycobiome. *mSystems* 2019:4. doi: 10.1128/mSystems.00004-19

Van Zuuren, E. J., Kramer, S., Carter, B., Graber, M. A., and Fedorowicz, Z. (2011). Interventions for rosacea. *Cochrane Database Syst. Rev.* 2011:CD003262. doi: 10.1002/14651858.CD003262.pub4

Voegeli, R., Gierschendorf, J., Summers, B., and Rawlings, A. V. (2019). Facial skin mapping: from single point bio-instrumental evaluation to continuous visualization of skin hydration, barrier function, skin surface pH, and sebum in different ethnic skin types. *Int. J. Cosmet. Sci.* 41, 411–424. doi: 10.1111/ics.12562

Wang, Y., Yu, Q., Zhou, R., Feng, T., Hilal, M. G., and Li, H. (2021). Nationality and body location alter human skin microbiome. *Appl. Microb. Biotechnol.* 105, 5241–5256. doi: 10.1007/s00253-021-11387-8

Williams, G. C. (1996). *Natural Selection, Ecology, and Morphogenesis in Adaptation and Natural Selection*. Princeton: Princeton University Press, 56–91.

Williams, H. C., Dellavalle, R. P., and Garner, S. (2012). Acne vulgaris. *Lancet* 379, 361–372. doi: 10.1016/S0140-6736(11)60321-8

Yi, H., Ravilious, G. E., Galant, A., Krishnan, H. B., and Jez, J. M. (2010). From sulfur to homoglutathione: thiol metabolism in soybean. *Amino Acids* 39, 963–978. doi: 10.1007/s00726-010-0572-9

Yigider, A. P., Kayhan, F. T., Yigit, O., Kavak, A., and Cingi, C. (2016). Skin diseases of the nose. *Am. J. Rhinol. Allergy* 30, 83–90. doi: 10.2500/ajra.2016.30.4318

Youn, S. W., Na, J. I., Choi, S. Y., Huh, C. H., and Park, K. C. (2005). Regional and seasonal variations in facial sebum secretions: a proposal for the definition of combination skin type. *Skin Res. Technol.* 11, 189–195. doi: 10.1111/j.1600-0846.2005.00119.x

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 08 August 2022
DOI 10.3389/fmicb.2022.923451

OPEN ACCESS

EDITED BY
Zheng Zhang,
Shandong University, China

REVIEWED BY
Hanif Ullah,
COMSATS Institute of Information Technology, Pakistan
Regina Sharmila Dass,
Pondicherry University, India

*CORRESPONDENCE
Changhong Liu
chliu@nju.edu.cn

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

RECEIVED 19 April 2022
ACCEPTED 19 July 2022
PUBLISHED 08 August 2022

CITATION
Liu X, Zain ul Arifeen M, Xue Y and Liu C (2022) Genome-wide characterization of laccase gene family in *Schizophyllum commune* 20R-7-F01, isolated from deep sediment 2 km below the seafloor. *Front. Microbiol.* 13:923451. doi: 10.3389/fmicb.2022.923451

Genome-wide characterization of laccase gene family in *Schizophyllum commune* 20R-7-F01, isolated from deep sediment 2 km below the seafloor

Xuan Liu, Muhammad Zain ul Arifeen, Yarong Xue[1] and Changhong Liu*

State Key Laboratory of Pharmaceutical Biotechnology, Nanjing University, Nanjing, China

Laccases are ligninolytic enzymes that play a crucial role in various biological processes of filamentous fungi, including fruiting-body formation and lignin degradation. Lignin degradation is a complex process and its degradation in *Schizophyllum commune* is greatly affected by the availability of oxygen. Here, a total of six putative laccase genes (*ScLAC*) were identified from the *S. commune* 20R-7-F01 genome. These genes, which include three typical Cu-oxidase domains, can be classified into three groups based on phylogenetic analysis. *ScLAC* showed distinct intron-exon structures and conserved motifs, suggesting the conservation and diversity of *ScLAC* in gene structures. Additionally, the number and type of cis-acting elements, such as substrate utilization-, stress-, cell division- and transcription activation-related cis-elements, varied between *ScLAC* genes, suggesting that the transcription of laccase genes in *S. commune* 20R-7-F01 could be induced by different substrates, stresses, or other factors. The SNP analysis of resequencing data demonstrated that the *ScLAC* of *S. commune* inhabiting deep subseafloor sediments were significantly different from those of *S. commune* inhabiting terrestrial environments. Similarly, the large variation of conserved motifs number and arrangement of laccase between subseafloor and terrestrial strains indicated that *ScLAC* had a diverse structure. The expression of *ScLAC5* and *ScLAC6* genes was significantly up-regulated in lignin/lignite medium, suggesting that these two laccase genes might be involved in fungal utilization and degradation of lignite and lignin under anaerobic conditions. These findings might help in understanding the function of laccase in white-rot fungi and could provide a scientific basis for further exploring the relationship between the LAC family and anaerobic degradation of lignin by *S. commune*.

KEYWORDS

***Schizophyllum commune*, subseafloor, laccase, lignin degradation, fungi, phylogeny**

Introduction

Laccase (benzenediol: oxygen oxidoreductase, EC 1.10.3.2) is a metalloprotein belonging to the group of polyphenol oxidases containing copper atoms in the catalytic site and therefore also called blue multicopper oxidases (Baldrian, 2006). Laccase proteins have three conserved domains (Cu-oxidase, Cu-oxidase_2, and Cu-oxidase_3) that are utilized to identify canonical laccases (Bento et al., 2006; Kudanga et al., 2011). Such an arrangement of copper atoms distributed in three domains is present in most of the bacterial and fungal laccases. The catalytic capacity of laccases is actually non-specific but, in most cases, laccases oxidize a range of aromatic compounds, including phenolic moieties typically found in lignin, aromatic amines, benzenothiols, and hydroxyindols as well as non-aromatic compounds, using molecular oxygen as an electron acceptor (Shiba et al., 2000; Claus, 2004; Rodríguez and Toca Herrera, 2006; Chirivì et al., 2012). Laccases have been implicated in a variety of physiological functions in living organisms due to their non-specific catalytic abilities.

Laccase was initially discovered in the Japanese lacquer tree by Yoshida (1883), and since then, it has been found in all domains of life: higher plants, some insects, a few bacteria, and fungi (Solomon et al., 1996; Alexandre and Zhulin, 2000; Claus, 2003). Basic characteristics and functions of laccases are diverse both within and across biological kingdoms. In plants, laccases participate in the radical-based mechanisms of lignin polymer formation (Berthet et al., 2011; Hu et al., 2018), while in fungi, laccases are hypothesized to play a variety of physiological roles, such as stress defense, melanin synthesis (Hua et al., 2018), fruiting-body formation (Lettera et al., 2010; Zhang et al., 2015), and lignin degradation (Singh and Sharma, 2010; Coconi et al., 2018). Lignin degradation is the most important function of fungal laccase. Laccases can directly depolymerize the lignin macromolecule, either alone or in combination with other enzymes. Laccases catalyze the removal of an electron from natural lignin's phenolic hydroxyl groups, resulting in free phenoxy radicals, and further oxidizes to quinones. Additionally, laccases decarboxylate phenolic and methoxyphenolic acid structures of lignin and cause their demethylation or demethoxylation (Leonowicz et al., 2001). Laccases also are useful biocatalysts for a wide range of biotechnological applications due to their high non-specific oxidation capacity and the use of readily available molecular oxygen as an electron acceptor (Mayer and Staples, 2002). In addition, laccases have important application values in various industrial processes, including textile refining, dye decolorization, bioremediation, lignocellulose delignification, organic synthesis, and food processing (Bilal et al., 2017; Mtibaà et al., 2018; Zhang et al., 2018).

Schizophyllum commune belongs to the white rot fungi and is one of the most widely distributed mushrooms on Earth. It is an effective wood-degrading basidiomycete that can produce a large number of hydrolases such as xylanase (Gautam et al., 2018), pectinase (Mehmood et al., 2019), cellulase (Kumar B. et al., 2018), endoglucanase, glycoside hydrolase, and oxidoreductase (Tovar-Herrera et al., 2018). Genome sequencing of a terrestrial strain H4-8 revealed two laccase genes and four genes encoding a distant relative of laccase (Ohm et al., 2010). Our previous study indicated that *S. commune* was the predominant fungal species in deep subseafloor coal-bearing sediments ranging from ∼1.5 to ∼2.0 km below the seafloor (kmbsf), and could grow under both anaerobic and aerobic culture conditions (Liu et al., 2017; Zain Ul Arifeen et al., 2020). Compared with other environmental isolates, strain 20R-7-F01 of *S. commune* isolated from ∼20-million-year-old coal-bearing sediment at 1,966.3 kmbsf has a stronger ability to adapt to *in situ* environmental conditions, including carbon (energy) source, temperature, oxygen, and nitrogen source (Zain Ul Arifeen et al., 2020).

Although laccases were identified and classified in various *S. commune* strains (Kumar et al., 2015, Kumar V. P. et al., 2018; Zhao et al., 2018; Kirtzel et al., 2019), an investigation of the laccase gene family in *S. commune* at the whole-genome level is yet to be conducted. In this study, we identified all possible laccase-coding genes from the *S. commune* reference genome (20R-7-F01). We then analyzed the physical and chemical properties, gene structure, amino acid sequence, systematic evolution, and expression patterns of the gene family in media with or without lignin/lignite. The results could facilitate the understanding of the laccase function in white-rot fungi and provide a scientific basis for further exploring the relationship between the LAC family and the anaerobic degradation of lignin by *S. commune*.

Materials and methods

Strains and culture conditions

The fungal strains were isolated from subseafloor sediment, which was collected by drilling vessel at Site C0020 (41°10.5983′N, 142°12.0328′E) in the Pacific Plate off the Shimokita Peninsula, Japan, during the IODP Expedition 337, at a water depth of 1,180 m (Inagaki et al., 2015; Liu et al., 2017). Briefly, the sediment samples were ground into powder in an anaerobic chamber with a flame-sterilized hammer, placed evenly on three petri dishes containing specific media that simulated to the *in situ* environmental conditions, and incubated at 30°C for 7∼14 days (Liu et al., 2017). *S. commune* strains 6R-2-F01, 15R-5-F01, 20R-7-F01, and 24R-3-F01 were obtained from the sediment samples at to 1,496; 1,924; 1,966, and 1,993 mbsf, respectively. Two terrestrial strains CFCC_7252 and CFCC_86625 were purchased from China Forestry Culture Collection Center, which were isolated from Populus wood in Songshan, Beijing and Jurong, Jiangsu of China, respectively.

Strain MCCC_3A00233 collected from marine sediment of the Atlantic Ocean was purchased from Third Institute of Oceanography, State Oceanic Administration, People's Republic of China. and the other five terrestrial strains (225DK, 227DK, MF, Hom2-8, and 207) were obtained from NCBI and JGI database. Details of the habitat and culture conditions of *S. commune* strains have been described previously (Liu et al., 2022). All the fungal strains were maintained on potato dextrose agar (PDA) at 30°C. For DNA and RNA isolations, the fresh mycelia of *S. commune* were inoculated into a 250-ml conical flask containing 150 ml PD (200 g/L potato, 20 g/L glucose) and incubated in a shaking chamber at 30°C, 200 rpm for 7 days.

Identification of laccase gene family members in *Schizophyllum commune* 20R-7-F01

The *S. commune* 20R-7-F01 genome was assembled using SMRT Analysis and deposited in GenBank under the accession number VCHW00000000. Laccase members contain Cu-oxidase, Cu-oxidase_2, and Cu-oxidase_3 (PF00394, PF07731 and PF07732) domains. The three domains were searched in the *S. commune* 20R-7-F01 genome using HAMMER software (Finn et al., 2011), and protein sequences with three Cu-oxidase domains in the LAC domain were recognized as members of the LAC family. The laccase gene was named using the prefix Sc for *S. commune* followed by the *LAC* gene family abbreviation and numbered sequentially according to their position on unitigs.

Physical map of *Schizophyllum commune* 20R-7-F01 laccase genes and properties of laccase proteins

Using the *S. commune* 20R-7-F01 genome, the unitig length and the starting position of genes on unitigs were obtained. After statistical analysis, the physical distribution map of their unitigs was visualized using Mapchart 2.32 software (Voorrips, 2002). The theoretical isoelectric point (pI) and molecular weight (MW) of ScLAC proteins were analyzed using the Compute pI/MW tool on the ExPASy server[1] (Wilkins et al., 1999). Subcellular locations of the ScLAC members were determined using the online software CELLO[2] (Yu et al., 2006). Signal peptides of each laccase were predicted using SignalP algorithm[3] (Nielsen et al., 1997). Prediction of transmembrane regions was performed with TMHMM Server[4] (Krogh et al., 2001). The glycosylation sites of the ScLAC members were predicted by NetNGlyc 1.0[5] (Gupta and Brunak, 2002).

Analysis of gene structure and motif composition

The sequence of laccase genes and their coding region were first transformed into FASTA format then matched, and intron/exon structure was determined by comparing the coding sequence of each *ScLAC* gene with its genomic sequence using the Gene Structure Display Server 2.0[6] (Hu B. et al., 2015). In addition, the upstream regions (1.5 kb) of the *ScLAC* gene sequences were extracted and used for the search of cis-elements using YEASTRACT[7] (Monteiro et al., 2020). Conserved motifs of laccase proteins were identified statistically using MEME[8] (Bailey et al., 2009), and the maximum number of motifs to find was set at 10. Visualization of motif compositions was executed with TBtools V1.09 (Chen et al., 2020).

Sequence alignment and phylogenetic analysis

The identified ScLAC amino acid sequences were aligned separately against each other using ClustalW in MEGA7.0 (Kumar et al., 2016). The conserved regions of ScLAC were used to build the phylogenetic tree. The unrooted phylogenetic tree was created using MEGA7.0 by a neighbor-joining algorithm with bootstrap replication of 100 times. The final phylogenetic tree was visualized and edited in iTOL[9] (Letunic and Bork, 2016).

Genome resequencing and variant calling

The genome resequencing and variant detection for *S. commune* strains were carried out according to our previous methods (Liu et al., 2022). Briefly, the genome DNA of *S. commune* strains was extracted and fragmented to generate an approximately 300 bp library insert size and sequenced on an Illumina HiSeq 2500 platform at BGI Genomic (Shenzhen, China). The filtered resequencing reads were mapped to the reference genome of *S. commune* 20R-7-F01 for SNP and variant detection.

1 http://web.expasy.org/
2 http://cello.life.nctu.edu.tw/
3 https://services.healthtech.dtu.dk/service.php?SignalP-5.0
4 https://services.healthtech.dtu.dk/service.php?TMHMM-2.0
5 https://services.healthtech.dtu.dk/service.php?NetNGlyc-1.0
6 http://gsds.gao-lab.org/index.php
7 http://www.yeastract.com/index.php
8 http://meme-suite.org/tools/meme
9 http://itol.embl.de/

Transcriptome analysis

Total RNA was extracted from mycelia of strain 20R-7-F01 that were cultured in bottles containing lignin and lignite medium, and incubated under anaerobic (i.e., LigWO1-3 and CoalWO1-3) and aerobic (i.e., LigO1-3 and CoalO1-3) condition for seven days (30°C), respectively. Each treatment included three replicates. Lignite was collected from coal mine in Xinjiang. It contains N 1.15%, C 68.7%, H 4.123%, S 1.642%, organic component 98.17%, and inorganic component 1.83%, and vitrinite reflectance was 0.49%. Lignin alkali was purchased from Sigma (CAS# 8068-05-1), which contains 5% moisture. After sampling, all mycelia were immediately frozen in a liquid nitrogen tank and delivered to the Personal Biotechnology Company (Shanghai, China) for mRNA extraction, cDNA library construction, and sequencing. After trimming of low-quality reads ($Q < 20$) and adapter contamination, the clean reads were mapped to the assembled genome of strain 20R-7-F01 using TopHat (Trapnell et al., 2009). Gene prediction was performed using Cufflinks (Roberts et al., 2011). To compare the gene expression level in different libraries, the transcript level of each expressed gene was calculated and normalized to the reads per kilobase of exon model per million mapped reads (RPKM). We used DESeq software for differential analysis of gene expression (Anders and Huber, 2010). Genes with an adjusted *p*-value ≤ 0.01 and an absolute value of log2 (expression-fold change) ≥ 1 were deemed to be differentially expressed (Hu L. et al., 2015). The Pheatmap software package in R language was used to perform bidirectional cluster analysis of differential genes and samples. Distances were calculated using the Euclidean method and clustered by complete linkage.

Results

Laccase gene family of *Schizophyllum commune* 20R-7-F01

To identify the laccase genes in *S. commune* 20R-7-F01, we searched the genome with HAMMER software for Cu-oxidase, Cu-oxidase_2, and Cu-oxidase_3 domains (PF00394, PF07731, and PF07732). Six putative laccase genes (*ScLAC1* to *ScLAC6*) were identified (**Table 1**) and mapped to six of the 162 *S. commune* 20R-7-F01 unitigs (**Figure 1**), indicating that the *ScLAC* gene family did not have the characteristics of tandem replication or clustering.

ScLAC proteins

Basic information on all *S. commune* 20R-7-F01 laccases, including gene name, physical location, amino acid length, molecular weight, pI value, subcellular localization, signal peptide and transmembrane topology, were presented in **Table 1**. The length of laccase proteins ranged from 374 aa (ScLAC6) to 1,137 aa (ScLAC3) residues, and the predicted molecular weights were between 41.08 kDa (ScLAC6) and 125.25 kDa (ScLAC3). The predicted pI-values of the laccase proteins were found to be in the range of 4.62 (ScLAC2) to 6.56 (ScLAC6), indicating that they belonged to acidic proteins. The predicted subcellular locations revealed that the six laccase proteins were located in cytoplasm, nucleus, and mitochondria, and were also found extracellularly. ScLAC1, ScLAC4, and ScLAC6 were predicted to be localized only in the extracellular space, whereas ScLAC3 was located either in the nucleus or mitochondria, ScLAC2 was located in the cytoplasm, and ScLAC5 was located in either the cytoplasm or extracellular. ScLAC4–ScLAC6 were probably signal proteins, while ScLAC1–ScLAC3 may not contain any signal regions. ScLAC6 had transmembrane topology, while the other five laccases did not contain transmembrane domains. Additionally, variable N-glycosylation sites were predicted to be present in all ScLAC proteins (**Table 1**), indicating that ScLAC family exhibited potential post-translational modifications.

Gene structure, motif compositions, and phylogeny of *ScLAC*

To reveal the structural diversity of *S. commune* 20R-7-F01 laccase genes, we constructed the exon/intron organization and searched for conservative motifs based on the phylogenetic tree of all *S. commune* 20R-7-F01 laccase alignments (**Figure 2**). Phylogenomic analysis showed that the *S. commune* 20R-7-F01 laccase gene family was clustered into three branches, of which *ScLAC1* and *ScLAC2* were one clade, *ScLAC4* and *ScLAC6* were another clade, and *ScLAC5* and *ScLAC3* were the last clade (**Figure 2**). In addition, to evaluate the number of laccase genes in the genome of *S. commune* 20R-7-F01, the total number of laccase genes was determined in other Agaricales. The total number of laccase genes varied significantly among species, ranging from 4 in *Hebeloma cylindrosporum* and *Postia placenta* to 55 in *Dendrothele bispora* (**Supplementary Figure 1**). The amount of laccase in Schizophyllaceae was relatively small compared to other species. In addition, the total number of laccases and protein-coding genes were normalized by genome assembly (in Mb) to avoid potentially misleading comparisons due to differences in genome size and total number of genes among the investigated species. No positive correlation was found between genome size or total number of predicted genes and the number of laccase genes in the corresponding genome (**Supplementary Figure 1**). For instance, *D. bispora* showed the highest number of laccases (55), but *Moniliophthora perniciosa* had the highest proportion of laccases per total number of genes (0.22%).

TABLE 1 Basic information of *Schizophyllum commune* laccases.

Gene name	Gene ID	Physical location	AAL (aa)	MW (D)	pI	SL	SP	TR	N-Glyc
ScLAC1	unitig_8.g1172	3898655–3900801	565	62,635.26	4.81	Extracellular	N	N	7
ScLAC2	unitig_405.g35	137770–140761	511	56,942.17	4.62	Cytoplasmic	N	N	4
ScLAC3	unitig_432.g31	101263–107421	1,137	125,249.48	6.55	Nuclear, mitochondrial	N	N	6
ScLAC4	unitig_430.g21	92411–94931	564	61,367.09	5.62	Extracellular	Y	N	5
ScLAC5	unitig_21.g21	70020–72642	651	71,418.18	4.89	Cytoplasmic, extracellular	Y	N	10
ScLAC6	unitig_35.g69	325705–328463	374	41,084.33	6.56	Extracellular	Y	Y	4

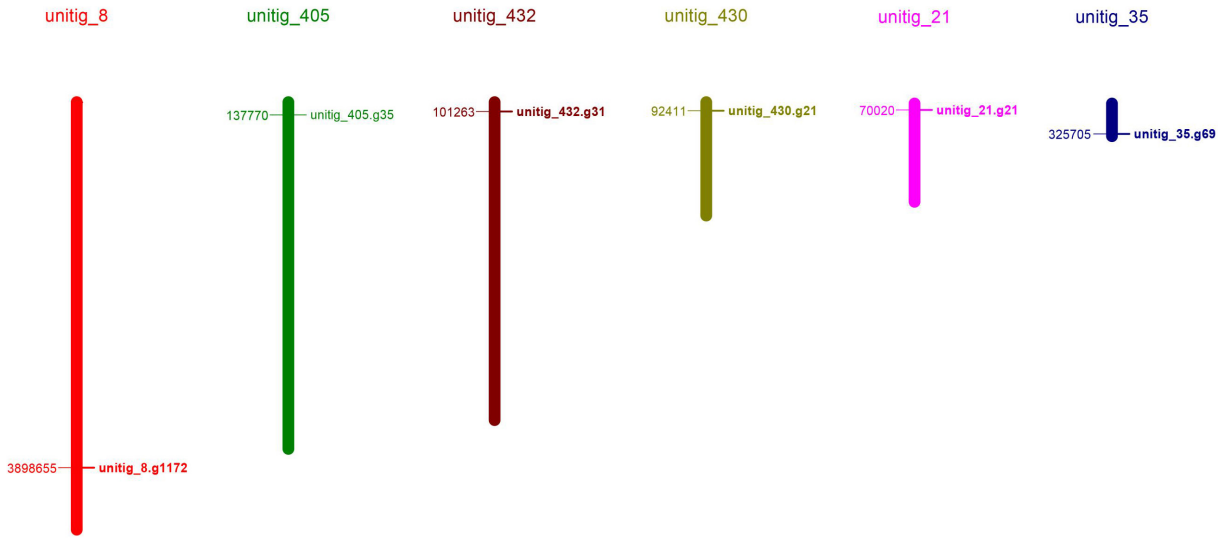

FIGURE 1
Physical map of the *Schizophyllum commune* 20R-7-F01 laccase genes.

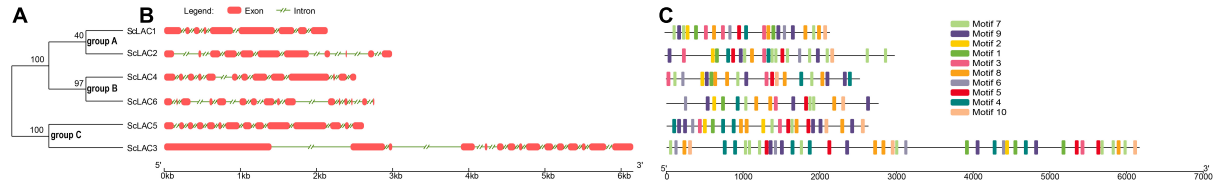

FIGURE 2
Phylogenetic relationships, gene structure, and motif compositions of the *Schizophyllum commune* 20R-7-F01 laccase gene family. **(A)** A neighbor-joining tree of six ScLAC protein sequences constructed using MEGA v7.0. **(B)** The structure of the six ScLAC genes. Red squares correspond to exons and shrinked green lines indicate introns. **(C)** Schematic motif composition of six ScLAC genes. The colored boxes represent the different motifs, indicated in the top right-hand corner. The scales at the bottom of the image indicate the estimated exon/intron and motif length in kb.

The number of introns of *ScLAC* family members varied from 8 to 15. Surprisingly, nearly all of the closest genes on the phylogenetic tree showed remarkably different gene structures. For instance, the introns and exons of *ScLAC5* were most closely arranged, whereas its nearby paralogous gene *ScLAC3* had the longest intron, although their evolutionary relationship reached a 100% bootstrap value. Additionally, *ScLAC6* had the most introns; its coding sequences were divided into 15 parts by introns. In short, *ScLAC3*, *ScLAC5*, and *ScLAC6* were more complicated than the other laccase genes with respect to their structure.

To further reveal the conserved motifs of the ScLAC proteins, we analyzed six ScLAC proteins and identified 10 motifs using the MEME program (Figure 2). As expected, the motif compositions of peer groups had different structures and organizations, which indicated the possibility of functional

divergence among those proteins. Although 10 motifs were found in every ScLAC protein, there were some differences in the number of occurrences. For instance, motif-4 was repeated three and six times in ScLAC5 and ScLAC3, respectively. This difference in motif number across ScLAC proteins indicated that different ScLAC proteins may have different functions.

Cis-regulatory elements predicted in the *ScLAC* promoters

To obtain further insights into the possible regulatory patterns of ScLAC, we analyzed the cis-acting elements of the 1.5 kb regulatory sequence upstream of the six *ScLAC* gene sequences using the Yeastract database. The promoter regions of *ScLAC1–ScLAC6* included various functional cis-acting elements (Figure 3 and Supplementary Table 1) associated with substrate utilization, stress, cell division, and transcription activation. Among them, *ScLAC6* had the most cis-elements, including 16 stress-related, 15 substrate utilization-related, nine cell division-related, and eight amino acid transcription-related cis-elements. In addition, these laccases also contained specific cis-elements; for example, *ScLAC2* contained one specific cis-element, named Nrg2p, which mediated glucose repression and negatively regulated filamentous growth, while *ScLAC3* contained four specific cis-elements, which negatively regulated nitrogen catabolic gene expression and were involved in induction of *CLN3* transcription in response to glucose (Supplementary Figure 2 and Supplementary Table 1). The differences in the number and types of cis-acting elements in *S. commune* 20R-7-F01 suggested that the transcription

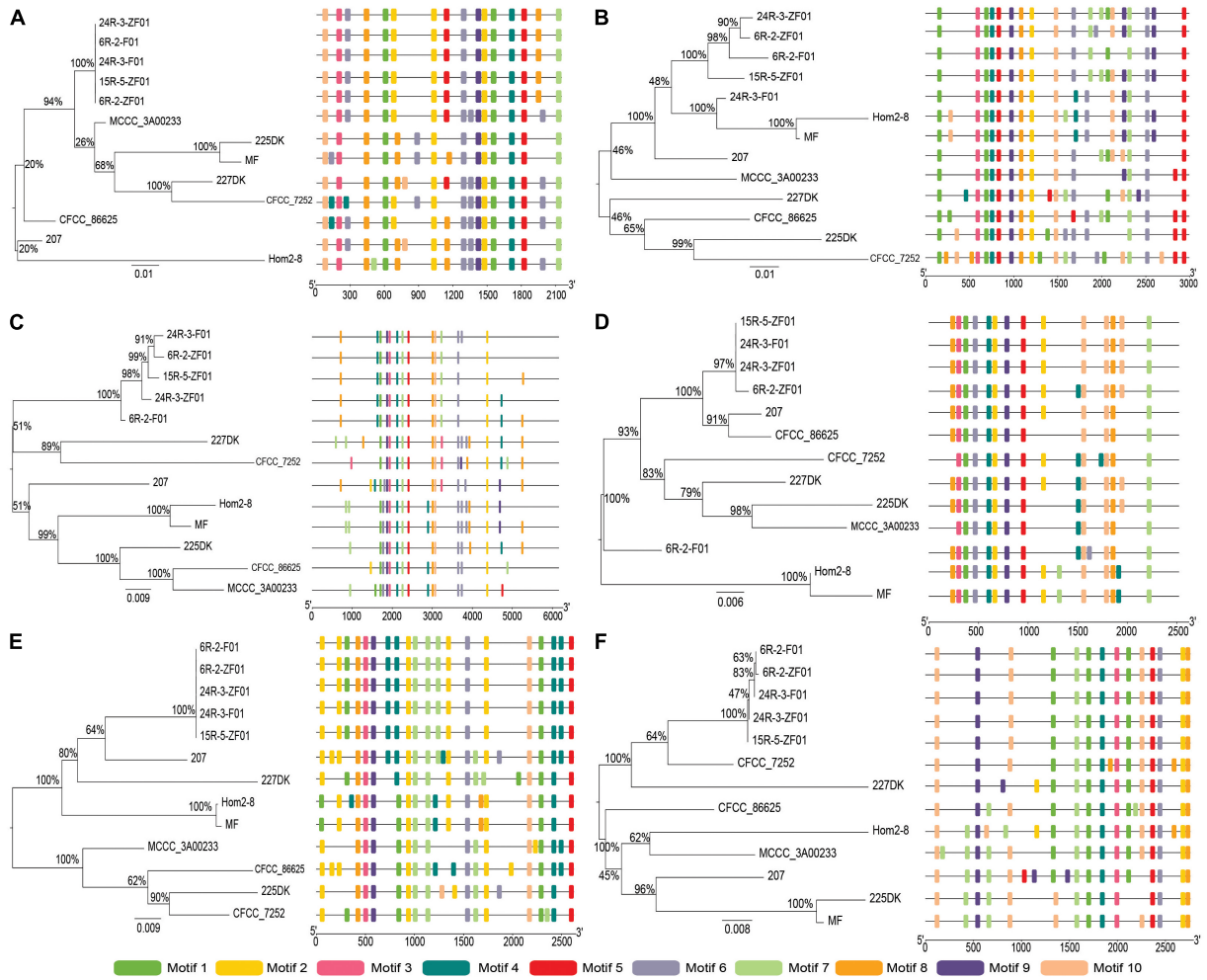

FIGURE 3
Phylogenetic and motif compositions of the *Schizophyllum commune* population laccase gene family. **(A–F)** Represent neighbor-joining tree of ScLAC1-ScLAC6 protein sequences in *S. commune* population constructed using MEGA v7.0, and schematic motif composition of ScLAC1-ScLAC6 genes in *S. commune* population. The colored boxes represent the different motifs, indicated in the top right-hand corner. The scales at the bottom of the image indicate the motif length in kb.

of laccase genes may be regulated by substrates, stresses, or other factors.

Laccase differences between subseafloor and terrestrial environments

To understand the evolutionary relationships between subseafloor and terrestrial laccase genes, five subseafloor and eight terrestrial strains were re-sequenced. Phylogenetic analysis based on SNP mutation sites showed that the *ScLAC* of *S. commune* strains inhabiting deep subseafloor sediments differed significantly from that of *S. commune* strains inhabiting terrestrial environments (**Figure 3**). The average number of SNP mutation sites of the five *ScLAC* in the terrestrial strains was greater than those in subseafloor strains, with the exception of *ScLAC3*, although it contained the most SNP mutation sites (**Supplementary Tables 2, 3**). In addition, the number and arrangement of conserved motifs in laccases between subseafloor and terrestrial strains also showed various differences, suggesting that *ScLAC* possessed a diverse structure. Among them, the conserved domains of ScLAC2–ScLAC4 were different in both subseafloor and terrestrial strains, indicating that the evolution of these three laccase genes was not only related to habitat but also related to strains (**Figure 3**).

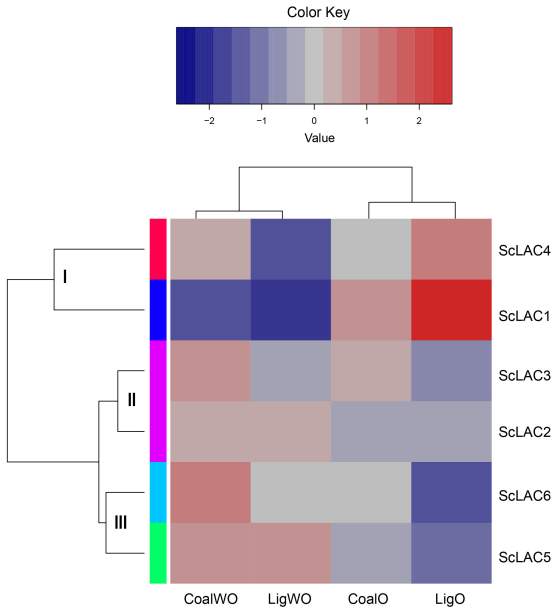

FIGURE 4
Heatmap of the expression profiles of laccase genes from *Schizophyllum commune* 20R-7-F01 in various carbon sources with or without oxygen. The heatmaps with hierarchical clustering were visualized using the software heatmap2 and the values were log2-transformed with normalization. The blue and red elements indicate low and high relative expression levels, respectively.

Transcriptome analysis of six putative laccase in lignite/lignin degradation

RNA-seq analysis of strain 20R-7-F01 cultured in lignin/lignite-containing medium under aerobic and anaerobic conditions at 30°C for 7 days showed that the six laccase genes could be classified into three groups (I, II, and III) (**Figure 4**). The relative expression levels of *ScLAC1* and *ScLAC4* were lower under anaerobic conditions than under aerobic conditions. In contrast, the laccase genes of group II (*ScLAC2* and *ScLAC3*) and group III (*ScLAC5* and *ScLAC6*) tended to be induced by anaerobic conditions (**Figure 4**). Additionally, compared with aerobic condition, the expression of *ScLAC5* and *ScLAC6* genes was upregulated by 2.48- and 2.10-fold in the anaerobic conditions (**Supplementary Table 4**), suggesting that these two laccase genes may be involved in anaerobic utilization and degradation of lignite and lignin by fungi.

Discussion

Laccases as ligninolytic enzymes play important roles in various biological processes of fungi, including lignin degradation and fruiting-body formation, are typically encoded by gene families (Vasina et al., 2015). Through genome-sequencing analysis, we identified six laccase genes in subseafloor *S. commune* 20R-7-F01, all of which contained three conserved Cu-oxidase domains. However, significant differences were found among these six genes; for instance, very few amino acid sequence similarities were found, and their exon-intron structures were diversified. This suggests that the laccase genes of subseafloor *S. commune* have vast evolutional and functional diversity.

The subcellular localization of proteins is invaluable for understanding their functions and interactions with other proteins (Peng and Gao, 2014). Based on subcellular localization analysis, we found that the ScLAC were located in extracellular, cytoplasmic and nuclear, and mitochondrial. The wide distribution of ScLAC in *S. commune* indicated that these enzymes might have distinct roles in response to various environments (Yang et al., 2021).

The amino acid sequence of fungal laccase generally contains a signal peptide sequence at the N-terminus to guide transmembrane transfer (Jiao et al., 2018). However, some fungal strains have no signal peptide sequence in the laccase gene. For instance, LeLAC3 of *Lentinula edodes* contained a signal peptide sequence in strain D703PP-9, but was absent in strain W1-26 (Sakamoto et al., 2015; Yan et al., 2019). The deficiency of signal peptide sequence was also reported in *Flammulina velutipes* and *Setosphaeria turcica*

(Wang et al., 2015; Liu et al., 2019), as well as in plants (Xu et al., 2019). Our study also proved that ScLAC4–ScLAC6 in subseafloor *S. commune* 20R-7-F01 possessed the signal peptide, while ScLAC1–ScLAC3 did not. These data suggest that the laccase genes differ not only between species but also within species.

In fungi, laccase genes differentiate into many paralogous genes and perform various functions throughout the fungal life cycle (Kumar et al., 2003). They are usually clustered in the form of scaffolds; for instance, the 12 laccase genes in *Pleurotus ostreatus* and 13 laccase genes in *L. edodes* were mapped into six and seven scaffolds, respectively (Sakamoto et al., 2015; Jiao et al., 2018). Here, we identified six laccase genes scattered in six unitigs, which were thought to be only specific to *S. commune* 20R-7-F01 genome; similarly, six laccase genes were distributed on five scaffolds of *S. commune* strain H4-8 genome. Therefore, we inferred that this difference might be related to species and strain differences.

Gene structure and protein motif analyses can provide a theoretical basis for the function and classification of laccase family genes (Wang et al., 2019). In general, genes in the same group and subgroup should have a similar conserved domain and motif distribution with closely related members in the phylogenetic tree, revealing the functional similarity between proteins in the same subgroup proteins (Yang et al., 2021). Inconsistent with the results of previous studies, we found that the most closely related members of laccase genes in the phylogenetic tree of *S. commune* had highly diverse motif compositions, and that the conserved motifs of laccases between subseafloor and terrestrial strains were highly diverse. The number and length of introns and exons in *ScLAC* genes were significantly different. In general, groups B and C had more exons and introns than group A (**Figure 2**). A small number of introns in a gene usually are the result of genetic evolution, which can rapidly regulate genes in response to stress (Stival Sena et al., 2014). Introns are another source of sequence variation (Jacob and Smith, 2017; Naro and Sette, 2017), and intron retention may increase the diversity of proteins and the complexity of genes expression (Zhang et al., 2004).

Cis-elements play significant roles in the regulatory process to respond to multiple abiotic stresses (Feng et al., 2016). The various cis-elements found in the promoter regions of *ScLAC* genes were classified into four major groups: substrate utilization-related, stress-related, cell division-related, and amino acid transcription-related cis-elements. These cis-elements may be recognized by some transcription factors and were thus involved in the regulation and expression of *ScLAC* genes. The presence of multiple cis-elements suggests that *ScLAC* genes could be involved in fungal response to multiple stresses. Laccases are thought to play an important role in fruiting-body formation (Lettera et al., 2010; Zhang et al., 2015) and our recent investigation found that the biosynthesis of amino acids also helps in the formation of fruiting bodies (Zain Ul Arifeen et al., 2021). Thus, the activation of amino acid transcription-related cis-elements in *ScLAC* genes could explain the possible role of *ScLAC* genes in fruiting-body formation.

Lignocellulose degradation by *S. commune* is an important but complex process, which needs to be thoroughly understood. *S. commune* utilizes more than 150 genera of woody plants and can also colonize softwood and grass silage (Ohm et al., 2010). As a model mushroom, *S. commune* H4-8 has complete genome sequence and annotation, and possesses the most extensive polysaccharide decomposition mechanism. The genome of strain H4-8 is rich in the glycoside hydrolase family (hemicellulose and pectin degradation) and polysaccharide lyase family (pectin degradation), which enables it to degrade all plant cell wall components, including lignin (Ohm et al., 2010; Sornlake et al., 2017). Fungi are known to possess a variety of lignin degrading enzymes including lignin peroxidase, manganese peroxidase, dye decolorizing peroxidase, multifunctional peroxidase, and laccase (Floudas et al., 2012). Among these enzymes, laccases are the primary tool lignin degradation in most basidiomycetes (white-rot fungi) and litter-decomposing saprotrophic fungi (Janusz et al., 2020). Laccase catalyzes the one-electron oxidation of substituted phenols, aniline, and aromatic thiols to corresponding free radicals, and reduces molecular oxygen to water (Qi et al., 2015). The broad substrate specificities of laccases, coupled with their use of molecular oxygen as the final electron acceptor rather than the hydrogen peroxide used by ligninolytic peroxidases, makes these enzymes suitable for lignin degradation. However, laccase can only directly degrade phenolic compounds with low-redox-potential, but cannot oxidize the most recalcitrant aromatic hydrocarbons. Nevertheless, some low-molecular-weight compounds produced by fungal degradation of lignin can act as redox mediators to promote the oxidation of refractory substrates (e.g., the non-phenolic lignin moiety) by laccases (Eggert et al., 1996; Camarero et al., 2005).

Basically, laccase use molecular oxygen as the final electron acceptor, and its activity is driven by the concentration of available oxygen (Qi et al., 2015). However, it has been proved that laccase can also oxidize catechol, *o*-aminophenol, *p*-aminophenol, *o*-phenylenediamine, and *p*-phenylenediamine under anaerobic conditions, with activities of 0.978, 0.707, 0.437, 3.603, and 1.039 mg μmol^{-1} min^{-1}, respectively (Xie et al., 1999). Shleev et al. (2005) observed direct electron transfer (DET) between the gold electrode and the laccase of *Trametes hirsuta* under anaerobic conditions. Our previous study also proved that laccase may be involved in the anaerobic degradation of phenanthrene by *S. commune* 20R-7-F01 (Zain Ul Arifeen et al., 2022). Based on the transcriptome analysis of *S. commune* 20R-7-F01 during lignin/lignite degradation, we found for the first time that the expression of *ScLAC1* was significantly downregulated under anaerobic conditions, while the expression of *ScLAC5* and *ScLAC6* was significantly up-regulated, compared with that under aerobic conditions

(**Figure 4** and **Supplementary Table 4**). These data suggested that *ScLAC5* and *ScLAC6* may play an important role in the utilization of lignite/lignin and other carbon sources by fungi in anaerobic environment. However, the anaerobic catalytic mechanism of laccase and its effect on fungi to obtain nutrition and energy in the anaerobic subseafloor environments need to be further studied.

Conclusion

A total of six putative laccase genes (*ScLAC*) with three typical Cu-oxidase domains were identified in *S. commune* 20R-7-F01 genome. The physical locations of these genes were mapped in six of 162 unitigs of *S. commune* 20R-7-F01. The theoretical pI of deduced ScLAC proteins widely ranged from 4.62 to 6.56. The MW of the ScLAC proteins ranged from 41.08 to 125.25 kDa and the length varied between 374 and 1,137 amino acids. Based on phylogenetic analysis, the six *ScLAC* genes were classified into three groups with distinct intron-exon structures and conserved motif. All of the ScLAC had cis-elements related to substrate utilization, stress, cell division, and activates transcription of amino acid in the promoter regions, while the number and type of cis-elements had difference between each other. The phylogenetic tree of resequencing data shows that there are many differences in the number and arrangement of conserved motifs between the ScLAC gene of S. commune strains inhabiting deep subseafloor sediments and the ScLAC gene of strains inhabiting terrestrial environments. The expressions of *ScLAC5* and *ScLAC6* genes were significantly upregulation under anaerobic conditions, implying that these two laccase genes might be involved in anaerobic utilization and degradation of lignite and lignin by fungi.

In summary, we identified all possible laccase-coding genes from the *S. commune* reference genome (20R-7-F01) and analyzed the physical and chemical properties, gene structure, amino acid sequence, and systematic evolution, also studied the expression patterns of the gene family under anaerobic and aerobic by growing in lignin/lignite medium. Our data and analysis could facilitate the understanding of the laccase function of white-rot fungi and provide a scientific basis for further exploring the relationship between the *ScLAC* genes family and the anaerobic degradation of lignin by *S. commune*.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA858196; https://www.ncbi.nlm.nih.gov/, PRJNA738972.

Author contributions

XL performed data analysis and wrote the first draft of the manuscript. MZ helped with data analysis and edited the manuscript. YX cultivated strains of *S. commune* and edited the manuscript. CL conceived the study and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China (nos. 91951121, 41973073, and 41773083).

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.923451/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Comparative analysis of the total number of laccases genes and their distribution across the different subclades in the genome of Agaricales. Percentage of laccases in total number of genes was calculated as following: (total number of laccases/total number of predicted genes in the genome) ×100. Genome size and number of predicted genes were retrieved from NCBI or JGI and refer to the current version of the assembled genome of each species.

SUPPLEMENTARY FIGURE 2
Venn analysis of predicted cis-elements in the promoter regions of laccase genes from *Schizophyllum commune* 20R-7-F01.

References

Alexandre, G., and Zhulin, I. B. (2000). Laccases are widespread in bacteria. *Trends Biotechnol.* 18, 41–42. doi: 10.1016/S0167-7799(99)01406-7

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Baldrian, P. (2006). Fungal laccases–occurrence and properties. *FEMS Microbiol. Rev.* 30, 215–242. doi: 10.1111/j.1574-4976.2005.00010.x

Bento, I., Carrondo, M. A., and Lindley, P. F. (2006). Reduction of dioxygen by enzymes containing copper. *J. Biol. Inorg. Chem.* 11, 539–547. doi: 10.1007/s00775-006-0114-9

Berthet, S., Demont-Caulet, N., Pollet, B., Bidzinski, P., Cézard, L., Le Bris, P., et al. (2011). Disruption of LACCASE4 and 17 results in tissue-specific alterations to lignification of *Arabidopsis thaliana* stems. *Plant Cell* 23, 1124–1137. doi: 10.1105/tpc.110.082792

Bilal, M., Asgher, M., Parra-Saldivar, R., Hu, H., Wang, W., Zhang, X., et al. (2017). Immobilized ligninolytic enzymes: An innovative and environmental responsive technology to tackle dye-based industrial pollutants-A review. *Sci. Total Environ.* 576, 646–659. doi: 10.1016/j.scitotenv.2016.10.137

Camarero, S., Ibarra, D., Martínez, M. J., and Martínez, A. T. (2005). Lignin-derived compounds as efficient laccase mediators for decolorization of different types of recalcitrant dyes. *Appl. Environ. Microbiol.* 71, 1775–1784. doi: 10.1128/AEM.71.4.1775-1784.2005

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chirivì, C., Fontana, G., Monti, D., Ottolina, G., Riva, S., and Danieli, B. (2012). The quest for new mild and selective modifications of natural structures: Laccase-catalysed oxidation of ergot alkaloids leads to unexpected stereoselective C-4 hydroxylation. *Chemistry* 18, 10355–10361. doi: 10.1002/chem.201201076

Claus, H. (2003). Laccases and their occurrence in prokaryotes. *Arch. Microbiol.* 179, 145–150. doi: 10.1007/s00203-002-0510-7

Claus, H. (2004). Laccases: Structure, reactions, distribution. *Micron* 35, 93–96. doi: 10.1016/j.micron.2003.10.029

Coconi, L. N., Fernández, F., Loske, A. M., and Gómez-Lim, M. A. (2018). Enhanced delignification of lignocellulosic biomass by recombinant fungus *Phanerochaete chrysosporium* overexpressing laccases and peroxidases. *J. Mol. Microbiol. Biotechnol.* 28, 1–13.

Eggert, C., Temp, U., Dean, J. F., and Eriksson, K. E. (1996). A fungal metabolite mediates degradation of non-phenolic lignin structures and synthetic lignin by laccase. *FEBS Lett.* 391, 144–148. doi: 10.1016/0014-5793(96)00719-3

Feng, K., Yu, J., Cheng, Y., Ruan, M., Wang, R., Ye, Q., et al. (2016). The SOD gene family in tomato: Identification, phylogenetic relationships, and expression patterns. *Front. Plant Sci.* 7:1279. doi: 10.3389/fpls.2016.01279

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Floudas, D., Binder, M., Riley, R., Barry, K., Blanchette, R. A., Henrissat, B., et al. (2012). The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* 336, 1715–1719. doi: 10.1126/science.1221748

Gautam, A., Kumar, A., Bharti, A. K., and Dutt, D. (2018). Rice straw fermentation by *Schizophyllum commune* ARC-11 to produce high level of xylanase for its application in pre-bleaching. *J. Genet. Eng. Biotechnol.* 16, 693–701. doi: 10.1016/j.jgeb.2018.02.006

Gupta, R., and Brunak, S. (2002). Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac. Symp. Biocomput.* 7, 310–322.

Hu, B., Jin, J., Guo, A. Y., Zhang, H., Luo, J., and Gao, G. (2015). GSDS 2.0: An upgraded gene feature visualization server. *Bioinformatics* 31, 1296–1297. doi: 10.1093/bioinformatics/btu817

Hu, L., Li, H., Chen, L., Lou, Y., Amombo, E., and Fu, J. (2015). RNA-seq for gene identification and transcript profiling in relation to root growth of bermudagrass (*Cynodon dactylon*) under salinity stress. *BMC Genomics* 16:575. doi: 10.1186/s12864-015-1799-3

Hu, Q., Min, L., Yang, X., Jin, S., Zhang, L., Li, Y., et al. (2018). Laccase Ghlac1 modulates broad-spectrum biotic stress tolerance via manipulating phenylpropanoid pathway and jasmonic acid synthesis. *Plant Physiol.* 176, 1808–1823. doi: 10.1104/pp.17.01628

Hua, S., Zhang, B., Fu, Y., Qi, B., Li, Y., Tian, F., et al. (2018). Enzymatic gene expression by *Pleurotus tuoliensis* (Bailinggu): Differential regulation under low temperature induction conditions. *World J. Microbiol. Biotechnol.* 34:160.

Inagaki, F., Hinrichs, K. U., Kubo, Y., Bowles, M. W., Heuer, V. B., Hong, W. L., et al. (2015). Exploring deep microbial life in coal-bearing sediment down to ˜2.5 km below the ocean floor. *Science* 349, 420–424. doi: 10.1126/science.aaa6882

Jacob, A. G., and Smith, C. W. J. (2017). Intron retention as a component of regulated gene expression programs. *Hum. Genet.* 136, 1043-1057. doi: 10.1007/s00439-017-1791-x

Janusz, G., Pawlik, A., Świderska-Burek, U., Polak, J., Sulej, J., Jarosz-Wilkołazka, A., et al. (2020). Laccase properties, physiological functions, and evolution. *Int. J. Mol. Sci.* 21:966.

Jiao, X., Li, G., Wang, Y., Nie, F., Cheng, X., Abdullah, M., et al. (2018). Systematic analysis of the *Pleurotus ostreatus* laccase Gene (PoLac) family and functional characterization of PoLac2 involved in the degradation of cotton-straw lignin. *Molecules* 23:880.

Kirtzel, J., Scherwietes, E. L., Merten, D., Krause, K., and Kothe, E. (2019). Metal release and sequestration from black slate mediated by a laccase of *Schizophyllum commune*. *Environ. Sci. Pollut. Res. Int.* 26, 5–13. doi: 10.1007/s11356-018-2568-z

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315

Kudanga, T., Nyanhongo, G. S., Guebitz, G. M., and Burton, S. (2011). Potential applications of laccase-mediated coupling and grafting reactions: A review. *Enzyme Microb. Technol.* 48, 195–208. doi: 10.1016/j.enzmictec.2010.11.007

Kumar, B., Bhardwaj, N., Alam, A., Agrawal, K., Prasad, H., and Verma, P. (2018). Production, purification and characterization of an acid/alkali and thermo tolerant cellulase from *Schizophyllum commune* NAIMCC-F-03379 and its application in hydrolysis of lignocellulosic wastes. *AMB Express* 8:173.

Kumar, S. V., Phale, P. S., Durani, S., and Wangikar, P. P. (2003). Combined sequence and structure analysis of the fungal laccase family. *Biotechnol. Bioeng.* 83, 386–394. doi: 10.1002/bit.10681

Kumar, S. V., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Kumar, V. P., Naik, C., and Sridhar, M. (2015). Production, purification and characterization of novel laccase produced by *Schizophyllum commune* NI-07 with potential for delignification of crop residues. *Appl. Biochem. Microbiol.* 51, 432–441. doi: 10.1134/S0003683815040080

Kumar, V. P., Naik, C., and Sridhar, M. (2018). Morphological and phylogenetic identification of a hyper laccase producing strain of *Schizophyllum commune* NI-07 exhibiting delignification potential. *Indian J. Biotechnol.* 17, 302–315.

Leonowicz, A., Cho, N. S., Luterek, J., Wilkolazka, A., Wojtas-Wasilewska, M., Matuszewska, A., et al. (2001). Fungal laccase: Properties and activity on lignin. *J. Basic Microbiol.* 41, 185–227. doi: 10.1002/1521-4028(200107)41:3/4<185::AID-JOBM185>3.0.CO;2-T

Lettera, V., Piscitelli, A., Leo, G., Birolo, L., Pezzella, C., and Sannia, G. (2010). Identification of a new member of *Pleurotus ostreatus* laccase family from mature fruiting body. *Fungal Biol.* 114, 724–730. doi: 10.1016/j.funbio.2010.06.004

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44, W242–W245. doi: 10.1093/nar/gkw290

Liu, C. H., Huang, X., Xie, T. N., Duan, N., Xue, Y. R., Zhao, T. X., et al. (2017). Exploration of cultivable fungal communities in deep coal-bearing sediments from ∼1.3 to 2.5 km below the ocean floor. *Environ. Microbiol.* 19, 803–818. doi: 10.1111/1462-2920.13653

Liu, N., Cao, Z. Y., Cao, K. K., Ma, S. X., Gong, X. D., Jia, H., et al. (2019). Identification of laccase-like multicopper oxidases from the pathogenic fungus *Setosphaeria turcica* and their expression pattern during growth and infection. *Eur. J. Plant Pathol.* 153, 1149–1163. doi: 10.1007/s10658-018-01632-8

Liu, X., Huang, X., Chu, C., Xu, H., Wang, L., Xue, Y. R., et al. (2022). Genome, genetic evolution, and environmental adaptation mechanisms of *Schizophyllum commune* in deep subseafloor coal-bearing sediments. *iScience* 25:104417.

Mayer, A. M., and Staples, R. C. (2002). Laccase: New functions for an old enzyme. *Phytochemistry* 60, 551–565. doi: 10.1016/S0031-9422(02)00171-1

Mehmood, T., Saman, T., Irfan, M., Anwar, F., Ikram, M. S., and Tabassam, Q. (2019). Pectinase production from *Schizophyllum commune* through central composite design using citrus waste and its immobilization for industrial exploitation. *Waste Biomass Valori.* 10, 2527–2536. doi: 10.1007/s12649-018-0279-9

Monteiro, P. T., Oliveira, J., Pais, P., Antunes, M., Palma, M., Cavalheiro, M., et al. (2020). YEASTRACT+: A portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic Acids Res.* 48, D642–D649. doi: 10.1093/nar/gkz859

Mtibaà, R., Barriuso, J., de Eugenio, L., Aranda, E., Belbahri, L., Nasri, M., et al. (2018). Purification and characterization of a fungal laccase from the ascomycete Thielavia sp. and its role in the decolorization of a recalcitrant dye. *Int. J. Biol. Macromol.* 120, 1744–1751. doi: 10.1016/j.ijbiomac.2018.09.175

Naro, C., and Sette, C. (2017). Timely-regulated intron retention as device to fine-tune protein expression. *Cell Cycle* 16, 1321–1322. doi: 10.1080/15384101.2017.1337983

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10, 1–6. doi: 10.1093/protein/10.1.1

Ohm, R. A., de Jong, J. F., Lugones, L. G., Aerts, A., Kothe, E., Stajich, J. E., et al. (2010). Genome sequence of the model mushroom *Schizophyllum commune*. *Nat. Biotechnol.* 28, 957–963. doi: 10.1038/nbt.1643

Peng, C., and Gao, F. (2014). Protein localization analysis of essential genes in prokaryotes. *Sci. Rep.* 4:6001.

Qi, Y. B., Wang, X. L., Shi, T., Liu, S., Xu, Z. H., Li, X., et al. (2015). Multicomponent kinetic analysis and theoretical studies on the phenolic intermediates in the oxidation of eugenol and isoeugenol catalyzed by laccase. *Phys. Chem. Chem. Phys.* 17, 29597–29607. doi: 10.1039/C5CP03475B

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 27, 2325–2329. doi: 10.1093/bioinformatics/btr355

Rodríguez, C. S., and Toca Herrera, J. L. (2006). Industrial and biotechnological applications of laccases: A review. *Biotechnol. Adv.* 24, 500–513. doi: 10.1016/j.biotechadv.2006.04.003

Sakamoto, Y., Nakade, K., Yoshida, K., Natsume, S., Miyazaki, K., Sato, S., et al. (2015). Grouping of multicopper oxidases in *Lentinula edodes* by sequence similarities and expression patterns. *AMB Express* 5:63.

Shiba, T., Xiao, L., Miyakoshi, T., and Chen, C. L. (2000). Oxidation of isoeugenol and coniferyl alcohol catalyzed by laccases isolated from *Rhus vernicifera* Stokes and *Pycnoporus coccineus*. *J. Mol. Catal. B Enzym.* 10, 605–615. doi: 10.1016/S1381-1177(00)00184-3

Shleev, S., Christenson, A., Serezhenkov, V., Burbaev, D., Yaropolov, A., Gorton, L., et al. (2005). Electrochemical redox transformations of T1 and T2 copper sites in native *Trametes hirsuta* laccase at gold electrode. *Biochem. J.* 385, 745–754. doi: 10.1042/BJ20041015

Singh, A. D., and Sharma, R. K. (2010). Ligninolytic fungal laccases and their biotechnological applications. *Appl. Biochem. Biotechnol.* 160, 1760–1788. doi: 10.1007/s12010-009-8676-y

Solomon, E. I., Sundaram, U. M., and Machonkin, T. E. (1996). Multicopper oxidases and oxygenases. *Chem. Rev.* 96, 2563–2606. doi: 10.1021/cr950046o

Sornlake, W., Rattanaphanjak, P., Champreda, V., Eurwilaichitr, L., Kittisenachai, S., Roytrakul, S., et al. (2017). Characterization of cellulolytic enzyme system of *Schizophyllum commune* mutant and evaluation of its efficiency on biomass hydrolysis. *Biosci. Biotechnol. Biochem.* 81, 1289–1299. doi: 10.1080/09168451.2017.1320937

Stival Sena, J., Gigue're, I., Boyle, B., Rigault, P., Birol, I., Zuccolo, A., et al. (2014). Evolution of gene structure in the conifer *Picea glauca*: A comparative analysis of the impact of intron size. *BMC Plant Biol.* 14:95. doi: 10.1186/1471-2229-14-95

Tovar-Herrera, O. E., Martha-Paz, A. M., Pérez, L. Y., Aranda, E., Tacoronte-Morales, J. E., Pedroso-Cabrera, M. T., et al. (2018). Schizophyllum commune: An unexploited source for lignocellulose degrading enzymes. *Microbiologyopen* 7:e00637.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Vasina, D. V., Mustafaev, O. N., Moiseenko, K. V., Sadovskaya, N. S., Glazunova, O. A., Tyurin, À, et al. (2015). The *Trametes hirsuta* 072 laccase multigene family: Genes identification and transcriptional analysis under copper ions induction. *Biochimie* 116, 154–164. doi: 10.1016/j.biochi.2015.07.015

Voorrips, R. E. (2002). MapChart: Software for the graphical presentation of linkage maps and QTLs. *J. Hered.* 93, 77–78. doi: 10.1093/jhered/93.1.77

Wang, Q., Li, G., Zheng, K., Zhu, X., Ma, J., Wang, D., et al. (2019). The Soybean laccase gene family: Evolution and possible roles in plant defense and stem strength selection. *Genes* 10:701.

Wang, W., Liu, F., Jiang, Y., Wu, G., Guo, L., Chen, R., et al. (2015). The multigene family of fungal laccases and their expression in the white rot basidiomycete *Flammulina velutipes*. *Gene* 563, 142–149. doi: 10.1016/j.gene.2015.03.020

Wilkins, M. R., Gasteiger, E., Bairoch, A., Sanchez, J. C., Williams, K. L., Appel, R. D., et al. (1999). Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* 112, 531–552.

Xie, X. Y., Yan, C. N., Wu, D. Q., and Qu, S. S. (1999). Reactivity properties of laccase with various types of substrates. *J. Cent. China Norm. Univ.* 2, 246–249.

Xu, X., Zhou, Y., Wang, B., Ding, L., Wang, Y., Luo, L., et al. (2019). Genome-wide identification and characterization of laccase gene family in *Citrus sinensis*. *Genes* 689, 114–123.

Yan, L., Xu, R., Bian, Y., Li, H., and Zhou, Y. (2019). Expression profile of laccase gene family in white-rot basidiomycete *Lentinula edodes* under different environmental stresses. *Genes* 10:1045.

Yang, Y., Sossah, F. L., Li, Z., Hyde, K. D., Li, D., Xiao, S., et al. (2021). Genome-wide identification and analysis of chitinase GH18 gene family in *Mycogone perniciosa*. *Front. Microbiol.* 11:596719. doi: 10.3389/fmicb.2020.596719

Yoshida, H. (1883). Chemistry of lacquer (urushi). *J. Chem. Soc.* 43, 472–486. doi: 10.1039/CT8834300472

Yu, C. S., Chen, Y. C., Lu, C. H., and Hwang, J. K. (2006). Prediction of protein subcellular localization. *Proteins* 64, 643–651. doi: 10.1002/prot.21018

Zain Ul Arifeen, M., Chu, C., Yang, X. Y., Liu, J. Z., Huang, X., Ma, Y. N., et al. (2021). The anaerobic survival mechanism of *Schizophyllum commune* 20R-7-F01, isolated from deep sediment 2 km below the seafloor. *Environ. Microbiol.* 23, 1174–1185. doi: 10.1111/1462-2920.15332

Zain Ul Arifeen, M., Ma, Y. N., Wu, T. S., Chu, C., Liu, X., Jiang, J. P., et al. (2022). Anaerobic biodegradation of polycyclic aromatic hydrocarbons (PAHs) by fungi isolated from anaerobic coal-associated sediments at 2.5 km below the seafloor. *Chemosphere* 9:135062.

Zain Ul Arifeen, M., Yang, X. Y., Li, F. F., Xue, Y. R., Gong, P. X., and Liu, C. H. (2020). Growth behaviors of deep subseafloor *Schizophyllum commune* in response to various environmental conditions. *Acta Microbiol. Sin.* 60, 1882–1892.

Zhang, G. W., Song, H. D., and Chen, Z. (2004). Molecular mechanism of mRNA alternative splicing. *J. Genet. Genomics* 31, 102–107.

Zhang, J., Chen, H., Chen, M., Ren, A., Huang, J., Wang, H., et al. (2015). Cloning and functional analysis of a laccase gene during fruiting body formation in *Hypsizygus marmoreus*. *Microbiol. Res.* 179, 54–63. doi: 10.1016/j.micres.2015.06.005

Zhang, Y., Lv, Z., Zhou, J., Xin, F., Ma, J., Wu, H., et al. (2018). Application of eukaryotic and prokaryotic laccases in biosensor and biofuel cells: Recent advances and electrochemical aspects. *Appl. Microbiol. Biotechnol.* 102, 10409–10423. doi: 10.1007/s00253-018-9421-7

Zhao, W., Qin, T., Zhang, Q., and Xu, S. (2018). Research on laccase from *Schizophyllum commune*: Its enzymatic properties and application to advanced treatment of papermaking wastewater. *Environ. Sci. Technol.* 41, 67–72.

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 08 August 2022
DOI 10.3389/fmicb.2022.920780

OPEN ACCESS

EDITED BY
Zheng Zhang,
Shandong University, China

REVIEWED BY
Jiyuan Zhou,
Guangzhou Medical University, China
Alexander N. Ignatov,
Peoples' Friendship University of Russia, Russia

*CORRESPONDENCE
Huilin Sun
sun-hui-lin@126.com
Chao Liu
liuchaogzf@163.com
Ling Chen
lingpzy@163.com

†These authors have contributed equally to this work

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

RECEIVED 15 April 2022
ACCEPTED 01 July 2022
PUBLISHED 08 August 2022

CITATION
Huang L, Deng L, Liu C, Huang E, Han X, Xiao C, Liang X, Sun H, Liu C and Chen L (2022) Fecal microbial signatures of healthy Han individuals from three bio-geographical zones in Guangdong.
Front. Microbiol. 13:920780.
doi: 10.3389/fmicb.2022.920780

Fecal microbial signatures of healthy Han individuals from three bio-geographical zones in Guangdong

Litao Huang[1†], Liting Deng[2†], Changhui Liu[3†], Enping Huang[1], Xiaolong Han[3], Cheng Xiao[1], Xiaomin Liang[1], Huilin Sun[2*], Chao Liu[3*] and Ling Chen[1*]

[1]Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou, China, [2]The First Affiliated Hospital of Guangdong Pharmaceutical University, Guangzhou, China, [3]Guangzhou Forensic Science Institute, Guangzhou, China

Important forensic evidence traced from crime scenes, such as fecal materials, can help in the forensic investigation of criminal cases. Intestines are the largest microbial pool in the human body whose microbial community is considered to be the human "second fingerprint". The present study explored the potential for community characteristics of gut microbes in forensic medicine. Fecal microbiota profiles of healthy individuals from three representative Han populations (Guangzhou, Shantou and Meizhou) in Guangdong Province, China were evaluated using High-throughput sequencing of V3-V4 hypervariable regions of the 16SrRNA gene. Results of the present study showed that at the genus level, Shantou, Guangzhou, and Meizhou behaved as Enterotype1, Enterotype2, and Enterotype3, which were mainly composed of Bacteroides, Prevotella, and Blautia, respectively. Based on OTU abundance at the genus level, using the random forest prediction model, it was found that there might be potential for distinguishing individuals of Guangzhou, Meizhou, and Shantou according to their fecal microbial community. Moreover, the findings of the microbial community of fecal samples in the present study were significantly different from that of saliva samples reported in our previous study, and thus it is evident that the saliva and feces can be distinguished. In conclusion, this study reported the fecal microbial signature of three Han populations, which may provide basic data for the potential application in forensic practice, containing body fluid identification, and geographical inference.

KEYWORDS

forensic medicine, feces, gut microbiome, 16S rRNA gene sequencing, Guangdong Han individuals

Introduction

Human beings live in a world full of microbes, and co-evolution, co-adaptation as well as co-dependence are the relationships between them and indigenous microbiota (Turnbaugh et al., 2007; Blaser and Falkow, 2009). Microorganisms exist in many sites of the human body mainly in the intestines. They also have

a profound influence on human physiological metabolism and nutrition regulation (Hooper and Gordon, 2001). The human intestinal tract, a nutrient-rich microenvironment, carries 100 trillion (10^{14}) bacteria which are about 10 times more than the number of human cells (Hooper et al., 2002; Bäckhed et al., 2005). The colon is the main contributor to the total number of bacteria in the entire intestine with a density close to 10^{11}-10^{12}cells/ml (Ley et al., 2006; Sender et al., 2016). Non-invasive fecal samples (the part in the middle of feces that is not in contact with the air and the ground) are usually considered representative of colonic microorganisms for they are easy to obtain and do not harm the subject (Davenport et al., 2017). Next-generation sequencing (NGS) using 16SrRNA gene sequence analysis overcomes the shortcomings that most traditional microorganisms cannot be cultivated and performed much deeper microbial community analysis at a low cost (Weinstock, 2012).

Feces have been specified as the important evidence for specific crimes, including burglary, robbery, and sexual cases. Particularly, in anal sexual assault cases, fecal traces of the victim may be left on a condom at the crime scene (Johnson et al., 2005). When analyzed, the microbial community in feces may help in individual identification and tracing the source of tissues and body fluids. Quaak et al. distinguished individuals by researching the microbial profiles generated in fecal samples from 35 healthy volunteers of different ages. It was then proposed that individual identification can be carried out by applying the fecal microbial profile to the increase evidence value of the trace when there was no or only part of human STR in fecal samples (Quaak et al., 2017). Microarray was also performed to analyze 175 samples from healthy individuals, successfully distinguishing and identifying the oral cavity, feces, and skin samples. The study noted that it might be beneficial for presenting important corroborating evidence for the scene left by the victim and/or suspect, aiding in the reconstruction of a case process (Quaak et al., 2018).

Recent studies have shown that the human intestinal microbial community is not only affected by the host's own factors, but also by external factors (Wen and Duffy, 2017). In general, geography and environment have shown the main influence on intestinal microbes (He et al., 2018; Rothschild et al., 2018). Guangdong Province is located in the southernmost part of mainland China and is an important heritage site of Lingnan culture. Lingnan Han groups, consisting of the Guangfu, Hakka, and Chaoshan, account for a majority of Han people in Guangdong. They have a unique culture in terms of language, customs, and living habits. For instance, Guangfu people speak Cantonese, cook Cantonese cuisine, and live mainly in the Pearl River Delta area of Guangdong. Further, Hakkas people are concentrated in northern Guangdong, mainly Hakka dialect and Hakka cuisine, together with Chaoshan people living in eastern Guangdong have their own Chaoshan dialects and Chaoshan cuisine (Wang et al., 2010; Du et al., 2019). Guangdong's three Han characteristic population was recognized as a branch of Han Chinese, and the gut microbiome characterization and forensic potential of these three groups are poorly defined or still need to be explored. The current study aimed to reveal the differences in fecal microbiota between the groups. Indigenous Han individuals from Guangzhou, Meizhou, and Shantou were selected as the representative of Guangfu, Hakka, and Chaoshan individuals, respectively. The fecal samples were collected and characterized through high-throughput sequencing of the samples in the V3-V4 region of the 16SrRNA gene. The prospect of forensic application of fecal microbiota was valued.

Materials and methods

Sample collection

This study was approved by the Biomedical Ethics Committee of Southern Medical University, Guangzhou, China. After obtaining informed consent, a total of 59 fecal samples were collected from healthy Han individuals (aged between 16 and 62) who had lived in Guangzhou, Meizhou, and Shantou for more than three generations in Guangdong Province, China. A total of 19, 20, and 20 samples from people in Guangzhou, Meizhou, and Shantou, respectively were collected. Participants were balanced by age and sex, divided into age1 (16–32 years old) and age2 (33–62 years old) groups, male and female groups. The participants received adequate training and guidance on the sample collection process before fecal collection and one sample was then collected per participant. The exclusion criteria were (1) participants who reported antibiotic use/other treatments within 3 months. (2) participants were diagnosed with any inflammation-related bowel disease or gastrointestinal disease within 3 months. (3) participants who lived <1 year or left the province within 1 month. According to the above criteria, a total of 59 healthy individuals from the three regions were included, and all fecal samples collected were named "F" (Guangzhou sample numbered from 1 to 19, Shantou sample numbers were from 20 to 39, Meizhou sample numbers were from 40 to 59). The participants used a sterile spoon to dig out a fallen scoop (about 3–5g) of fecal samples, collected them in a sterile plastic container, and immediately stored them in a refrigerator at-80°C in the laboratory awaiting extraction of the fecal bacterial genomic DNA.

DNA extraction, PCR amplification, and sequencing

Bacterial genomic DNA in the samples was extracted using QIAamp DNA Stool Mini Kit (QIAGEN, Hilden, Germany), according to the manufacturer's instructions.

The concentration and purity of DNA were quantified by using an ultraviolet spectrophotometer and DNA extraction quality is checked by 1% agarose gel electrophoresis. Qualified DNA samples were amplified using bacterial 16S rRNA corresponding DNA sequence V3-V4 region universal primers 338F (5′- ACTCCTACGGGAGGCAGCA−3′) and 806R (5′-GGACTACHVGGGTWTCTAAT−3′) which contained a unique sequence tag to barcode each sample. PCR enrichment was performed in a 25μl reaction containing 12.5μl of 2×Q5 Master Mix, 0.2μM of each primer,120ng of the extracted DNA, and Nuclease-free water. PCR reaction amplification conditions were: initial denaturation at 98°C for 5 mins; followed by 15-21 cycles of denaturation at 98°C for 10 s, primer annealing at 57°C for 30 s, extension at 72°C for 30 s; and a final extension step at 72°C for 5 min. The PCR products were purified with AmpureXP beads and eluted in the Elution buffer. Libraries were built with NEB Next UltraTM DNA Library Prep Kit for Illumina (New England Biolabs Inc, Ipswich, USA). And then the validated libraries were used for sequencing on the Illumina MiSeq platform (Illumina Corporation, San Diego, USA). The sequencing data have been deposited in NCBI BioProject PRJNA824624 with the Biosample accessions SAMN27409411-SAMN27409469.

Bioinformatics analysis

The raw reads obtained by sequencing are filtered to obtain high-quality data (clean reads) for downstream analysis. Using the software FLASH (Magoč and Salzberg, 2011) (Fast Length Adjustment of Short reads,v1.2.11), the paired reads obtained by double terminal sequencing are assembled into a sequence, that is, a tag, by using the overlapping relationship. Use CUTADAPT (Martin, 2011) to remove tags containing primers, refer to the gold database (v20110519) chimera database, and use the UCHIME method in the VSEARCH (v2.3.4) (Rognes et al., 2016) software to remove the tags containing the chimera. Use VSEARCH (v2.3.4) software to cluster Tags with a similarity> 97% into an OTU, and get the OTU representative sequence. Use RDP classifier (v2.2) (Wang et al., 2007) software to compare OTU representative sequence with Silva(v128) database for species annotation. Alpha diversity is used to analyze the species diversity in the sample, using mothur (v1.39.5) (Schloss et al., 2009) software to calculate 5 indicators, including Chao, Ace, Shannon, and Simpson. Beta diversity is used to measure the diversity between samples, calculated using QIIME (v1.80) (Caporaso et al., 2010) software. The rest of the graphics are implemented using R package (v3.0.3). Use LEfSe (LDA Effect Size) (v1.0) (Segata et al., 2011) to calculate the LDA score value. The significant flora must meet the threshold $p < 0.05$ and the LDA score value ≥2.0 (or ≤-2.0). Through the use of QIIME (v1.80) (Schloss et al., 2009) software, the use of similarity analysis (ANOSIM) for group comparison analysis, to find out the different components in the group.

Machine learning process

Random forest analysis was used to perform classification. This method constructed multiple decision trees by using the information contained in input features and predicted the classification of three regions by combining multiple weak classifiers (Breiman, 2001). According to the random forest method in the R package RandomForest (v4.6-14), the OTU data of intestinal microorganisms in the three regions was used to build a model for predicting the sample distribution in the areas. The RF classification method was divided into two steps: one was to build a decision tree based on randomly selected samples (the training set) which include 70% of the original data set (42 samples). The other one was to use the test set which was the remaining samples (17 samples) in the original data set to verify the decision tree (Svetnik et al., 2003). In addition, the receiver operating characteristic (ROC) curve was used to evaluate the constructed model, and the area under the ROC curve (AUC) was used to designate the ROC effect to evaluate the potential of intestinal microbial markers to predict different regions.

Results

Correlation with age and sex of the subjects

The present study explored the relationship between the composition of the gut microbial community and age as well as sex in the entire population. The results of ANOSIM analysis of the present study based on Bray–Curtis distance showed that there was no significant difference in the gut microbial community between age 1 and age 2 group ($p = 0.49$), and the male and female group ($p = 0.30$).

Whole sequencing data

Fecal samples of 59 healthy individuals from Guangzhou, Meizhou, and Shantou, Guangdong Province were subjected to high-throughput sequencing of 16SrRNA gene. After filtering, a data set consisting of 4256.44Mbp of effective and high-quality 16SrRNA gene sequences were generated, including 16,740,484 reads (median=221,912 reads, ranging from 79,892 to 599,496 reads; Supplementary Table 1). A cluster analysis of 97% similarity was performed to determine a total of 3,419 OTUs. All the valid sequences were annotated with species at different taxonomic levels, which yielded a total of 3,419 OTUs, belonging to 13 phyla, 15 classes, 21 orders, 35 families, 119

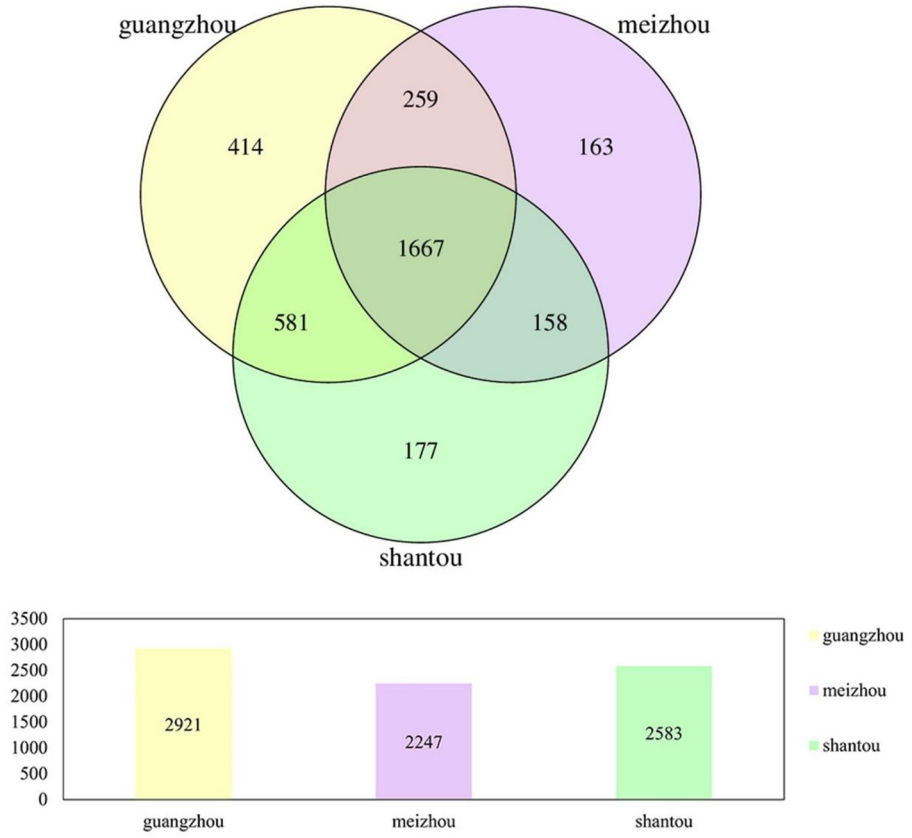

FIGURE 1
Venn diagrams of bacterial OTUs in all fecal samples from people in Guangzhou, Meizhou and Shantou.

genera, and 22 species. The Venn diagram showed that the number of unique OTUs in Guangzhou, Meizhou, and Shantou was 414, 163, and 177, respectively, with 1667 OTUs shared by all the samples in the present study (Figure 1).

Richness and diversity of microbial communities

Microbial complexity in the feces was estimated based on alpha-diversity indices (Chao, Ace, Simpson, and Shannon), and the results showed that there was no significant difference in the diversity among all individuals in each group (Figures 2E–H). Pairwise diversity of the three groups in the present study, the indices of Chao and Ace represented the species richness. The results of the present study showed that individuals from Guangzhou and Shantou had significantly higher index values as compared with those from Meizhou (Figures 2A,B). Results of the Simpson diversity index in the current study revealed that the three regions had similar statistical index values, indicating no significant difference in species diversity (Figure 2C; $p >$ 0.05). In addition, the sparse curve of the Shannon index showed a trend toward saturation as presented in Figure 2D which illustrated sufficient sequencing depth.

Overview of bacterial community composition

The average relative abundance of the three groups at the phylum and genus level was also evaluated to further intuitively uncover the microbial composition characteristics in the three regional groups as presented in Figure 3. It was found that phylum Firmicutes was the most predominant phyla in Guangzhou, Meizhou, and Shantou, with relative abundances of 46.7, 43.4, and 62.5%, respectively. This was followed by phylum Bacteroidetes, which contributed 43.1, 38.2, and 16.1% of the total sequences. Further, it was noted that Bacteroides had the highest abundance in the bacterial communities of fecal samples at the genus level, accounting for 28.7, 31.7, and 12.7% in Guangzhou, Meizhou, and Shantou, respectively. On the other hand, Faecalibacterium

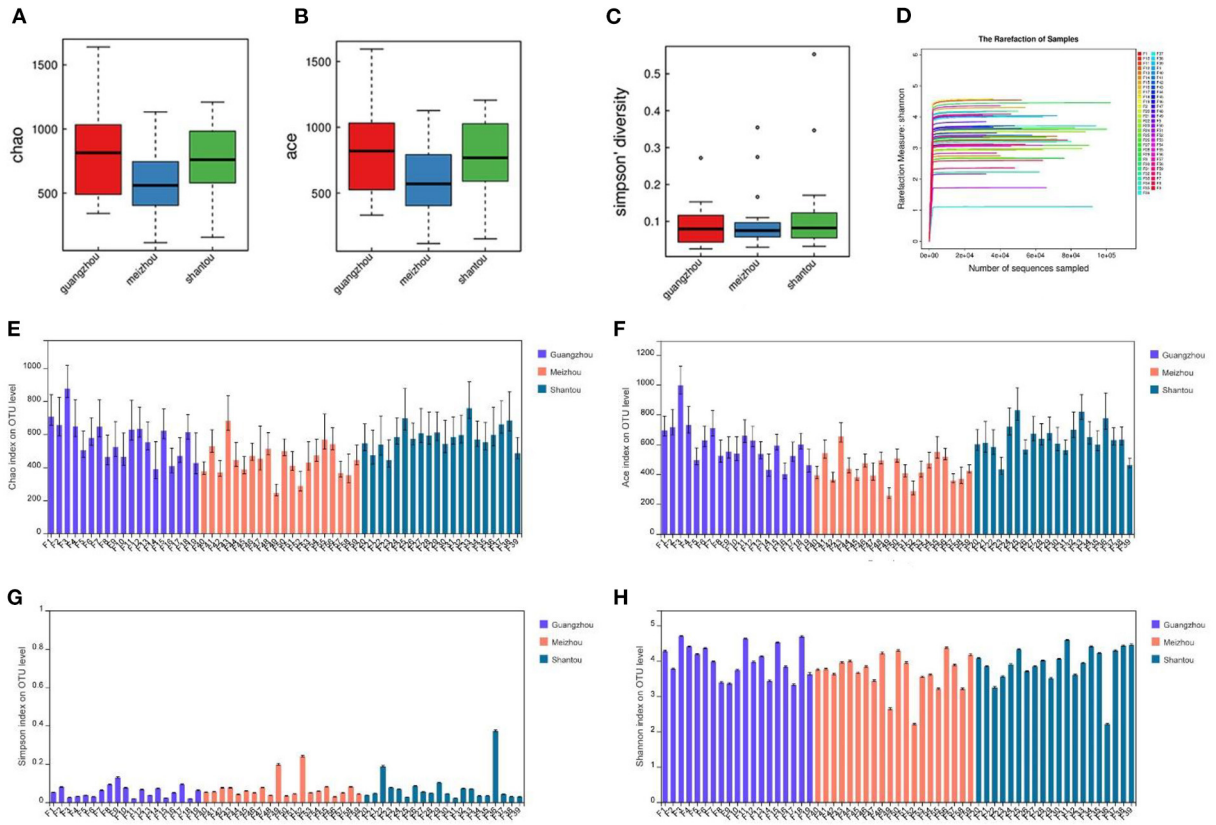

FIGURE 2
Differences in bacterial alpha diversity among the three regions: **(A,E)** Chao. **(B,F)** Ace. **(C,G)** Simpson diversity. **(D,H)** Shannon index.

accounted for 7.4, 7.9, and 8.9% in Guangzhou, Meizhou, and Shantou, respectively. The remaining top 10 bacterial genera were Blautia, Eubacterium_rectale_group, Bifidobacterium, Roseburia, Prevotella_9, Megamonas, Escherichia-Shigella, and Fusobacterium. Besides, it was found that the relative abundance of Bifidobacterium was 1.54%, 1.04%, and 5.09% in Guangzhou, Meizhou, and Shantou, respectively.

Genus-level core intestinal flora and comparison of feces and saliva

The intestinal core microbiome was determined at the genus level and defined as bacteria with >0.1% abundance in ≥80% of the respective samples (Dehingia et al., 2015). It was found that there were six main genera in the fecal samples of all individuals, which constituted a genera-level phylogenetic core, including Bacteroides, Blautia, Eubacterium_hallii_group, Faecalibacterium, Lachnoclostridium, and Roseburia (Supplementary Table 2). Further, these fecal samples were used to compare with saliva samples we previously published (Yao et al., 2021) and the results of the comparisons were as shown in Supplementary Figure 1. The data of the present study on principal coordinate analysis (PCoA) based on genus-level abundance revealed that there was a clear distinction between fecal samples and saliva samples. Further, the linear discriminant analysis (LDA) histogram reflected that at the genus level, the relative abundance of Bacteroides, Faecalibacterium, Blautia, and Bifidobacterium was higher in the fecal samples, whereas the relative abundance of Streptococcus, Gemella, Porphyromonas, and Haemophilus was higher in the saliva samples.

Beta diversity of bacterial communities

Beta diversity was assessed by PCoA and ANOSIM analysis using the Bray–Curtis distance method at the operational classification unit (OTU) level to further indicate the similarity between microbial communities. Although there were some slight overlaps in individual samples, the samples of Guangzhou and Meizhou groups, Shantou and Meizhou groups were roughly clustered. The similar structure of the intestinal microbiota community was found in the fecal samples between Guangzhou and Shantou, indicating

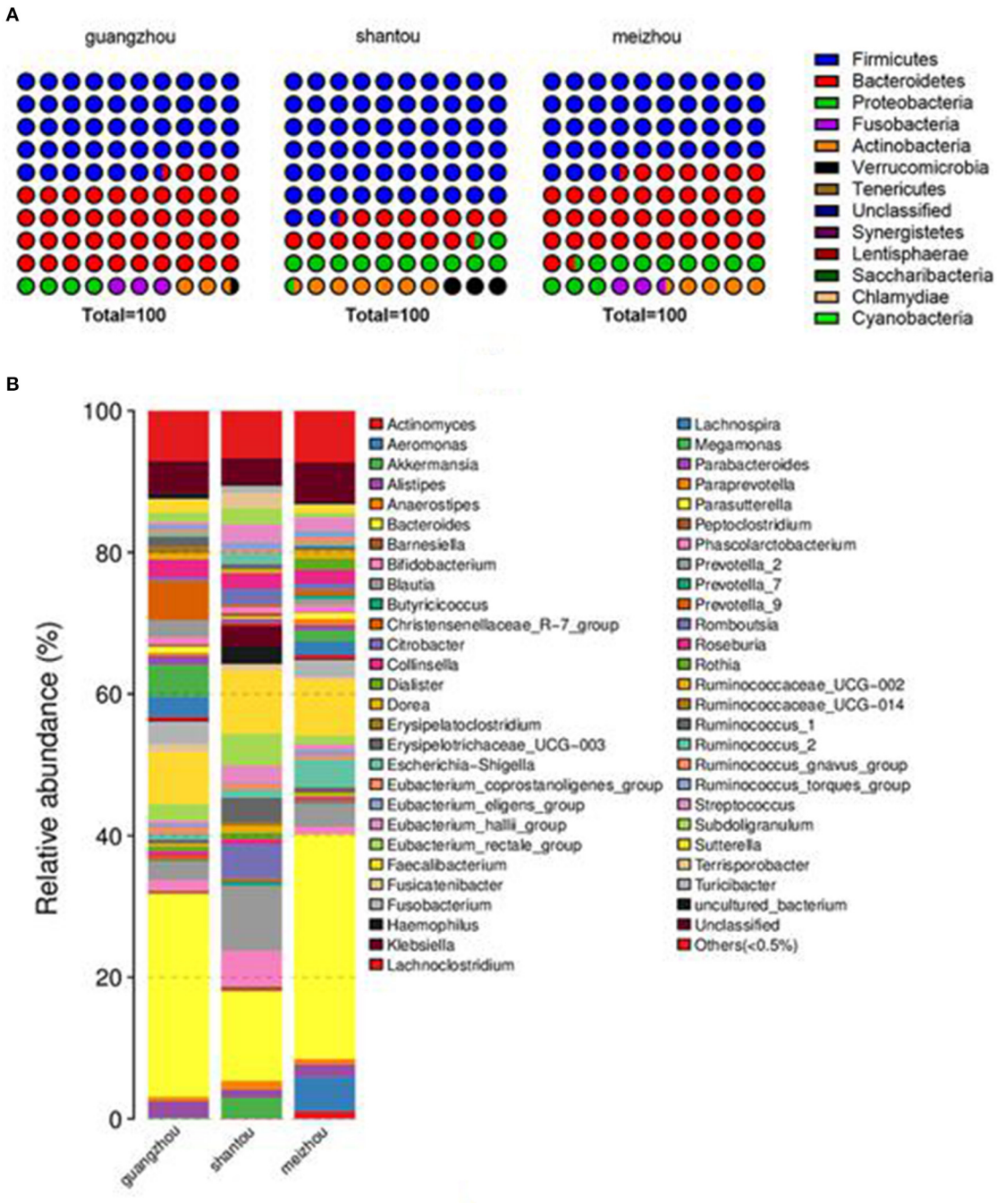

FIGURE 3
Distribution of intestinal microbes at different taxonomic levels in Guangzhou, Meizhou and Shantou populations. Two levels of dominant taxa are shown (Others: <0.5% relative abundance). **(A)** Distribution at the phylum level. **(B)** Distribution at the genus level.

an overlap in community structure (Figures 4A–C). The samples of the Meizhou population formed an "out-group," which was generally not confounding with the samples of the Guangzhou or Shantou populations (Figure 4D). The ANOSIM analysis was performed on the three geographical groups (Supplementary Figure 2), and the results of this study demonstrated that the differences between the groups were greater than the differences within the groups, and the groupings were statistically significant ($R = 0.3254$, $p = 0.0010$).

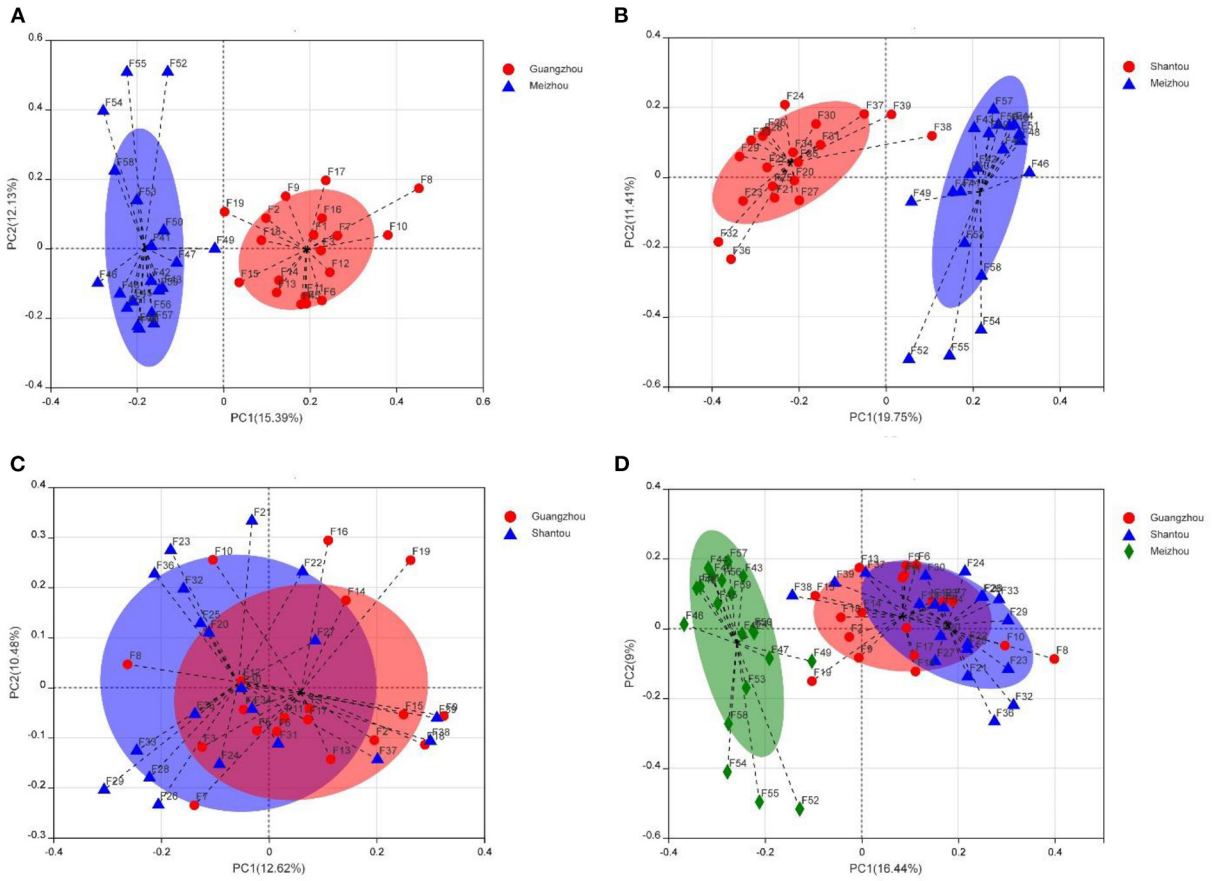

FIGURE 4
Taxonomic diversity of microbiomes from samples from Guangzhou, Meizhou, and Shantou. The principal coordinate analysis (PCoA) graph analysis is based on the Bray-Curtis distance at the operational classification unit (OTU) level, and each sample is represented by a point. **(A)** Guangzhou vs. Meizhou. **(B)** Shantou vs. Meizhou. **(C)** Guangzhou vs. Shantou. **(D)** Guangzhou vs. Shantou vs. Meizhou.

Comparison of differences among three regions

The linear discriminant analysis effect size (LEfSe) test for biomarkers was used to find the taxa with the strongest effect on region differentiation. The Cladogram chart showed that there were at least two significant differences in the phylum, class, order, family, genus, and species level in the fecal samples from Guangzhou and Meizhou (Figure 5A). The composition of the microbial community of the fecal samples from Shantou at the phylum level was not significantly different from that of Guangzhou and Meizhou. In addition, at least three significantly different microorganisms were found at the level of class, order, family, genus, and species levels. Further, a total of 96 differentially abundant taxa were found in the three regions shown in the histogram of LDA value distribution (Figure 5B). At the phylum level, the significant differences in the samples of the Guangzhou and Meizhou populations were mainly Bacteroidetes, and Firmicutes, respectively. The top five microorganisms with significant differences at the genus level in the three regions included Prevotella-9, Megamonas, Fusobacterium, Lachnospira, and Prevotella_2 in Guangzhou, Bacteroides, Actinomyces, Paraprevotella, Bulleidia, Bilophila in Shantou, and Blautia, Bifidobacterium, Erysipelotrichaceae_UCG_003, Klebsiella, Citrobacter in Meizhou.

Random forest

During the construction of a random forest model based on the composition of gut microbes, top 230 OTUs markers were set as the best set. The markers performed well and were on the training set (n = 42, 14 samples in Guangzhou, Shantou and Meizhou). The validation set of the random forest model (n = 17, 5 Guangzhou samples, 6 Meizhou samples, and 6 Shantou samples) showed that 12 of the 17 validation samples were correctly classified, and 100% of the Meizhou samples were

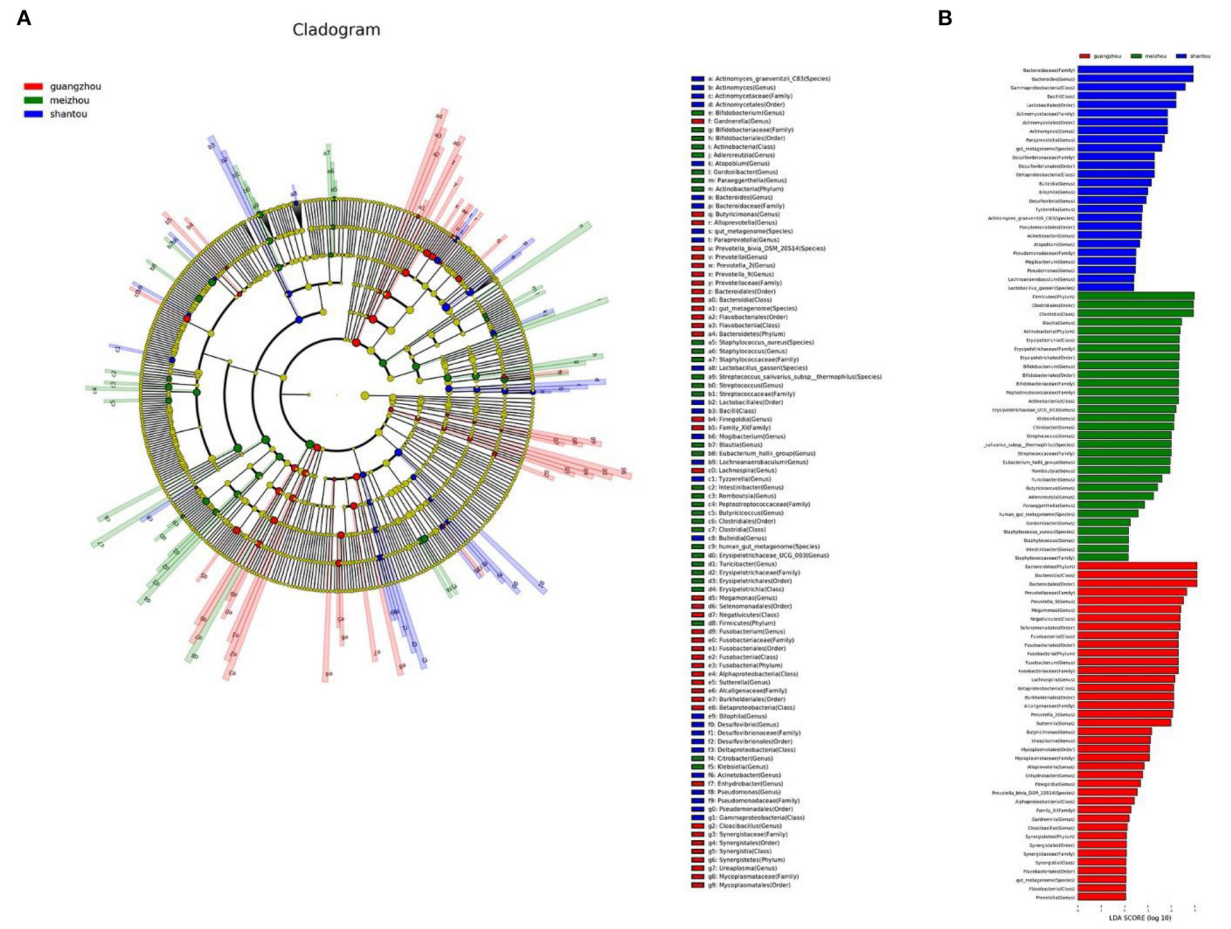

FIGURE 5
Differentially abundant taxa between the three regions. These different genera from phylum to genus were identified by linear discriminant analysis (LDA) using LEfSe. **(A)** Cladogram result graph. **(B)** Linear discriminant analysis (LDA) value distribution histogram. Red: Guangzhou; green: Meizhou; blue: Shantou.

correctly predicted, whereas 2 Guangzhou samples (F7 and F8) were identified as Shantou samples and 3 Shantou samples (F23, F25, and F34) were identified as Guangzhou samples, with an overall accuracy of 70.59%. The performance of the model was evaluated using ROC analysis. The AUC of the area under the curve in Guangzhou, Shantou, and Meizhou were 0.88, 0.73, and 1.00, respectively (Figure 6).

Discussions

The present study explored the correlation between the gut microbiota of the entire population and age as well as sex. Further, the ANOSIM analysis showed that there were no statistical difference between the intestinal microbial community structures between 16 and 32 as well as between 33 and 62 years of age. Previous studies had shown that Bifidobacterium was dominant in infants and a larger proportion of Bacteroides was dominant in elderly individuals (Claesson et al., 2011; Yatsunenko et al., 2012). On the other hand, Firmicutes and Bacteroidetes as the dominant bacteria were mainly dominant in adults. The established microbiota composition remained unchanged when there was no change in long-term eating habits and pathophysiology (Adak and Khan, 2019). In the current study, the small difference between the two age groups could be associated with most young individuals in the current study (45 cases, 76.27% between 25–45 years old), with only one individual who was over 60 years old. In addition, it was evident from the results of this study that there were no statistical differences in fecal microbiota between males and females. The finding of the present study was consistent with the results of a study carried out by Arumugam et al. that found that sex had no effect on the structure of the gut microbes of individuals from six different countries (Arumugam et al., 2011). Moreover, several other studies have also shown that sex factors have less influence on the gut microbial community than

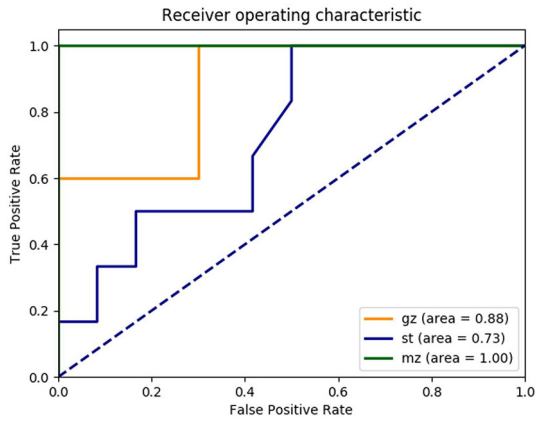

FIGURE 6
Receiver operating characteristic (ROC) curves trained on the OTU abundance demonstrate the performance of distinguishing fecal samples from Guangzhou, Shantou, and Meizhou. Yellow line: Guangzhou; blue line: Shantou; and green line: Meizhou.

other factors (Kovacs et al., 2011; Human Microbiome Project Consortium, 2012).

The analyses performed at the phylum level in the present study showed that the intestinal microbiota of this research was made up of the four most important phyla, including Firmicutes, Bacteroidetes, Proteobacteria, and Actinobacteria. It was evident that phylum Firmicutes and Bacteroides were the most abundant. This was similar to the results of previous studies (Jandhyala et al., 2015). Although the diversity of gut microbes at the phylum level was low, it was noted that they had significantly high diversity at the genus level. From the results of the current experiments, the predominant genera in all individuals was Bacteroides, followed by Faecalibacterium. A previous study reported that China had the highest abundance of Bacteroides at the genus level as compared with four other countries. This was consistent with the findings of the present study. Furthermore, the study reported that Japan had higher levels of Bifidobacterium whereas the abundance of Prevotella and Faecalibacterium was relatively higher in Korea (Nam et al., 2011). Previous studies had also indicated that Faecalibacterium was more dominant in the populations of Hadza, Italy, and the United States. Furthermore, Prevotella was a significant genus found among the Indian tribes, Mongolians, American Indians, and Malawi tribes (Dehingia et al., 2015). This difference in dominant genus originates from the variations in the intestinal microbiome, whereas the changes in the intestinal microbiome may be caused by geography and ethnicity among other factors (Dwiyanto et al., 2021).

One of the main interests of human gut microbial research was toward the core microbiota. The bacterial genera of Faecalibacterium, Eubacterium, Clostridium, Blautia, Ruminococcus, and Roseburia were found to be the core gut microbiota in the representative populations of the world (Dehingia et al., 2015). In the current study, six genera-level core intestinal bacteria of the gut microbiota, ubiquitously in unrelated individuals from Guangdong, which were Bacteroides, Blautia, Eubacterium_hallii_group, Faecalibacterium, Lachnoclostridium, and Roseburia. A microbial analysis report from nine provinces in China revealed a total of nine core bacteria (Balutia, Clostridium, Ruminocossus, Faecalibacterium, Subdoligranulum, Roseburia, Coproccus, Bacteroides, and Phascolarctobacterium) (Zhang et al., 2015). In healthy western individuals, Bifidobacterium, Bacteroides, Faecalibacterium, Ruminococcus, Blautia, Dorea, Eubacterium, and Coprococcus were the core intestinal bacteria genus (Martínez et al., 2013). Further, the intestinal core flora shared by these people were Bacteroides, Blautia, and Faecalibacterium. In addition, more than 45% of the common bacterial genera could be detected in both feces and oral cavities (Segata et al., 2012). It is worth mentioning that the establishment of the intestinal saliva microbial communities was similar. According to a study by Schmidt et al., transmission to, and subsequent colonization of the large intestine by oral microbes commonly occurred in healthy individuals. Although it has been previously reported that *Streptococcus salivarius* and *S. mutans* were particularly found in saliva (Tagg and Ragland, 1991). A study conducted by Kai-NanZou et al. showed that the bacteria in the intestines overlapped with those in feces (Zou et al., 2016). These results indicated that the identification of sample types using a single microbial marker may be misjudged. The findings of fecal samples in the present study were compared with those of saliva samples in our previously published study (Yao et al., 2021). In addition, the results showed that fecal and saliva samples can be distinguished, which could avoid the defect of single microbial markers to identify both saliva and feces samples.

The PCoA displayed regional differences in intestinal microorganisms between Meizhou and the other two regions. Different geographic origins of humans may result in diverse compositions of the gut microbiome, due to distinctive genetic backgrounds or life environments (Li and Zhao, 2015). Guangfu and Chaoshan populations occupied the two rich areas of the Pearl River delta plain and Chaoshan plain, respectively. The barren and backward mountainous areas of northern and eastern Guangdong were the basic distribution areas of the Hakka people. Several studies have demonstrated that geographic location plays an important role in shaping the intestinal microbial community, and dietary habits could also affect the composition and distribution of intestinal microbes (De Filippo et al., 2010; Zhang et al., 2013; Singh et al., 2017). Through a return visit to the volunteers in the three regions, they simply recorded their eating habits. The Meizhou area was dominated by greasy food, whereas the Guangzhou and Shantou areas were dominated by intake of a light diet (Song et al., 2005; Zhong et al., 2017; Wang et al., 2019).

A high-fat diet had been shown to reduce the diversity and richness of human gut microbial communities, which was negatively correlated with the abundance of Bifidobacterium. Furthermore, Caesar et al. reported that Bacteroides increased in mice fed with lard (De Filippo et al., 2010; Caesar et al., 2015; Khine et al., 2019). The results of the present study suggested that the intestinal microbes in Meizhou had the lowest abundance of Bifidobacterium and microbial alpha diversity, whereas Bacteroides showed the highest abundance among the three regions. This might be related to the fact that Hakka ancestors lived in mountainous areas with inconvenient transportation, expended much physical strength on their daily labor, and needed to supplement foods with rich fat sources such as lard, developing a diet that preferred greasy foods. Therefore, diet may also be an important factor affecting the microbial differences in fecal samples from the three geographical regions. The dietary associations seen here paralleled a recent study comparing European and African, Europeans consuming high-fat foods formed a typical taxonomy dominated by Bacteroidetes, while Africans consuming low-fat diets had higher microbial diversity (De Filippo et al., 2010). At the same time, a study of American populations showed that the gut flora of individuals with a typical western diet high in animal fat and protein was dominated by Bacteroides (Wu et al., 2011). There are, of course, many differences between the three regions that might influence the gut microbiome, but dietary differences provide an attractive potential explanation.

According to the results of the ANOSIM analysis, there were significant differences in the intestinal bacterial community composition in samples from the three regions. A previous study identified three intestinal types: Bacteroides (Enterotype 1), Prevotella (Enterotype 2), and Ruminococcus (Enterotype 3) (Arumugam et al., 2011), which could afford a strong discriminatory classification ability among European individuals, although other studies had reported that Enterotype 3 was an uncertain bacterial composition (Liang et al., 2017). Hyun Seok et al. showed that structure of gut microbiota variations across the geographical location. The characterization of population distribution according to the three enterotype classifications showed that the distributions of Enterotype 2 and Enterotype 1 differed by region. Samples from the U.S. and Japan had large numbers of Enterotype 1, while samples from Amazon natives in Venezuela, as well as from Malawi and Tanzania in Africa had large numbers of Enterotype 2 (Oh et al., 2022). In the present study, linear discriminant analysis (LDA) using LEfSe showed that Shantou, Guangzhou, and Meizhou belonged to Enterotype 1, Enterotype 2, and Enterotype 3, which were mainly composed of Bacteroides, Prevotella, and Blautia, respectively.

The present study attempted to construct a prediction model on the basis of OTU abundance of a genus of intestinal microbes for biogeographic inference. According to the parameter importance ranking of random forest, the most important characteristic differences in classification were mainly Bacteroides, Lactobacillus, and Prevotella-9. Similar to LEfSe analysis, it might be inferred that the main flora of intestinal microbes could be used as a factor in predicting geographic location. Likewise, a study conducted by De Filippo et al. found that Firmicutes and Bacteroides could distinguish children in rural Europe and Africa has significantly demonstrated that Prevotella was a powerful tool for discriminatory classification (De Filippo et al., 2010). The present study found that through verification, the accuracy of the predictions in the three regions was very high, especially in the Meizhou area, where the AUC was 1. All the samples from Meizhou in the verification set were correctly classified, whereas the performance of Guangzhou and Shantou was not satisfactory (the Guangzhou sample and the Shantou sample misjudged each other). Further, the finding of this study was similar to the results of PCoA. It might be possible that a combination of geography, dietary, and other factors play an important role (Yatsunenko et al., 2012). This needs to be understood by further research.

This study provides the first insight into the gut microbiome data of the three characteristic Han populations in Guangdong, which can enrich gut flora information of Chinese ethnic groups. And joint analysis of geography and diet might be helpful to provide enlightening information for forensic science. In addition, due to the complexity of the population composition and living environment of Guangdong Province, so the representativeness of researching samples from the selected three regions is limited. In our current study, individual differences need to be analyzed with large sample size, and the research is still limited to the relative abundance at the genus level. In the future, the sample size will be expanded, sample table information will be recorded in detail (recording used water sources, Food Frequency Questionnaire (FFQ), and other factors), and fecal microbiome analysis will be performed in depth based on microbial species level and sequence. In order to observe the flora differences in different regions of Guangdong Province, follow-up studies will further explore the gut flora of multi-ethnic and multiregional populations.

Conclusion

In conclusion, the current study used high-throughput sequencing methods to study the characteristics of the fecal microbial community of healthy Han individuals living in three regions of Guangdong Province. The results of the current study showed that the composition of intestinal microbes was mainly composed of Bacteroides, Faecalibacterium, and Blautia at the genus level. The feces could be significantly distinguished from saliva samples according to microbial differences at the genus level of both. Further, the populations in the three regions exhibited different enterotype classifications and the prediction model based on the random forest algorithm evidently showed

a significant effect in distinguishing individuals, which might be due to regional differences. In conclusion, microbial community information in feces may have the potential for forensic analysis of body fluid traceability and regionally specific.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, SAMN27409411-SAMN27409469.

Ethics statement

The studies involving human participants were reviewed and approved by Biomedical Ethics Committee of Southern Medical University, Guangzhou, China. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

LH: conceptualization, methodology, visualization, investigation, writing—original draft. LD and CL: validation, formal analysis. EH, XH, CX, and XL: resources, supervision, data curation. HS, CL, and LC: writing—review & editing. All authors discussed the results and contributed to the final manuscript.

Funding

This project was supported by the Open project of Natural Science Foundation of Guangdong Province (Grant no. 2020A1515010938), Science and Technology Program of Guangzhou, China (Grant no. 2019030016 and Grant no. 202102080308), and Medical Science and Technology Research Foundation of Guangdong Province (A2019443). We are grateful to all volunteers who contributed samples for this study.

Acknowledgments

The authors would like to acknowledge the support of the specific colleagues and the collaboration effort of the wider project team, which included Southern Medical University, The First Affiliated Hospital of Guangdong Pharmaceutical University and Guangzhou Forensic Science Institute.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.920780/full#supplementary-material

References

Adak, A., and Khan, M. R. (2019). An insight into gut microbiota and its functionalities. *Cell Mol. Life Sci.* 76, 473–493. doi: 10.1007/s00018-018-2943-4

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473, 174–180. doi: 10.1038/nature09944

Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., and Gordon, J. I. (2005). Host-bacterial mutualism in the human intestine. *Science.* 307, 1915–1920. doi: 10.1126/science.1104816

Blaser, M. J., and Falkow, S. (2009). What are the consequences of the disappearing human microbiota? *Nat. Rev. Microbiol.* 7, 887–894. doi: 10.1038/nrmicro2245

Breiman (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Caesar, R., Tremaroli, V., Kovatcheva-Datchary, P., Cani, P. D., and Bäckhed, F. (2015). Crosstalk between gut microbiota and dietary lipids aggravates WAT inflammation through TLR signaling. *Cell Metab.* 22, 658–668. doi: 10.1016/j.cmet.2015.07.026

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

Claesson, M. J., Cusack, S., O'Sullivan, O., Greene-Diniz, R., de Weerd, H., Flannery, E., et al. (2011). Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl. Acad. Sci. U. S. A.* 108 (Suppl 1), 4586–4591. doi: 10.1073/pnas.1000097107

Davenport, E. R., Sanders, J. G., Song, S. J., Amato, K. R., Clark, A. G., and Knight, R. (2017). The human microbiome in evolution. *BMC Biol.* 15, 127. doi: 10.1186/s12915-017-0454-7

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., et al. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U. S. A.* 107, 14691–14696. doi: 10.1073/pnas.1005963107

Dehingia, M., Devi, K. T., Talukdar, N. C., Talukdar, R., Reddy, N., Mande, S. S., et al. (2015). Gut bacterial diversity of the tribes of India and comparison with the worldwide data. *Sci. Rep*. 5, 18563. doi: 10.1038/srep18563

Du, W., Wu, W., Wu, Z., Guo, L., Wang, B., and Chen, L. (2019). Genetic polymorphisms of 32 Y-STR loci in Meizhou Hakka population. *Int. J. Legal. Med*. 133, 465–466. doi: 10.1007/s00414-018-1845-1

Dwiyanto, J., Hussain, M. H., Reidpath, D., Ong, K. S., Qasim, A., Lee, S., et al. (2021). Ethnicity influences the gut microbiota of individuals sharing a geographical location: a cross-sectional study from a middle-income country. *Sci. Rep*. 11, 2618. doi: 10.1038/s41598-021-82311-3

He, Y., Wu, W., Zheng, H. M., Li, P., McDonald, D., Sheng, H. F., et al. (2018). Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med*. 24, 1532–1535 doi: 10.1038/s41591-018-0164-x

Hooper, L. V., and Gordon, J. I. (2001). Commensal host-bacterial relationships in the gut. *Science* 292, 1115–1118. doi: 10.1126/science.1058709

Hooper, L. V., Midtvedt, T., and Gordon, J. I. (2002). How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu. Rev. Nutr*. 22, 283–307. doi: 10.1146/annurev.nutr.22.011602.092259

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*. 486, 207–214. doi: 10.1038/nature11234

Jandhyala, S. M., Talukdar, R., Subramanyam, C., Vuyyuru, H., Sasikala, M., and Nageshwar, R. D. (2015). Role of the normal gut microbiota. *World J. Gastroenterol*. 21, 8787–8803. doi: 10.3748/wjg.v21.i29.8787

Johnson, D. J., Martin, L. R., and Roberts, K. A. (2005). STR-typing of human DNA from human fecal matter using the QIAGEN QIAamp stool mini kit. *J. Forensic Sci*. 50, 802–808. doi: 10.1520/JFS2004428

Khine, W., Zhang, Y., Goie, G., Wong, M. S., Liong, M., Lee, Y. Y., et al. (2019). Gut microbiome of pre-adolescent children of two ethnicities residing in three distant cities. *Sci Rep*. 9, 7831. doi: 10.1038/s41598-019-44369-y

Kovacs, A., Ben-Jacob, N., Tayem, H., Halperin, E., Iraqi, F. A., and Gophna, U. (2011). Genotype is a stronger determinant than sex of the mouse gut microbiota. *Microb. Ecol*. 61, 423–428. doi: 10.1007/s00248-010-9787-2

Ley, R. E., Peterson, D. A., and Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*. 124, 837–848. doi: 10.1016/j.cell.2006.02.017

Li, L., and Zhao, X. (2015). Comparative analyses of fecal microbiota in Tibetan and Chinese Han living at low or high altitude by barcoded 454 pyrosequencing. *Sci. Rep*. 5, 14682. doi: 10.1038/srep14682

Liang, C., Tseng, H. C., Chen, H. M., Wang, W. C., Chiu, C. M., Chang, J. Y., et al. (2017). Diversity and enterotype in gut bacterial community of adults in Taiwan. *BMC Genomics*. 18, 932. doi: 10.1186/s12864-016-3261-6

Magoč, 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

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet J*. 17, 10–12. doi: 10.14806/ej.17.1.200

Martínez, I., Muller, C. E., and Walter, J. (2013). Long-term temporal analysis of the human fecal microbiota revealed a stable core of dominant bacterial species. *PLoS ONE*. 8, e69621. doi: 10.1371/journal.pone.0069621

Nam, Y. D., Jung, M. J., Roh, S. W., Kim, M. S., and Bae, J. W. (2011). Comparative analysis of Korean human gut microbiota by barcoded pyrosequencing. *PLoS ONE*. 6, e22109. doi: 10.1371/journal.pone.0022109

Oh, H. S., Min, U., Jang, H., Kim, N., Lim, J., Chalita, M., et al. (2022). *Proposal of a health gut microbiome index based on a meta-analysis of Korean and global population datasets*. *J Microbiol*. 60:533–49. doi: 10.1007/s12275-022-1526-0

Quaak, F., van Duijn, T., Hoogenboom, J., Kloosterman, A. D., and Kuiper, I. (2018). Human-associated microbial populations as evidence in forensic casework. *Forensic Sci. Int. Genet*. 36, 176–185. doi: 10.1016/j.fsigen.2018.06.020

Quaak, F. C., de Graaf, M. M., Weterings, R., and Kuiper, I. (2017). Microbial population analysis improves the evidential value of faecal traces in forensic investigations. *Int. J. Legal Med*. 131, 45–51. doi: 10.1007/s00414-016-1390-8

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

Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. *Nature*. 555, 210–215. doi: 10.1038/nature25973

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

Segata, N., Haake, S. K., Mannon, P., Lemon, K. P., Waldron, L., Gevers, D., et al. (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol*. 13, R42. doi: 10.1186/gb-2012-13-6-r42

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol*. 12, R60. doi: 10.1186/gb-2011-12-6-r60

Sender, R., Fuchs, S., and Milo, R. (2016). Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. *Cell*. 164, 337–340. doi: 10.1016/j.cell.2016.01.013

Singh, R. K., Chang, H. W., Yan, D., Lee, K. M., Ucmak, D., Wong, K., et al. (2017). Influence of diet on the gut microbiome and implications for human health. *J. Transl. Med*. 15, 73. doi: 10.1186/s12967-017-1175-y

Song, F. Y., Toshiro, T., Li, K., Yu, P., Lin, X. K., Yang, H. L., et al. (2005). Development of a semi-quantitative food frequency questionnaire for middle-aged inhabitants in the Chaoshan area, China. *World J. Gastroenterol*. 11, 4078–4084. doi: 10.3748/wjg.v11.i26.4078

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci*. 43, 1947–1958. doi: 10.1021/ci034160g

Tagg, J. R., and Ragland, N. L. (1991). Applications of BLIS typing to studies of the survival on surfaces of salivary streptococci and staphylococci. *J. Appl. Bacteriol*. 71, 339–342. doi: 10.1111/j.1365-2672.1991.tb03797.x

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature*. 449, 804–810. doi: 10.1038/nature06244

Wang, M., Liang, B., Zhang, W., Chen, K., Zhang, Y., Zhou, H., et al. (2019). Dietary Lead Exposure and Associated Health Risks in Guangzhou, China. *Int. J. Environ. Res. Public Health*. 16. doi: 10.3390/ijerph16081417

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, W. Z., Wang, C. Y., Cheng, Y. T., Xu, A. L., Zhu, C. L., Wu, S. F., et al. (2010). Tracing the origins of hakka and chaoshanese by mitochondrial DNA analysis. *Am. J. Phys. Anthropol*. 141, 124–130. doi: 10.1002/ajpa.21124

Weinstock, G. M. (2012). Genomic approaches to studying the human microbiota. *Nature*. 489, 250–256. doi: 10.1038/nature11553

Wen, L., and Duffy, A. (2017). Factors influencing the gut microbiota, inflammation, and Type 2 Diabetes. *J Nutr*. 147, 1468S—1475S. doi: 10.3945/jn.116.240754

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*. 334, 105–108. doi: 10.1126/science.1208344

Yao, T., Han, X., Guan, T., Zhai, C., Liu, C., Liu, C., et al. (2021). Exploration of the microbiome community for saliva, skin, and a mixture of both from a population living in Guangdong. *Int. J. Legal Med*. 135, 53–62. doi: 10.1007/s00414-020-02329-6

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature*. 486, 222–227. doi: 10.1038/nature 11053

Zhang, J., Guo, Z., Xue, Z., Sun, Z., Zhang, M., Wang, L., et al. (2015). A phylo-functional core of gut microbiota in healthy young Chinese cohorts across lifestyles, geography and ethnicities. *ISME J*. 9, 1979–1990. doi: 10.1038/ismej.2015.11

Zhang, J., Zheng, Y., Guo, Z., Qiao, J., Gesudu, Q., Sun, Z., et al. (2013). The diversity of intestinal microbiota of Mongolians living in Inner Mongolia, China. *Benef Microbes*. 4, 319–328. doi: 10.3920/BM2013. 0028

Zhong, Z., Liu, J., Li, B., Li, C., Liu, Z., Yang, M., et al. (2017). Serum lipid profiles in patients with acute myocardial infarction in Hakka population in southern China. *Lipids Health Dis*. 16, 246. doi: 10.1186/s12944-017-0636-x

Zou, K. N., Ren, L. J., Ping, Y., Ma, K., Li, H., Cao, Y., et al. (2016). Identification of vaginal fluid, saliva, and feces using microbial signatures in a Han Chinese population. *J. Forensic Leg. Med*. 43, 126–131. doi: 10.1016/j.jflm.2016. 08.003

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 17 August 2022
DOI 10.3389/fmicb.2022.900702

OPEN ACCESS

EDITED BY
Zheng Zhang,
Shandong University,
China

REVIEWED BY
Adrienne Narrowe,
Agricultural Research Service (USDA),
United States
Osiris Gaona,
National Autonomous University of Mexico,
Mexico
Yan Wang,
Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, China

*CORRESPONDENCE
Matthew S. Payne
matthew.payne@uwa.edu.au

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

RECEIVED 21 March 2022
ACCEPTED 18 July 2022
PUBLISHED 17 August 2022

CITATION
Sindi AS, Stinson LF, Lean SS, Chooi Y-H, Leghi GE, Netting MJ, Wlodek ME, Muhlhausler BS, Geddes DT and Payne MS (2022) Effect of a reduced fat and sugar maternal dietary intervention during lactation on the infant gut microbiome.
Front. Microbiol. 13:900702.
doi: 10.3389/fmicb.2022.900702

Effect of a reduced fat and sugar maternal dietary intervention during lactation on the infant gut microbiome

Azhar S. Sindi[1,2], Lisa F. Stinson[3], Soo Sum Lean[3], Yit-Heng Chooi[3], Gabriela E. Leghi[4], Merryn J. Netting[5,6,7], Mary E. Wlodek[3,8], Beverly S. Muhlhausler[4,9], Donna T. Geddes[3] and Matthew S. Payne[1,10]*

[1]Division of Obstetrics and Gynecology, The University of Western Australia, Perth, WA, Australia, [2]College of Applied Medical Sciences, Umm Al-Qura University, Makkah, Saudi Arabia, [3]School of Molecular Sciences, The University of Western Australia, Perth, WA, Australia, [4]School of Agriculture, Food and Wine, The University of Adelaide, Adelaide, SA, Australia, [5]Women and Kids Theme, South Australian Health and Medical Research Institute (SAHMRI), Adelaide, SA, Australia, [6]Discipline of Pediatrics, The University of Adelaide, Adelaide, SA, Australia, [7]Women's and Children's Hospital, Adelaide, SA, Australia, [8]Department of Obstetrics and Gynecology, University of Melbourne, Melbourne, VIC, Australia, [9]CSIRO, Adelaide, SA, Australia, [10]Women and Infants Research Foundation, Perth, WA, Australia

Objective: A growing body of literature has shown that maternal diet during pregnancy is associated with infant gut bacterial composition. However, whether maternal diet during lactation affects the exclusively breastfed infant gut microbiome remains understudied. This study sets out to determine whether a two-week of a reduced fat and sugar maternal dietary intervention during lactation is associated with changes in the infant gut microbiome composition and function.

Design: Stool samples were collected from four female and six male (n=10) infants immediately before and after the intervention. Maternal baseline diet from healthy mothers aged 22–37 was assessed using 24-h dietary recall. During the 2-week dietary intervention, mothers were provided with meals and their dietary intake was calculated using FoodWorks 10 Software. Shotgun metagenomic sequencing was used to characterize the infant gut microbiome composition and function.

Results: In all but one participant, maternal fat and sugar intake during the intervention were significantly lower than at baseline. The functional capacity of the infant gut microbiome was significantly altered by the intervention, with increased levels of genes associated with 28 bacterial metabolic pathways involved in biosynthesis of vitamins (p=0.003), amino acids (p=0.005), carbohydrates (p=0.01), and fatty acids and lipids (p=0.01). Although the dietary intervention did not affect the bacterial composition of the infant gut microbiome, relative difference in maternal fiber intake was positively associated with increased abundance of genes involved in biosynthesis of storage compounds (p=0.016), such as cyanophycin. Relative difference in maternal protein intake was negatively associated with *Veillonella parvula*

(p=0.006), while positively associated with *Klebsiella michiganensis* (p=0.047). Relative difference in maternal sugar intake was positively associated with *Lactobacillus paracasei* (p=0.022). Relative difference in maternal fat intake was positively associated with genes involved in the biosynthesis of storage compounds (p=0.015), fatty acid and lipid (p=0.039), and metabolic regulator (p=0.038) metabolic pathways.

Conclusion: This pilot study demonstrates that a short-term maternal dietary intervention during lactation can significantly alter the functional potential, but not bacterial taxonomy, of the breastfed infant gut microbiome. While the overall diet itself was not able to change the composition of the infant gut microbiome, changes in intakes of maternal protein and sugar during lactation were correlated with changes in the relative abundances of certain bacterial species.

Clinical trial registration: Australian New Zealand Clinical Trials Registry (ACTRN12619000606189).

KEYWORDS

maternal diet, breastfeeding, breast milk, infant gut microbiome, microbial metagenomics, metagenomic sequencing

Introduction

The early postnatal period is a critical window for the development of the infant gut microbiome (Derrien et al., 2019), which has been associated with the programming of lifelong health and disease risk (Arrieta et al., 2015, 2018; Walker, 2017). The infant gut microbiome has a role in immune system development and protection against colonization with pathogens (Olin et al., 2018). Perturbations to the infant gut microbiota have been associated with the development of chronic diseases, including allergic disorders and obesity (Prescott, 2013). Early life gut microbiome establishment is a relatively dynamic process that is influenced by a range of environmental and host factors, including maternal diet (Chu et al., 2016; Lundgren et al., 2018; Savage et al., 2018; Ponzo et al., 2019; Babakobi et al., 2020; García-Mantrana et al., 2020), mode of delivery (Bäckhed et al., 2015; Bokulich et al., 2016), feeding practices (breastfeeding, formula, and the introduction of solid food; Thompson et al., 2015; Timmerman et al., 2017; Ho et al., 2018; Stewart et al., 2018), antibiotic use (Bokulich et al., 2016), gestational age at delivery (Hill et al., 2017), host genetics (Benson et al., 2010), and geography (De Filippo et al., 2010; Lin et al., 2013). However, breastfeeding has been reported to be the single most important factor associated with infant gut microbiome composition and function (Stewart et al., 2018).

Several studies have investigated maternal contributions to infant health. Maternal diet during pregnancy has been associated with infant health outcomes, including allergic diseases (Chatzi et al., 2008, 2013); diet-associated alterations to the infant gut microbiome may be implicated in such cases. To date, there have been several observational studies investigating the effect of maternal diet during gestation on the infant gut microbiome (Chu et al., 2016; Lundgren et al., 2018; Savage et al., 2018; Ponzo et al., 2019; Babakobi et al., 2020; García-Mantrana et al., 2020). However, only one study has examined the effect of maternal diet during lactation on the infant gut microbiome. Unfortunately, the results of this study are limited by the fact that maternal dietary intake during lactation was combined with maternal dietary intake during pregnancy, so that the effect of the diet during the lactation period alone could not be analyzed (Babakobi et al., 2020). All previous studies examining the relationship between maternal diet and the infant gut microbiome suffer from a number of flaws related to the methods used to assess maternal dietary intake, time of maternal dietary assessment in relation to infant stool sample collection and underreporting of confounders such as antibiotics, vitamin supplements, and probiotics, all of which may impact results. The use of short amplicon sequencing by these studies also generally limits taxonomical resolution of bacterial communities to the genus level (Walsh et al., 2018) and provides no accurate information on their functional potential (Langille et al., 2013). As such, well-designed dietary intervention studies are required to better understand the effect of maternal diet during lactation on the infant gut microbiome.

The most likely mechanisms by which maternal diet affects the infant gut microbiome and for which the most robust evidence exists is the entero-mammary pathway (Jiménez et al., 2008; Jost et al., 2014; Rodríguez, 2014; Milani et al., 2015; Fernández et al., 2016; Asnicar et al., 2017; Duranti et al., 2017; Murphy et al., 2017; Kordy et al., 2020), wherein gut bacteria are transported to the lactating mammary gland and thereby contribute to the human milk (HM) microbiome. It is well established that diet is a key factor that shapes the gut microbiome (De Filippo et al., 2010; Wu et al., 2011; Fava et al., 2012; Yatsunenko et al., 2012; Lin et al., 2013; David et al., 2014; Graf et al., 2015; Kovatcheva-Datchary et al.,

2015; De Filippis et al., 2016; Mandal et al., 2016; Röytiö et al., 2017). In addition, several studies have also shown that maternal diet is associated with the HM microbiome (Williams et al., 2017; Padilha et al., 2019; Babakobi et al., 2020; Cortes-Macías et al., 2020; LeMay-Nedjelski et al., 2020). Since the maternal gut is considered one of the sources of microbes for HM, we hypothesized that maternal diet during lactation can influence the infant gut microbiome. Understanding the effect of the maternal diet during lactation on the early life gut microbiome may allow optimization of dietary recommendations for lactating women to better support infant health and development.

The aim of this study was to determine the effect of a 2-week of a reduced fat and sugar maternal dietary intervention on infant gut microbiome composition and function using shotgun metagenomic sequencing in 10 healthy infants.

Materials and methods

Participants

Healthy, Caucasian, primiparous mothers with ages ranging from 22 to 37 years were invited to participate in the study ($n = 10$). All infants included in our study were healthy, exclusively breastfed, and aged 1.8–4.9 months. Six of the infants were born vaginally, two by emergency caesarean section, two by elective caesarean section, and four were female. During the study period, there was no consumption of antibiotic by infants and mothers. Exclusion criteria were pre-existing maternal diabetes, maternal diseases known to affect gastric absorption (such as gastric ulcers), dietary restrictions (such as vegan, vegetarian, gluten-free, milk-free, or dairy-free diets), pregnancy complications (including gestational diabetes, preeclampsia, preterm delivery, and foetal growth restriction), multiple pregnancies, known congenital abnormalities or health issues in the infant that could significantly affect feeding behavior, and solid food introduction before the first study session. The study was approved by The University of Western Australia Human Research Ethics Committee (RA/4/20/4953) and registered on the Australian New Zealand Clinical Trials Registry (ACTRN12619000606189). All mothers provided informed consent and answered a background health and lifestyle questionnaire on enrolment.

Study design

During the first week of the study, mothers followed their normal diet (Figure 1). After assessment of baseline habitual diet, mothers commenced a 2-week dietary intervention. Diets during this period aimed to reduce intakes of discretionary foods, saturated fats and added sugars in comparison with the women's habitual diet. To increase participant compliance, all meals and snacks were provided to mothers *via* a home delivery service from Lite n' Easy, Queensland, Australia. All meals were designed to contain healthy amounts of fat and sugar according to Food Standards Australia New Zealand, and to meet the energy and nutritional requirements for lactating women (Food Standards Australia New Zealand, 2019). Home visits were conducted weekly during the dietary intervention phase to collect infant stool samples, inquire about any issues, and to undertake anthropometric measures on mothers and their infants.

This study was part of a larger study (Leghi et al., 2021), which aimed to determine the effect of a reduced fat and sugar maternal dietary intervention on HM production, and the associated macronutrient (lactose, protein, and fat) and metabolic hormone (insulin, leptin, and adiponectin) profiles.

Maternal and infant anthropometric measures

Maternal and infant anthropometric data were collected at enrolment, immediately prior to the dietary intervention, and at the conclusion of the dietary intervention. Maternal weight was

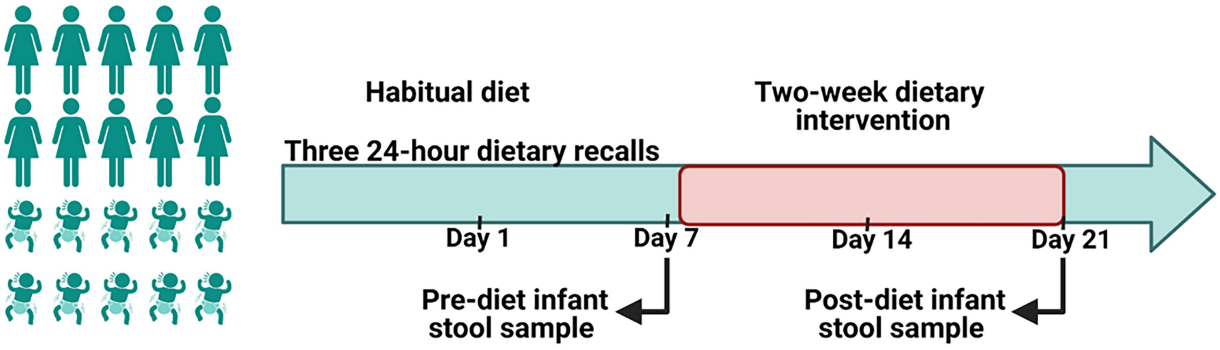

FIGURE 1
A schematic representation of the study design. A 2-week of a reduced fat and sugar maternal dietary intervention during lactation was performed to evaluate the effect of maternal diet on the infant gut microbiome. Before the intervention, mothers consumed their habitual diet, which was assessed using 24h dietary recalls. During the intervention maternal dietary intake was analyzed using FoodWorks 10; Xyris Software. Infant stool samples were collected immediately prior to the intervention (baseline) and at the end of the intervention.

measured using calibrated electronic scales (accurate to 0.1 kg). Maternal height was measured using a Stadiometer (accurate to 0.01 m). Maternal body composition was measured using bioimpedance spectroscopy (Impedimed SFB-7). Infant weight was measured using calibrated electronic scales (Medela Baby Scales, accurate to 2 g). Infant length was measured using a Stadiometer (accurate to 0.01 m). Head circumference was measured using flexible, non-stretchable measuring tape with increments of 0.1 cm and checked against a static measure. Infant body composition was measured using bioimpedance spectroscopy (Impedimed SFB-7).

Maternal dietary assessment

Baseline dietary intake (for 1 week prior to the dietary intervention) was assessed using the Automated Self-Administered 24-Hour Dietary Recall (ASA24) system (Subar et al., 2012). Mothers completed three online 24-h dietary recalls: two on weekdays and one on a weekend day. During the intervention period, mothers were asked to keep a food diary and to record any foods or drinks that they consumed other than those provided, as well as any food that they did not consume from the provided meals and snacks. Maternal energy and macronutrient intakes during the dietary intervention phase were analysed using FoodWorks 10; Xyris Software (Table 1). Baseline dietary intake data were missing for one mother-infant dyad. The relative difference of individual dietary factors between pre-and post-diet was calculated using the following formula: week 3 intake (g)–week 1 intake (g)/week 3 intake (g) × 100.

Infant stool sample collection

Participants serve as their own controls. Infant stool samples were collected immediately before and after the intervention. Due to variations in infant bowel habits, some samples were collected before or after the intended sampling day. Pre-diet samples were collected up to 3 days before, while for post-diet samples, all were collected within 2 days prior to 5 days after the intended sampling day, with the exception of one infant stool sample that was collected 11 days after (Table 2).

Participants self-collected infant stool samples by taking E-swabs of diapers. Stool swabs in 1 ml liquid Amies media were stored in the participant's home freezer (-20°C) before being collected at the next home visit and transferred to the laboratory where they were defrosted and vortexed for 5 s to release the sample from the swab into the liquid transport medium. Samples were then aliquoted and stored at −800°C until analysis.

DNA extraction and metagenomic analysis

DNA was extracted from pre-and post-diet stool samples using the QIAamp 96 PowerFecal QIAcube HT Kit (Qiagen) on the QIAcube HT system (Qiagen). The resulting DNA was quantitated using a high sensitivity dsDNA fluorometric assay (QuantIT, ThermoFisher, Q33120). Samples needed to reach a minimum of 0.2 ng/μl to pass quality control requirements. Libraries were prepared using a modified protocol, using the Illumina® DNA Prep, (M) Tagmentation (96 Samples) kit (Illumina #20018705), to allow for reduced reaction volumes. Libraries were indexed with IDT® for Illumina Nextera DNA Unique Dual Indexes Set A-D (Illumina #20027213–16). Pooled libraries were prepared for sequencing on the NovaSeq6000 (Illumina) with 2 × 150 bp paired-end chemistry. Sequencing was performed to a target depth of 3 Gbp (2 Gbp minimum, approximately 7–16 M paired-end reads) raw read generation before quality filtering. Data quality was guaranteed at 75% and above for reads >Q30 at the completion of the sequencing run. All sample preparation and sequencing was performed at Microba Life Sciences Limited.[1]

Metagenomic data processing

Metagenomic sequencing data quality control

Shotgun metagenomic sequencing data quality control was performed at Microba Life Sciences Limited (see footnote 1). Paired-end DNA sequencing data were demultiplexed and adaptor trimmed using Illumina BaseSpace Bcl2fastq2 (v2.20) accepting one mismatch in index sequences. Reads were then quality trimmed and residual adaptors removed using the software Trimmomatic v0.39 (Bolger et al., 2014) with the following parameters: -phred33 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 CROP:100000 HEADCROP:0 MINLEN:100. Human DNA was identified and removed by aligning reads to the human genome reference assembly 38 (GRCh38.p12, GCF_000001405) using bwa-mem v0.7.17 (Li, 2013) with default parameters except minimum seed length set to 31 (−k 31). Human genome alignments were filtered using SAMtools v1.7 (Li et al., 2009), with flags-ubh -f1-F2304. Any read pairs where at least one read mapped to the human genome with >95% identity over >90% of the read length were flagged as human DNA and removed. All samples were then randomly sub-sampled to a standard depth of 14 M reads, which was then rarefied to 11 M reads.

Quantification of microbial species, gene and pathway abundances

Species profiles were obtained with the Microba Community Profiler v2.0.2 (Parks et al., 2021) using the Microba Genome Database (MGDB) v2.0.0 as the reference. Reads were assigned to genomes within MGDB, and the relative cellular abundance of species clusters was estimated and reported. Quantification of gene and pathway abundance in the metagenomic samples was performed using the Microba Gene and Pathway Profiler (MGPP)

1 www.microba.com/research

10.3389/fmicb.2022.900702

TABLE 1 Estimation of maternal dietary factors before and during the dietary intervention.

Maternal ID	Week 1 (pre-diet intervention)					Week 2					Week 3				
	Protein (g)	Fat (g)	Saturated fat (g)	Sugar (g)	Fibre (g)	Protein (g)	Fat (g)	Saturated fat (g)	Sugar (g)	Fibre (g)	Protein (g)	Fat (g)	Saturated fat (g)	Sugar (g)	Fibre (g)
1	75.68	93.25	36.61	51.69	21.48	98.90	49.97	18.77	574.40	142.11	88.24	53.29	16.08	74.08	28.94
2	102.02	124.75	52.93	93.09	20.54	92.76	49.40	16.65	76.40	35.21	97.89	51.11	15.93	77.35	32.79
3	83.87	96.54	34.64	122.42	25.30	89.10	53.46	15.78	84.18	33.60	82.26	48.85	15.14	82.41	33.40
4	93.41	152.42	62.97	130.91	34.70	100.10	52.42	16.33	86.29	35.74	104.13	59.12	19.53	94.57	37.94
5	138.26	128.60	47.10	135.07	34.20	99.40	54.00	16.08	74.69	33.57	89.79	49.10	13.30	79.32	34.73
6[a]						90.06	43.72	17.67	82.14	28.70	81.50	46.39	16.34	79.09	26.02
7	121.53	155.86	70.22	194.14	32.59	94.32	54.95	19.55	92.18	31.43	92.82	48.53	17.38	80.70	30.63
8	97.79	98.39	34.18	97.22	28.16	89.13	55.70	20.28	86.78	42.06	85.99	54.00	16.66	80.86	30.49
9	127.52	189.62	72.56	135.01	47.95	104.35	59.40	16.80	80.23	34.24	94.62	56.43	15.77	88.15	35.22
10	69.73	119.17	52.92	147.00	28.82	92.86	48.28	14.99	86.21	34.74	91.99	47.55	15.76	85.12	32.76

[a]The three baseline 24-h dietary recalls were not completed by this mother.

v0.1.0 against the Microba Genes (MGENES) database v2.0.0. MGPP is a two-step process. In step 1, all open reading frames (ORFs) from all genomes in MGDB were clustered against UniRef90 (Suzek et al., 2015) release 2019/09 using 90% identity over 80% of the read length with MMSeqs2 Release 10-6d92c (Steinegger and Söding, 2018). Gene clusters were then annotated with the UniRef90 identifiers and linked to the Enzyme Commission (accessed *via* UniProt 2019/09) and Transporter Classification Database (Saier et al., 2016) annotations *via* the UniProt ID mapping service.[2] Enzyme Commission annotations were used to determine the encoding of MetaCyc (Caspi et al., 2020) pathways in each genome using enrichM[3] and pathways that were complete or near complete (completeness >80%), were classified as encoded. In step 2, all DNA sequencing read pairs that aligned with one or more bases to the gene sequence from any protein within an MGENES protein cluster were summed. DNA sequence reads were aligned directly to genome sequences. The genome sequences were annotated using full length ORFs, and the coordinates of these ORFs/genes recorded. The genes were annotated using the entire protein and clustered into protein clusters. When a DNA sequence read aligned to a genome, we required a minimum of one base overlap of the read to its' proximal annotated ORF, in order to count the read toward that protein sequence. These counts were then aggregated for each gene cluster. In essence, we relied on the specificity of the DNA. Read alignments were resolved to a single alignment for each read when possible. We typically were able to assign on average 85% of all reads in a sample. Any unresolved (multi mapped reads) were discarded. In this way, reads were counted only once. Abundances of encoded pathways of species reported as detected by MCP were calculated by averaging the read counts of all genes for each enzyme in that pathway. There was no taxonomy associated with the gene clusters.

Antibiotic resistance genes

Assembled reads were aligned against sequences from known antibiotic resistance genes (ARGs) using ABRicate[4] and starAMR.[5] To ensure accuracy, the Resfinder database and comprehensive antibiotic resistance database (CARD) were utilized in the search.

Statistical analyses

Univariate statistical tests were applied to a filtered set of features. In general, different criteria were applied for taxonomic data and functional data. For pre- *vs* post-diet comparison analyses, paired *t*-tests were used for comparisons of paired measurements of square root-transformed microbial and gene

2 www.uniprot.org/uploadlists/

3 https://github.com/geronimp/enrichM

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

5 https://github.com/phac-nml/staramr

TABLE 2 Infant stool sample collection times across the study.

Infant ID	Pre-diet collection time	Post-diet collection time
1	Day 7	Day 20
2	Day 7	Day 20
3	Day 5	Day 20
4	NA	Day 20
5	Day 5	Day 20
6	Day 4	Day 23
7	Day 7	Day 32
8	Day 5	Day 22
9	Day 4	Day 26
10	Day 6	Day 26

TABLE 3 Participant characteristics (n=10).

Variable	*N*% or Mean (Range) Pre-diet intervention	*N*% or Mean (Range) Post-diet intervention
Maternal age (years)	31.5 [22–37]	31.6 [22–38]
Infant age (months)	3.2 months [1.8–4.9]	4 months [2.5–5.8]
Maternal BMI, kg/m^2	24.9 [17–32.9]	24.5 [16.9–32.77]
BMI category:		
Normal (18.5–24.9)	3 (30%)	4 (40%)
Overweight (25–29.9)	4 (40%)	3 (30%)
Obesity class I (30–34.9)	1 (10%)	1 (10%)
Underweight (<18.5)	2 (20%)	2 (20%)
Maternal probiotic use[a]	1 (10%)	1 (10%)
Infant solid use	0 (0%)	0 (0%)
Mode of delivery:		
Vaginal	6 (60%)	
Emergency Caesarean section	2 (20%)	
Elective Caesarean section	2 (20%)	
Gestational age (weeks)	39.4 [38–41]	
Male infants	6 (60%)	

[a]one or two doses (not often).

abundance data. Rare taxonomic reads present in less than three samples and low abundance reads with a mean relative abundance within infants positive for that taxa of less than 0.5% were excluded. Rare functional reads present in less than three samples and low abundance reads with a maximum sample count less than 2 were excluded. Pearson correlation tests were used to compare continuous variables for analyses of associations between the difference in individual maternal dietary factors and square root-transformed microbial and gene abundance data. Square root transformation was used to normalize data distribution. Rare taxonomic features (species) where read counts were 0 in all samples and low abundance features with a maximum sample count of less than 100 reads were excluded. Rare functional features where read counts were 0 in all samples and low abundance features with a maximum sample count of less than two reads were excluded. p values were corrected for multiple hypothesis testing using the Benjamini and Hochberg false discovery rate correction. Corrected values of $p < 0.05$ are considered statistically significant. No confounders were included in the analysis due to the within individual design and the homogeneity of the group. Redundancy Analysis (RDA) was used to visualize relationships between samples in two-dimensions for identifying sample clusters based on the maternal fiber, protein, sugar, and fat intake.

Results

Participants

Maternal and infant characteristics of the study participants are shown in Table 3.

Effect of the dietary intervention on maternal dietary factor intake and body composition

During the intervention, compared to baseline values, maternal fat, saturated fat, and sugar intake decreased significantly by 59.6, 67.5, and 32.9%, respectively. However, for one participant sugar intake increased during the intervention (Figure 2). There were, however, no significant differences detected for maternal protein and fiber intake. Several changes in maternal body composition were also identified post-intervention (Figure 2), with significant reductions in maternal weight ($p = 0.049$), maternal fat mass ($p = 0.005$), fat mass index ($p = 0.004$), percentage of fat mass ($p = 0.036$), and maternal fat mass to fat-free mass ratio ($p = 0.022$).

Pre- *vs* post-diet infant gut microbiome composition

Number of reads that passed QC as well as the number of reads that mapped for each of the protocols for all samples are reported in (Supplementary Table 1). The most abundant bacterial genera identified in stool samples were *Bifidobacterium* spp. (24.3%), *Bacteroides* spp. (17.5%), and *Clostridium* spp. (11.5%; Figure 3). At the species level, these were *Clostridium neonatale* (11%), *Bifidobacterium longum* (10.4%), and *Bifidobacterium infantis* (8.04%; Figure 4). Seven *Bifidobacterium* spp. were identified, with high inter-individual variability in their abundance. The highest mean abundances were for *B. longum* (10.4%), *B. infantis* (8.04%), and *Bifidobacterium breve* (4.5%). *Bifidobacterium infantis* was present in only one infant stool sample (at 80.3% relative abundance) and its' relative abundance

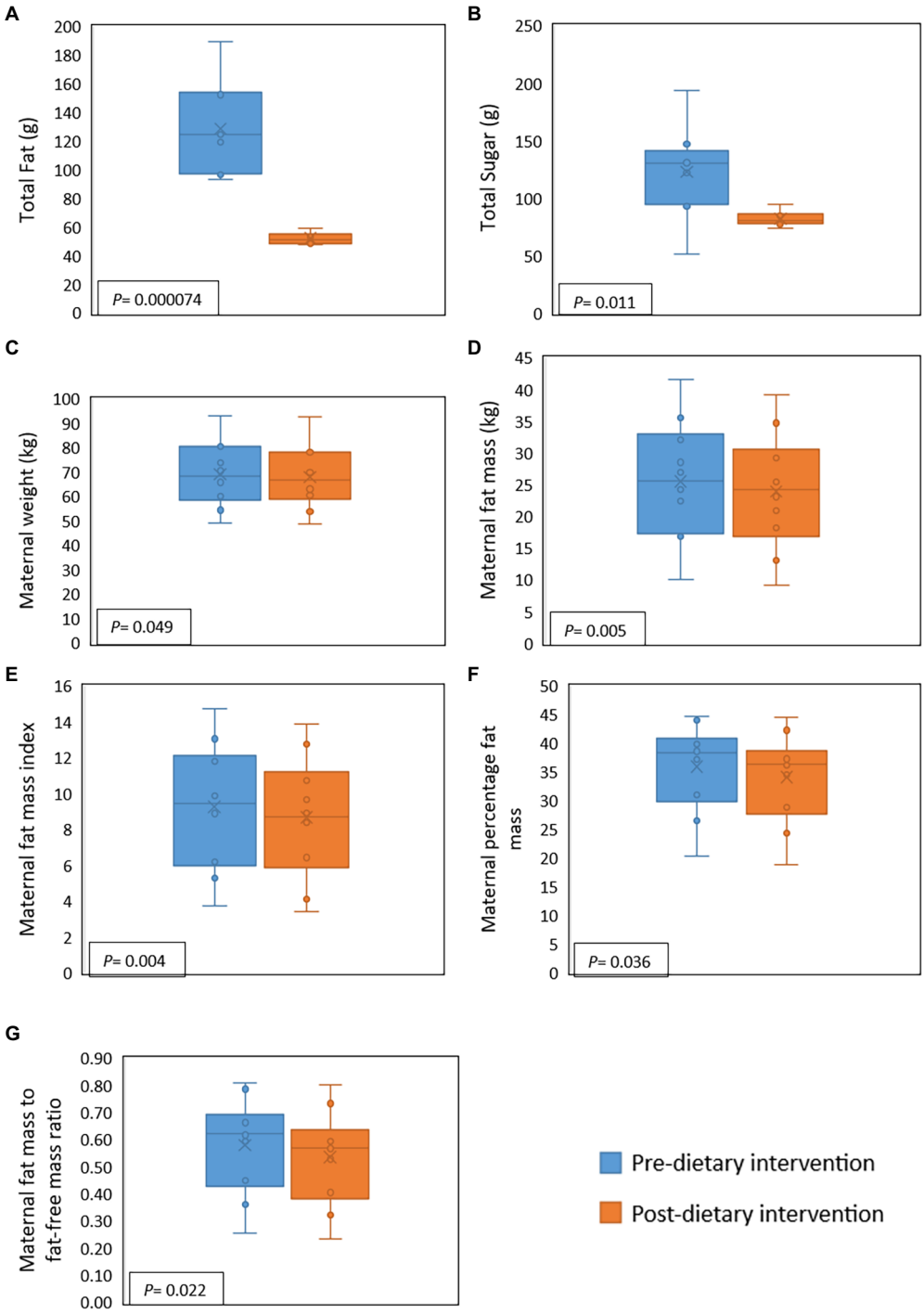

FIGURE 2
Effect of a 2-week maternal dietary intervention on maternal dietary intakes and body composition. Total maternal fat intake **(A)** and total maternal sugar intake **(B)** were significantly reduced by the dietary intervention. Maternal weight **(C)**, fat mass **(D)**, fat mass index **(E)**, percentage of fat **(F)**, and fat mass to fat-free mass ratio **(G)** were also significantly decreased after the dietary intervention. *X* represents the mean value, while the solid line represents the median.

did not change after the dietary intervention (80.4%). Other identified *Bifidobacterium* spp. were *Bifidobacterium bifidum*, *Bifidobacterium dentium*, *Bifidobacterium adolescentis*, and *Bifidobacterium animalis*, with relative abundances ranging from 0.05 to 10.6% and being present in both pre- and post-diet samples of at least one infant.

Pre- and post-diet samples were statistically compared using paired *t*-tests to identify any differences in infant gut microbial

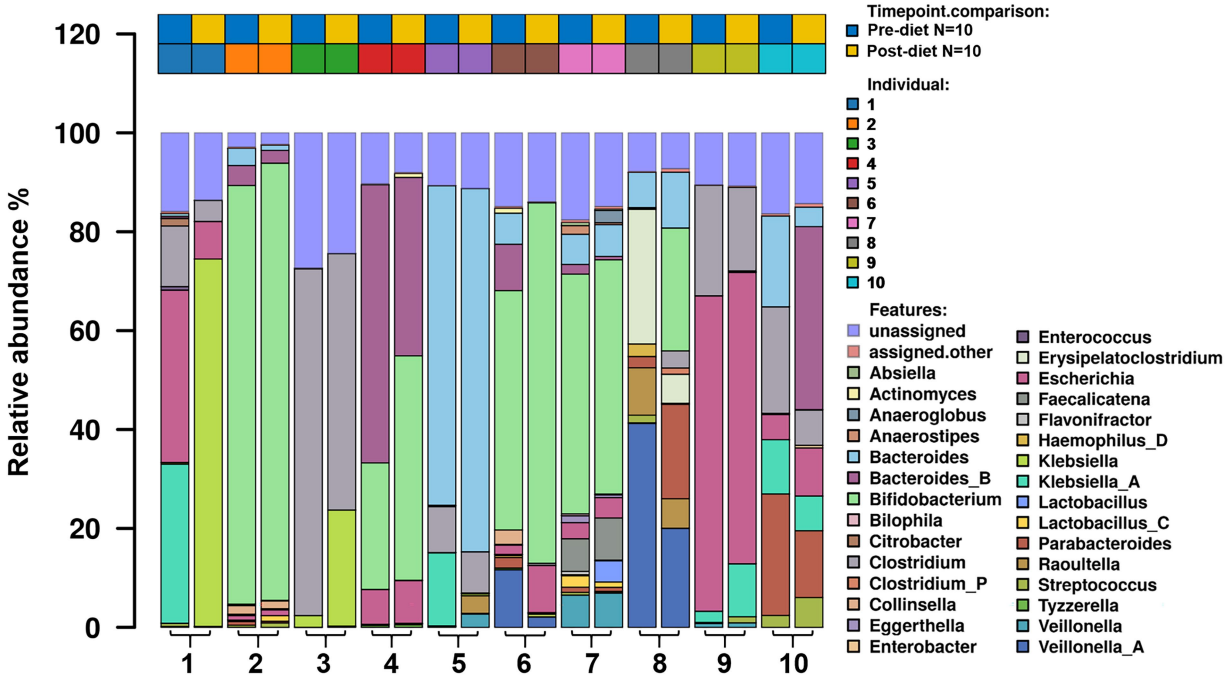

FIGURE 3
The relative abundance of bacterial genera in the infant gut microbiome pre- and post-dietary intervention. Only the top 30 most abundant genera are shown.

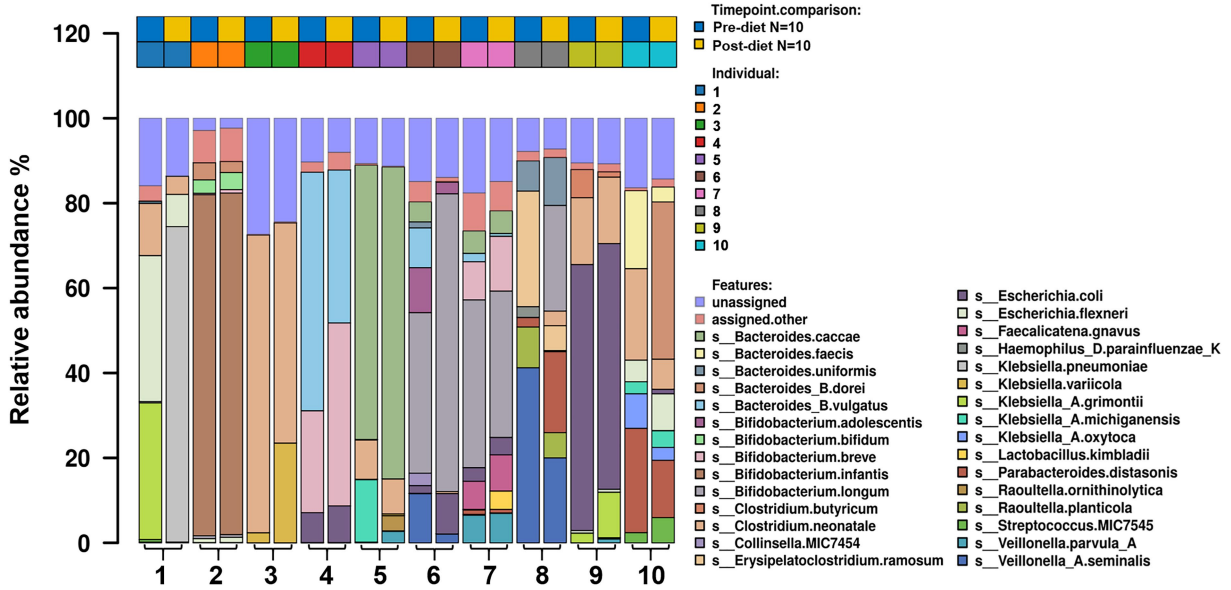

FIGURE 4
The relative abundance of bacterial species in the infant gut microbiome pre- and post-dietary intervention. Only the top most abundant species are shown.

composition within each infant. No statistically significant differences were identified in the relative abundance of any bacterial species between pre-*vs* post-diet samples, potentially due to the low participant numbers and high level of inter-individual variation. However, some microbial compositional changes were identified within each infant. For example, *B. breve* was present in three pre-diet samples (0.35, 24, and 8.9%), and its relative abundance increased in post-diet samples (0.82, 43.1, and 12.9%). It is also worth noting that one infant, whose mother had an increased sugar intake during the intervention, showed the

greatest difference in gut microbiome composition between pre- and post-diet, with decreased relative abundance of *Klebsiella grimontii* (32.2 vs. 0%), *Escherichia flexneri* (34.4 vs. 7.6%), and *Clostridium neonatale* (12.3 vs. 4.3%), and increased relative abundance of *Klebsiella pneumoniae* (0.3 vs. 74.3%; Figure 4). There was no difference in infant gut microbiome alpha-diversity in pre-*vs* post-diet samples (Figures 5A,B; richness $p = 0.08$; Shannon index $p = 0.63$). Nor was there any difference in Bray–Curtis distances between pre-*vs* post-diet samples (Figure 5C; Adonis, $p = 0.99$).

Pre- *vs* post-diet infant gut microbiome functional potential

While post-diet samples did not differ statistically from pre-diet samples in terms of microbial composition, alterations were identified in the functional potential of these microbial communities, with significant differences in multiple bacterial metabolic pathways detected. Overall, 808 gene clusters were significantly different between pre- and post-diet samples (Supplementary Table 2). Significant increases in the abundances of genes involved in 28 bacterial metabolic pathways were detected (Table 4). For instance, post-diet samples showed a significant increase in the potential for biosynthesis of co-factor prosthetic group electron carriers and vitamins ($p = 0.003$), metabolic regulators ($p = 0.003$), amino acids ($p = 0.005$), aromatic compounds ($p = 0.008$), carbohydrates ($p = 0.01$), and fatty acids and lipids ($p = 0.01$). However, while all gene clusters and metabolic pathways that were significantly different post-intervention were rendered insignificant after adjustment for FDR. *p*-values ranged from 0.06 to 0.078, indicating a strong trend toward significance (Table 4). The infant whose mother had an increased sugar

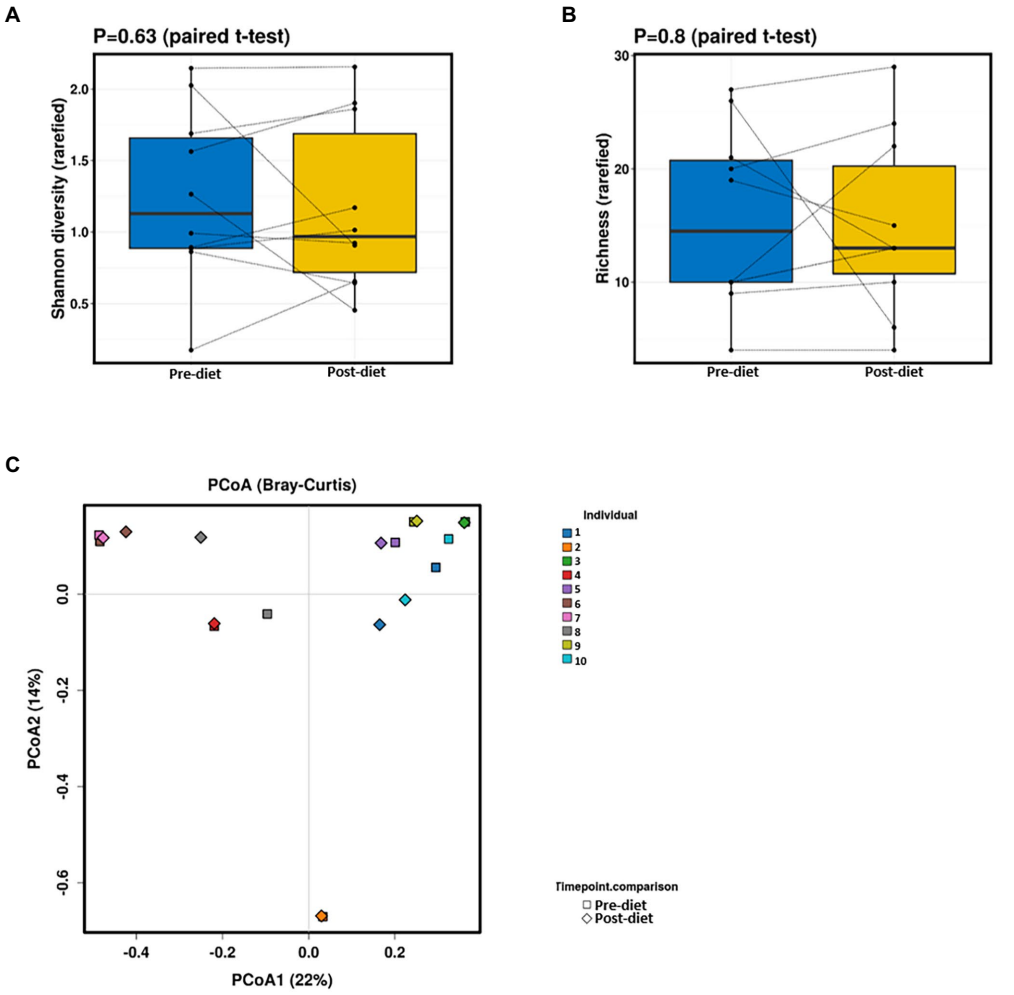

FIGURE 5
No significant differences in infant gut alpha-diversity [Shannon diversity index **(A)** or richness **(B)**] were detected between pre- and post-diet samples (blue and yellow, respectively). Principal coordinates analysis (PCoA) **(C)** of Bray–Curtis distances showed no significant differences between pre-diet and post-diet samples.

TABLE 4 Significantly different bacterial metabolic pathways identified in infant stool samples pre- and post-diet intervention (calculated using paired *t*- test).

Function	*p*-value	FDR
Cofactor prosthetic group electron carrier and vitamin biosynthesis	0.003	0.066
Metabolic regulator biosynthesis	0.003	0.066
Amino acid biosynthesis	0.005	0.066
Unclassified pathways	0.008	0.066
Aromatic compound biosynthesis	0.008	0.066
Carbohydrate biosynthesis	0.009	0.066
Fatty acid and lipid biosynthesis	0.01	0.066
Carbohydrate degradation	0.01	0.066
Aldehyde degradation	0.01	0.072
Glycolysis	0.01	0.072
Amino acid degradation	0.01	0.072
Fermentation	0.01	0.072
Reactive oxygen species degradation	0.02	0.072
Secondary metabolite degradation	0.02	0.072
Inorganic nutrient metabolism	0.02	0.072
Secondary metabolite biosynthesis	0.02	0.072
Hormone biosynthesis	0.02	0.072
Alcohol degradation	0.02	0.073
Glycan degradation	0.03	0.078
Cofactor prosthetic group electron carrier degradation	0.03	0.078
Entner–Doudoroff pathways	0.03	0.078
TCA cycle	0.04	0.078
Cell structure biosynthesis	0.04	0.078
Antibiotic resistance	0.04	0.078
Pentose phosphate pathways	0.04	0.078
Nucleoside and nucleotide biosynthesis	0.04	0.078
Carboxylate degradation	0.04	0.078
Fatty acid and lipid degradation	0.04	0.078

intake during the dietary intervention showed the most substantial difference in gut microbiome functional potential between pre-and post-diet samples, with a general increase in the relative abundance of genes involved in most functional metabolic pathways (Figure 6).

The detected changes in functional potential may possibly be explained by changes in the taxonomical composition of infant stool samples between pre- and post-diet. The dietary intervention resulted in an increase in the mean abundances of *Bacteroides dorei* (0.4 vs. 4%), B. breve (3.3 vs. 5.7%), *B. longum* (7.8% vs. 13%), and *Klebsiella variicola* (0.24 vs. 2.4%), and a decrease in the mean abundances of *Bacteroides vulgatus* (6.8 vs. 3.7%), *B. adolescentis* (1.1 vs. 0.28%), *C. neonatale* (13 vs. 9.1%), *E. flexneri* (4.1 vs. 1.8%), and *K. grimontii* (3.5 vs. 1.1%). The functional potential of these organisms are shown in (Table 5).

Maternal dietary factors and infant gut microbiome composition and function

Although infant gut microbiome composition did not differ significantly in pre-*vs* post-diet samples, correlations were identified between changes in individual dietary factors throughout the dietary intervention, the abundance of certain bacterial taxa, and the functional potential of the associated infant gut microbiome. We calculated the difference in individual dietary factors between pre-and post-diet as relative difference (delta). One mother-infant dyad was excluded from this analysis due to a lack of baseline dietary intake data.

Fiber

No significant correlations were identified between relative difference in dietary fiber intake and infant gut microbiome composition. At a functional level, relative difference in dietary fiber intake was correlated with significant changes in 28 gene clusters in the infant gut microbiome (Supplementary Table 3). An increased abundance of genes involved in the storage compound biosynthesis metabolic pathway were observed ($p = 0.018$; Figure 7). Interestingly, the abundance of the gene for the enzyme responsible for cyanophycin synthesis (cyanophycin synthase) was positively correlated with the relative difference in dietary fiber ($p = 0.025$). Cyanophycin acts as a temporary nitrogen reserve and accumulates in the form of granules in the cytoplasm during phosphate or sulfur starvation (Ziegler et al., 2002).

Protein

Relative difference in dietary protein content was negatively correlated with the relative abundance of *Veillonella parvula* (mean relative abundance 2.16%, $p = 0.006$), while positively correlated with the relative abundance of *Klebsiella michiganensis* (mean relative abundance 3.58%, $p = 0.047$; Figure 8). Functionally, relative difference in dietary protein was correlated with significant changes in 51 gene clusters (Supplementary Table 3). However, no significant correlations were identified with any bacterial metabolic pathways.

Sugar

Relative difference in dietary sugar was positively correlated with the relative abundance of *Lactobacillus paracasei* (mean relative abundance 0.83%, $p = 0.021$; Figure 9). Interestingly, the

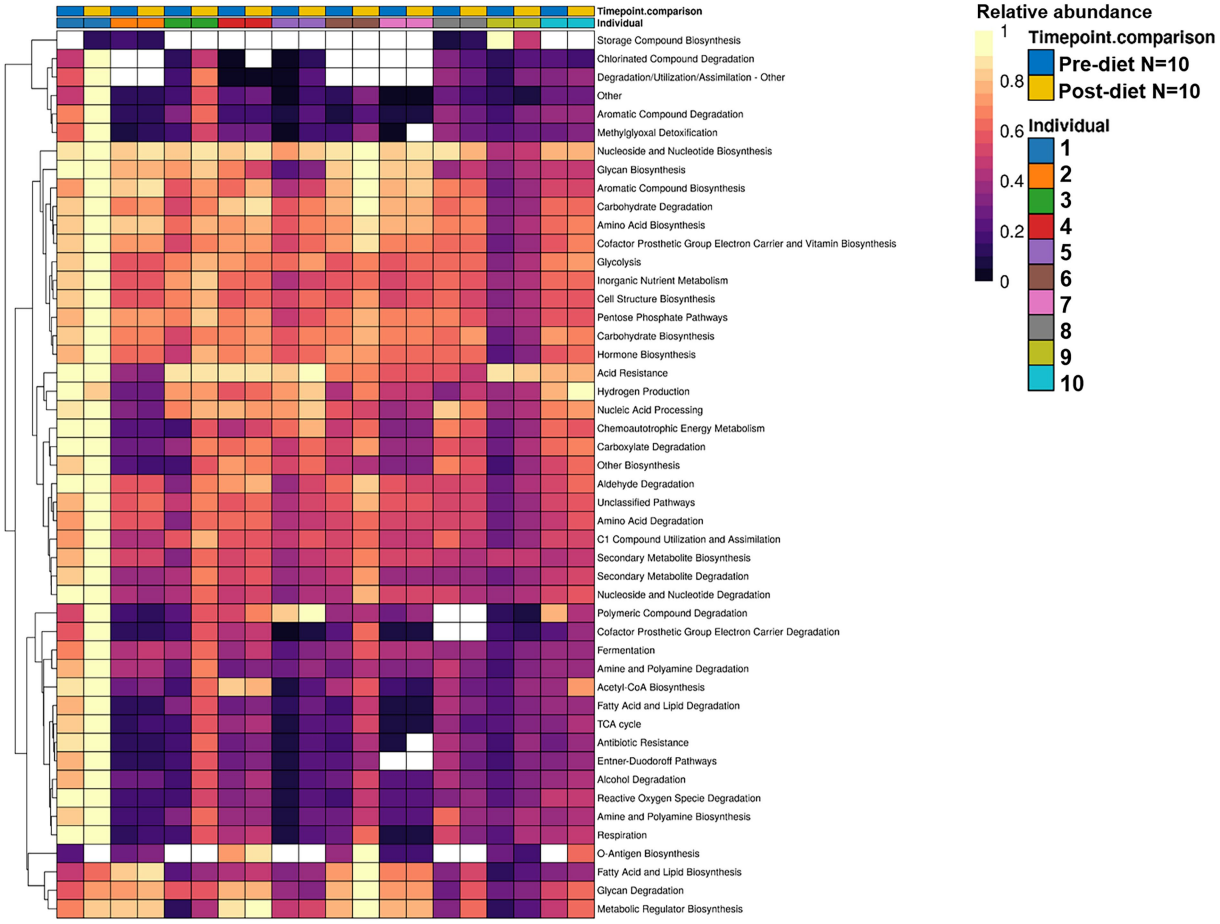

FIGURE 6
Hierarchically clustered heatmap showing the differential relative abundance of bacterial functional metabolic pathways across pre-diet vs. post-diet infant stool samples. A significant increase in the relative abundances of genes involved in 28 bacterial metabolic pathways was detected in post-diet compared to pre-diet samples.

RDA showed significant clustering of infant gut bacterial communities according to change in relative difference in dietary sugar ($p = 0.046$; Figure 10). In addition, a downward trend (although not significant) was observed in the bacterial richness of infant stool samples and increased relative difference in dietary sugar ($p = 0.06$; Figure 11). Relative difference in dietary sugar was correlated with significant changes in 150 gene clusters in the infant gut microbiome (Supplementary Table 3). However, no significant correlations were identified for any bacterial metabolic pathways.

Fat

No significant correlations were identified between relative difference in dietary fat intake and infant gut microbiome composition. However, associations were identified between relative difference in dietary fat and the functional potential of the infant gut microbiome. Relative difference in dietary fat was correlated with significant changes in 140 gene clusters (Supplementary Table 3). These gene clusters fell into three metabolic pathways: storage compounds biosynthesis, fatty acid and lipid biosynthesis, and metabolic regulator biosynthesis, all of which were positively correlated with the relative difference in dietary fat ($p = 0.039$, 0.016, and 0.038, respectively; Figure 12).

Antibiotic resistance genes

Forty unique ARGs were detected in the infant gut microbiome across the pre-and post-diet samples. For both pre-and post-diet samples, ARGs were found across eight different antibiotic classes. The mean number of ARGs per infant in pre-and post-diet samples was 4.8 and 4.7, respectively (Figure 13). The most common ARGs identified in pre-diet samples potentially conferred resistance to tetracycline, while the most common in post-diet samples potentially conferred resistance to tetracycline, erythromycin, and azithromycin (Table 6). Overall, the most common ARGs in both pre-and post-diet samples were tet(Q), tet(O), and msr(D) (Table 6).

TABLE 5 Significantly different bacterial metabolic pathways identified in infant stool samples pre- and post-dietary intervention and the corresponding organisms that potentially account for the functional changes.

Function	Corresponding organisms
Co-factor prosthetic group electron carrier and vitamin biosynthesis	*Bacteroides dorei, Bacteroides vulgatus, Bifidobacterium adolescentis, Bifidobacterium breve, Bifidobacterium longum, Clostridium neonatale, Escherichia flexneri, Klebsiella variicola,* and *Klebsiella grimontii*
Metabolic regulator biosynthesis	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Amino acid biosynthesis	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Aromatic compound biosynthesis	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Carbohydrate biosynthesis	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Fatty acid and lipid biosynthesis	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Carbohydrate degradation	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Aldehyde degradation	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Glycolysis	*B. dorei, B. vulgatus, B. adolescentis, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Amino acid degradation	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Fermentation	*B. dorei, B. adolescentis, B. vulgatus, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Reactive oxygen species degradation	*B. dorei, B. vulgatus, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Secondary metabolite degradation	*B. dorei, B. vulgatus, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Inorganic nutrient metabolism	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Secondary metabolite biosynthesis	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Hormone biosynthesis	*None*
Alcohol degradation	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Glycan degradation	*B. dorei, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Co-factor prosthetic group electron carrier degradation	None
Entner–Doudoroff pathways	*E. flexneri, K. variicola,* and *K. grimontii*
TCA cycle	*C. neonatale, E. flexneri,* and *K. grimontii*
Cell structure biosynthesis	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Antibiotic resistance	*E. flexneri, K. variicola,* and *K. grimontii*
Pentose phosphate pathways	*B. adolescentis, B. vulgatus, B. longum, E. flexneri, K. variicola,* and *K. grimontii*
Nucleoside and nucleotide biosynthesis	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Carboxylate degradation	*B. dorei, B. vulgatus, B. adolescentis, B. breve, B. longum, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*
Fatty acid and lipid degradation	*B. dorei, B. vulgatus, B. breve, C. neonatale, E. flexneri, K. variicola,* and *K. grimontii*

Discussion

We show that a maternal dietary intervention consisting of pre-prepared reduced fat and sugar meals during lactation for 2 weeks significantly alters the functional potential of the infant gut microbiome. The dietary intervention did not, however, affect the bacterial composition of the infant gut microbiome. In addition, changes in individual dietary factors over the course of the diet were correlated with the abundance of certain bacterial taxa, as well as the functional potential of the infant gut microbiome.

The dietary intervention resulted in detection of an increased abundance of genes involved in 28 bacterial metabolic pathways. These metabolic pathways are involved in the biosynthesis and degradation of co-factors, prosthetic groups, electron carriers, vitamins, amino acids, fatty acids, lipids, carbohydrates, and secondary metabolites. Previous studies have shown that infant diet modulates the functional capacity of the infant gut microbiome (Stewart et al., 2018) and that maternal diet modulates the functional capacity of the HM microbiome (Seferovic et al., 2020). For example, Stewart et al. reported that breast milk intake (partially or exclusively) was significantly associated with increased lipid and carbohydrate metabolic pathways in the infant gut microbiome and that breast milk was the single strongest factor responsible for modulation of the infant gut microbiome (Stewart et al., 2018). In comparison, Seferovic et al. showed that a maternal dietary intervention in lactating women can modify the functional capacity of the HM microbiome (Seferovic et al., 2020). A high-fat vs. a high-carbohydrate diet and a high-glucose vs. a high-galactose diet were associated with significant increases in multiple bacterial metabolic pathways (many of them involved in amino acid biosynthesis) in HM. However, similar to our study, the taxonomic composition of the HM microbiome was minimally affected by the maternal dietary intervention. The small sample size in our study may have reduced the power to detect an effect of maternal diet on HM microbiome composition; therefore, further studies with a larger cohort and a longer duration of dietary intervention may reveal more effects of maternal diet that were not detectable in our study. The significant change in the infant gut microbiome function is

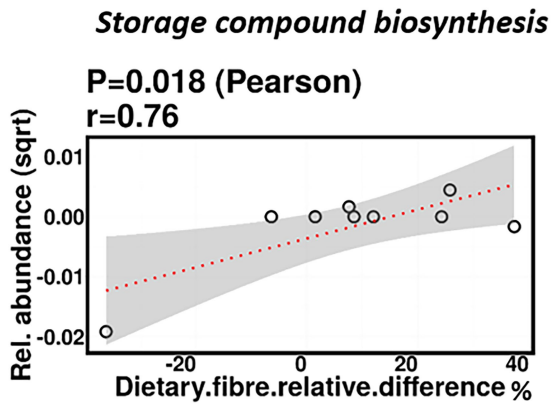

FIGURE 7
Change in maternal fiber intake % was positively correlated with the abundance of genes involved in storage compound biosynthesis.

unlikely driven by the macronutrients level in HM, as the dietary intervention did not change fat, protein, and lactose concentrations in HM (Leghi et al., 2021). Thus, this change might be driven by other HM biochemical components such as human milk oligosaccharides (HMOs), which have been shown to alter the functional capacity of the HM microbiome (Seferovic et al., 2020).

To our knowledge, no previous study has investigated the effects of maternal dietary intervention during lactation on the functional potential of the infant gut microbiome in humans. Studies in pregnant and lactating mice have reported that consumption of a high-fat diet is associated with changes in offspring gut microbiome function (Wankhade et al., 2017; Srinivasan et al., 2018); however, no studies have looked at the effects of a low-fat diet on offspring gut microbiome function. Srinivasan et al. reported that a maternal high-fat diet before, during, and after pregnancy is associated with increased pathways involved in fructose and mannose metabolism and decreased pathways related to indole alkaloid biosynthesis, α-linolenic acid metabolism, and carotenoid metabolism in the offspring gut microbiome (Srinivasan et al., 2018), while Wankhade et al. reported that high-fat diet consumption by dams during pregnancy and lactation was associated with an increase in pathways involved in regulating microbial replication and repair in the offspring gut microbiome (Wankhade et al., 2017). In contrast, our study showed that a reduced fat and sugar maternal dietary intervention resulted in increased presence of pathways involved in genetic material synthesis, such as nucleoside and nucleotide biosynthesis and cell structure biosynthesis. It should be noted, however, the functional inference reported in the above animal studies (Wankhade et al., 2017; Srinivasan et al., 2018), was generated using PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states), a computational tool for indirect analysis of function predicted from 16S rRNA gene sequencing. This method is somewhat limited in that it is not actually detecting the presence of microbial genetic material in samples, beyond the 16S rRNA gene, but instead inferring function based on taxonomy. This type of functional inference is thereby not as accurate as metagenomic data.

The significant increase in the abundance of genes involved in biosynthesis and degradation of vitamins, amino acids, fatty acids, lipids, and carbohydrates that we observed after the intervention is in agreement with the known role of the gut microbiome in the human metabolism of these dietary components (Rowland et al., 2018; Schoeler and Caesar, 2019). Several gut bacterial species have been associated with amino acid and carbohydrate metabolism and short chain fatty acid (SCFA) synthesis. In addition, the human gut microbiome plays an important role in synthesizing vitamins, especially those that humans cannot synthesize, such as thiamin (vitamin B1; LeBlanc et al., 2013). Indeed, after the intervention, there was a significant increase in the abundance of genes involved in the thiamine diphosphate biosynthesis I pathway ($p = 0.028$), responsible for thiamin synthesis. The mechanism by which the maternal dietary intervention generated these alterations is unclear, since no significant changes in the infant gut microbiome composition were detected after the intervention, and HM macronutrient content was not affected by the intervention (Leghi et al., 2021). However, there were clear compositional differences in the microbiome between pre-and post-diet samples for one individual, whose mother interestingly had an increase in sugar intake during the intervention. Increased sugar intake may have changed a certain component in HM such as HMOs (Seferovic et al., 2020), which could drive this marked change in the infant gut microbiome composition. Future research should consider increasing the sample size and investigating potential associations with HMO profile to validate these findings.

We also identified correlations between the relative change in individual dietary factors and the abundance of genes involved in some bacterial metabolic pathways. Relative difference in dietary fibre was correlated with a significant increase in genes associated with the storage compound biosynthesis metabolic pathway. No such finding has been reported by previous studies investigating the maternal dietary effect on the HM and infant gut microbiomes. However, only one study has linked total maternal fibre intake with changes in the beta diversity of the overall composition of HM KEGG bacterial metabolic pathways (LeMay-Nedjelski et al., 2020). Additionally, relative difference in dietary fat was correlated with a significant increase in genes associated with three bacterial metabolic pathways: storage compounds, fatty acid and lipid, and metabolic regulatory compounds biosynthesis. These results are consistent with those from a previous study in mice, where a maternal high-fat diet was reported to increase bacterial metabolic pathways involved in lipid metabolism and bile acid secretion synthesis of the offspring gut microbiome function in a sex-specific manner (Wankhade et al., 2018). Other studies of non-lactating mice reported that high-fat diet consumption is associated with a significant increase in the abundance of genes

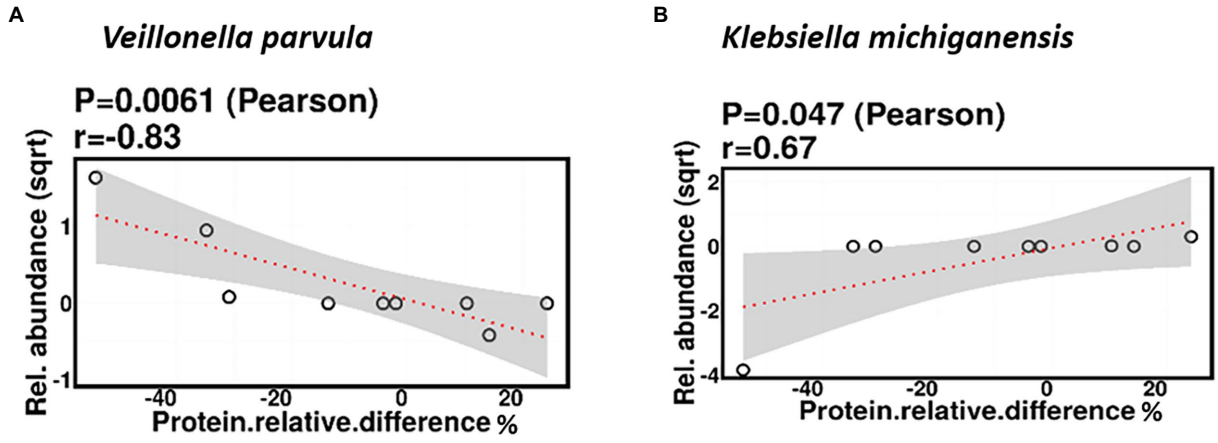

FIGURE 8
Change in maternal protein intake % was negatively correlated with the relative abundance of *Veillonella parvula* **(A)** and positively correlated with the relative abundance of *Klebsiella michiganensis* **(B)**.

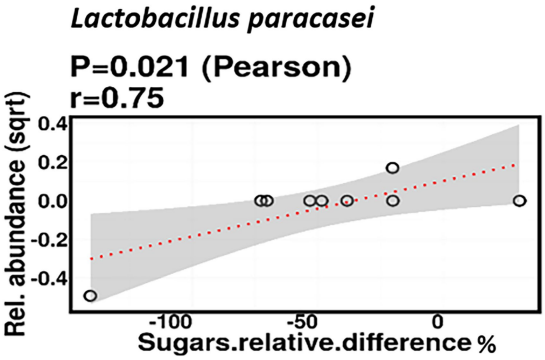

FIGURE 9
Change in maternal sugar intake % was positively correlated with the relative abundance of *Lactobacillus paracasei*.

associated with fatty acid metabolism (Xiao et al., 2017) and fatty acid biosynthesis in the gut microbiome compared to a high-sugar diet (Shan et al., 2019) and control diet (Shang et al., 2017). In contrast to our findings, Hildebrandt et al. reported that when mice on a standard chow diet switched to a high-fat diet, they showed an increased abundance of genes involved in signal transduction and membrane transport and a decreased abundance of genes associated with amino acid and carbohydrate metabolism (Hildebrandt et al., 2009). The biological significance of these findings is unknown; however, it has been reported that bacteria can store the excess of certain nutrients, including lipids, in the form of storage granules as a source of metabolic precursors or as an energy reserve (Murphy and Vance, 1999). Collectively, our results show that maternal diet during lactation influences the infant gut microbiome functional capacity; however, further studies are needed to investigate the mechanisms driving these changes and whether these changes have positive or negative health effects.

The 2-week maternal dietary intervention had no statistically significant effect on the composition of the infant gut microbiome. This is in agreement with the results of the only previous study to investigate the association between maternal diet during lactation and the infant gut microbiome composition (Babakobi et al., 2020). This finding is also in agreement with other dietary interventional studies, where short-term dietary interventions either did not induce changes or failed to significantly alter the gut microbiome composition in healthy adults (Korem et al., 2017), adults at high risk for developing metabolic disorders (Roager et al., 2019), children (Shulman et al., 2017), and mice (Dimova et al., 2017). However, given the high level of inter-individual variation in the infant gut microbiome, the low participant numbers in our study may have limited our ability to detect statistically significant changes in bacterial composition.

We also identified significant correlations between relative differences in individual dietary factors and the composition and functional potential of the infant gut microbiome. These appear to be driven by 2–4 participants that exhibited the greatest changes in dietary components. Decreased relative difference in dietary protein was correlated with an increase in the relative abundance of *V. parvula*. This finding is unexpected as *Veillonella* spp. are known for their role in amino acid hydrolysis and fermentation (Dai et al., 2011). However, no previous study has reported such an association. Relative difference in dietary protein was positively correlated with *Klebsiella michiganensis* relative abundance. This finding is in agreement with the amino acid fermenting function of *Klebsiella* spp. (Dai et al., 2011; Cai et al., 2020). However, no previous study has reported such an association. In terms of sugar, decreases in maternal pre- and post-diet sugar levels were associated with decreases in the relative abundance of *L. paracasei*. This finding is consistent with the known sugar metabolizing functions of *Lactobacillus* spp. (Makarova et al., 2006). The positive correlation between *Lactobacillus* spp. abundance and the

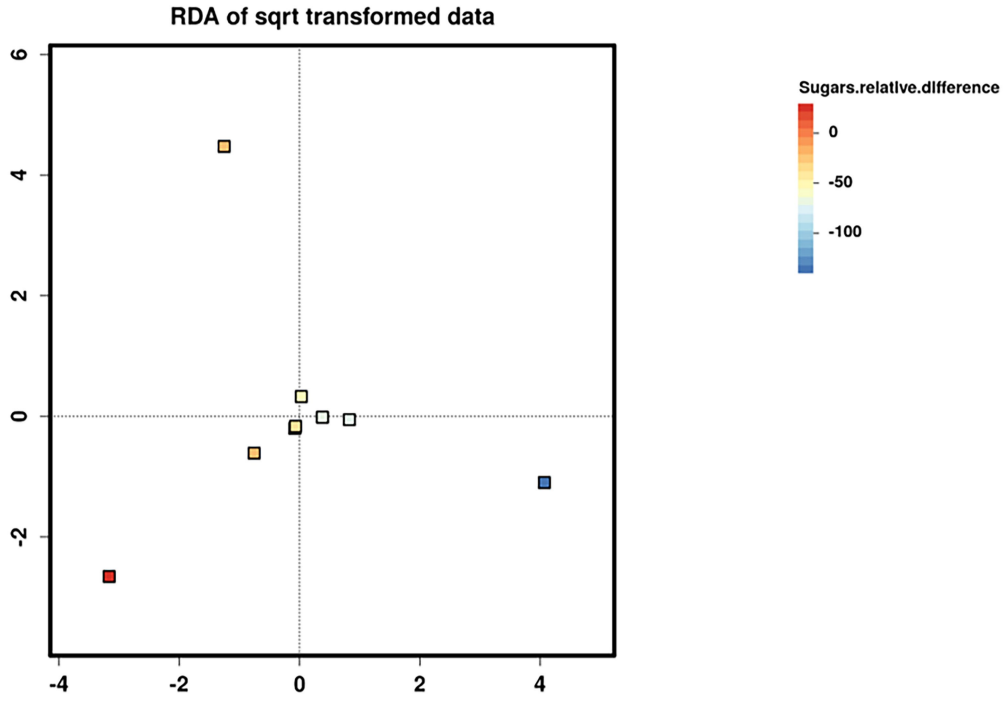

FIGURE 10
Redundancy analysis (RDA) biplots showing the two first axes of ordination for nine infant stool microbiome samples. Samples are coloured according to the change in maternal sugar intake with the dietary intervention.

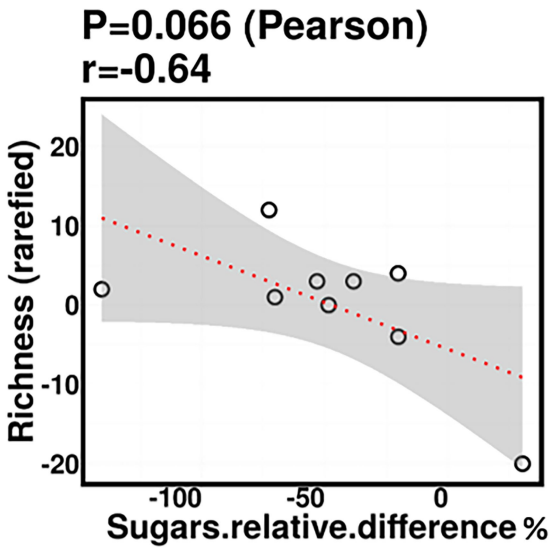

FIGURE 11
Infant stool bacterial richness was negatively correlated with change in maternal sugar intake with the dietary intervention; however, this was not statistically significant.

change in maternal sugar intake is inconsistent with previous studies that associated a high-sugar diet with decreased abundance of *Lactobacillus* spp. in mice (Sen et al., 2017; Yue et al., 2019; Khan et al., 2020). Nevertheless, the increased abundance of *Lactobacillus* spp. in these infants may be beneficial due to the probiotic potential of members of this genus (Martín et al., 2005; Shokryazdan et al., 2014).

The trend toward a negative association between infant stool bacterial richness and relative difference in maternal sugar intake is consistent with observations in mice, wherein mice who consumed a high-sugar diet had decreased alpha-diversity compared to those who consumed a normal diet (Sen et al., 2017; Do et al., 2018). In humans, this finding is in line with those of De Filippo et al., where they showed that gut microbiota richness in African children is higher than European children (De Filippo et al., 2010, 2017). They suggested that the low richness of the gut microbiota in European children could be due to Western diet consumption (high in sugar, animal protein, and fat). However, this finding is based on long-term dietary habits, and there was no direct assessment of dietary intake association with gut microbiome richness. Overall, our results suggest that change in maternal intake of fiber, protein, and sugar during lactation may affect infant gut bacterial composition; however, further studies on a larger cohort need to be conducted to confirm these findings.

Analysis of the infant gut resistome showed that the most prevalent ARGs in pre-and post-diet samples potentially conferred resistance to tetracycline. This finding agrees with previous studies where a high prevalence of resistance to tetracycline was detected in the faecal samples of healthy infants (Gueimonde et al., 2006; Rose et al., 2017; Pärnänen et al., 2018) and those at a high risk of eczema (Loo et al., 2020), suggesting

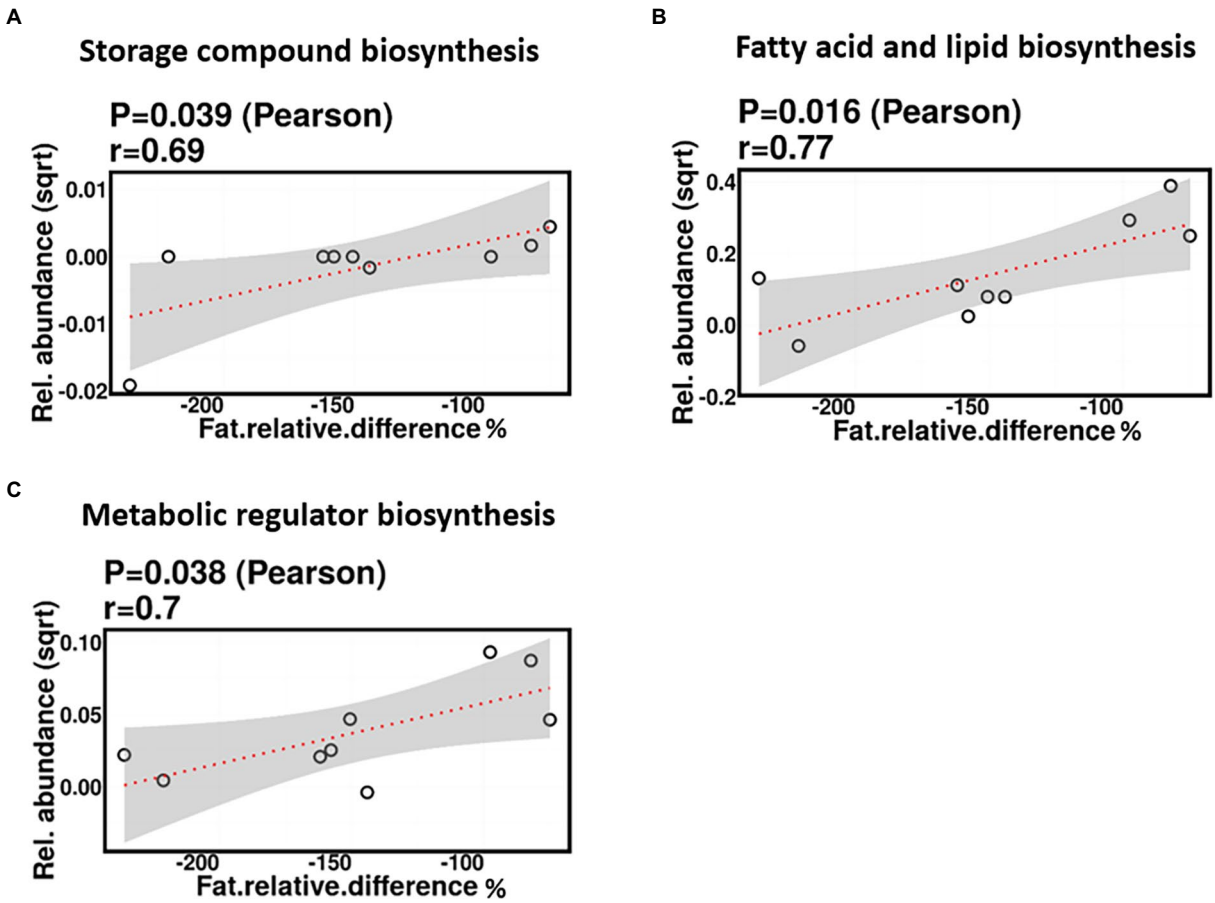

FIGURE 12
Change in maternal fat intake was positively correlated with the abundance of genes involved in storage compound biosynthesis **(A)**, fatty acid and lipid biosynthesis **(B)**, and metabolic regulator biosynthesis **(C)**.

their presence may be unremarkable. The mean number of ARGs per infant did not differ between pre-and post-diet samples (4.8 and 4.7), respectively. In contrast, one study compared the gut resistome of 35 obese children before and after a microbiota-targeted dietary intervention. A diet composed of traditional Chinese medicinal foods, whole grains, and prebiotics resulted in a significant reduction in ARGs within these children (Wu et al., 2016). To date, no study has investigated whether or not maternal diet could affect ARGs of the infant gut microbiome. Infant gut ARGs have been shown to be associated with different factors such as mode of delivery and infant sex (Lebeaux et al., 2021). Maternal diet may also be a factor that could potentially influence the infant gut resistome. Importantly, Pärnänen et al. characterized the HM, infant, and maternal gut microbiomes using metagenomic sequencing in order to identify potential sources of infant gut ARGs. Their results showed that infant gut ARGs resembled those of their own mother's gut and HM, suggesting vertical transmission from mothers to infants (Pärnänen et al., 2018). If maternal diet can impact the maternal gut resistome, it may, in turn, influence the HM and infant gut resistome. Therefore, future research should consider exploring the relationship between maternal diet and infant gut ARGs in a larger cohort.

Strengths and limitations

The key strength of this study is the use of a home delivery service to deliver meals to participants, which increases dietary intervention compliance and consistent dietary intake. In addition, the use of shotgun metagenomics allows characterization of functional capacity for the microbial communities. Another strength is that our study design included baseline (pre-diet) samples to establish a baseline microbiome, combined with use of strict inclusion criteria where all participants with factors that have been shown to impact maternal gut, HM, and infant gut microbiomes were excluded, including probiotic and antibiotic use. A number of limitations do, however, need to be acknowledged. The sample size is small, and the duration of the intervention is relatively short; thus, is it unknown whether the same results may be obtained if a similar dietary intervention were applied on a

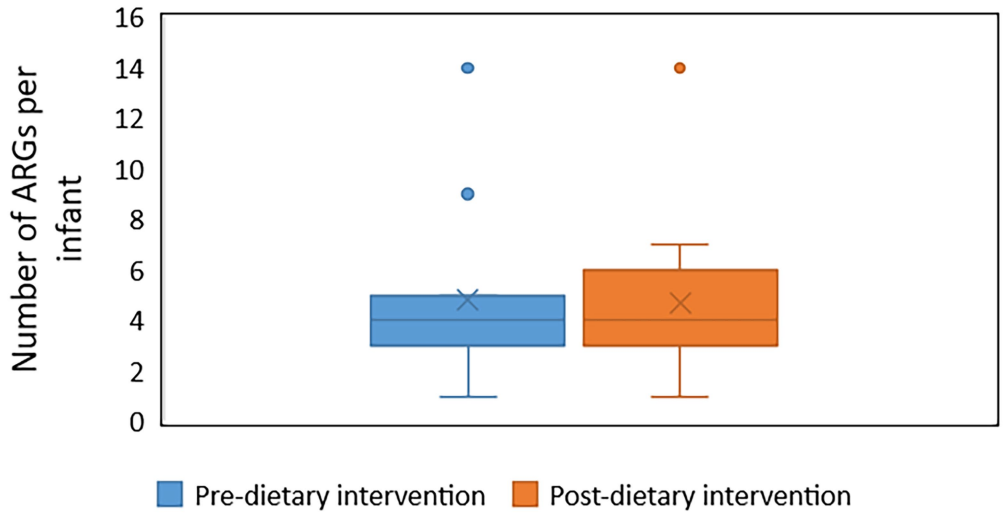

FIGURE 13
The number of antibiotic resistance genes (ARGs) in infant stool samples collected pre-and post-dietary intervention. X represents the mean value.

TABLE 6 Most commonly detected antibiotic resistance genes and their associated antibiotic class in pre-and post-dietary intervention infant faecal samples.

Antibiotic resistance genes	Antibiotic class	Related pathogens	Frequency within samples	
			Pre-diet intervention	Post-diet intervention
mef(A)	Erythromycin, azithromycin	*Streptococcus pneumoniae*	2	2
blaTEM-1B	Ampicillin	*Escherichia coli*	2	2
blaOXY-6-2	Ampicillin	*Klebsiella oxytoca*	2	1
tet(Q)	Tetracycline	*Butyrivibrio fibrisolvens*	6	4
tet(W)	Tetracycline	*Bacteroides fragilis*	2	2
aph(3′)-Ia	Kanamycin	*E. coli*	2	0
*cfx*A4	Ampicillin	*B. fragilis*	2	2
blaACI-1	Ampicillin	*Acidaminococcus intestini*	2	2
cfxA3	Ampicillin	*Pseudomonas aeruginosa*	2	0
tet(O)	Tetracycline	*Campylobacter coli* or *Campylobacter jejuni*	3	2
erm(F)	Erythromycin, azithromycin	*B. fragilis*	1	2
msr(D)	Erythromycin, azithromycin	*S. pneumoniae*	1	3

larger cohort and/or for a longer period. Additionally, time of sampling might be a confounding variable, as maternal weight/BMI typically decreases in the early months postpartum, and the infant gut microbiome develops temporally. The effects of other potential confounders such as infant sex and mode of delivery were minimized by including an equal number of these variables. In analyses of relative abundances, every increase or decrease in a certain microbe's abundance can lead to changes in that from other community members. This is an inherent limitation of using relative abundances to compare between different samples, however, despite this, this method is still well-accepted as an effective way to document the microbial composition of different sample types. Metabolic pathway analyses were not conducted, instead, all reference to these is based on the relative abundance of associated genes, which may or may not have been actively expressed.

Conclusion

This pilot study is the first to report the effect of a controlled maternal dietary intervention during lactation on the breastfed infant gut microbiome. While the effect of maternal diet during pregnancy on the composition of the infant gut microbiome has been previously investigated (Chu et al., 2016; Lundgren et al.,

2018; Savage et al., 2018; Ponzo et al., 2019; Babakobi et al., 2020; García-Mantrana et al., 2020), here, we show that maternal diet during lactation significantly alters the functional potential of the infant gut microbiome. While there were no significant differences in the overall bacterial composition of infant stool samples taken before or after the dietary intervention, we did find associations between changes in individual dietary factor intakes of protein and sugar during lactation and changes in the relative abundances of certain taxa in the infant gut microbiome. The impact of these changes on infant health and development remains to be further investigated. We speculate that the maternal diet may have altered the HM microbiome or other milk components that likely mediated the observed changes in the infant gut microbiome. Future dietary interventional studies investigating the relationship between maternal diet during lactation and the infant gut microbiome should consider examining the HM microbiome and other milk components such as HMOs, antimicrobial peptides, and SCFAs. This may shed light on the mechanism/s by which maternal diet impacts the infant gut microbiome. Results obtained from such studies may allow optimization of dietary recommendations for lactating women to better support breastfed infant health and development.

Data availability statement

The data presented in the study are deposited in the SRA repository, accession number PRJNA819247.

Ethics statement

The studies involving human participants were reviewed and approved by The University of Western Australia Human Research Ethics Committee (RA/4/20/4953). The patients/participants provided their written informed consent to participate in this study.

Author contributions

LS, DG, MW, BM, and MP: study design. AS: sample collection and processing, statistical analyses, and writing original draft. AS, LS, DG, and MP: interpretation. SL: detection of antibiotic resistance genes. GL, MN, LS, DG, MW, BM, and MP: writing review and editing. All authors contributed to the article and approved the submitted version.

Funding

AS, LS, and DG are supported by an unrestricted research grant from Medela AG, administered by The University of Western Australia. MP is supported by a National Health and Medical Research Council Project grant (APP1144040). Umm Al-Qura University, Saudi Arabia provides a PhD scholarship for AS. The funding bodies were not involved in the design of the study, collection/analysis/interpretation of data, writing of the manuscript, or in the decision to publish the results.

Acknowledgments

We would like to acknowledge Kirsty O'Hehir for all of her assistance with study recruitment and implementation of the dietary intervention.

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.900702/full#supplementary-material

References

Arrieta, M.-C., Arévalo, A., Stiemsma, L., Dimitriu, P., Chico, M. E., Loor, S., et al. (2018). Associations between infant fungal and bacterial dysbiosis and childhood atopic wheeze in a nonindustrialized setting. *J. Allergy Clin. Immunol.* 142, 424–434.e10. doi: 10.1016/j.jaci.2017.08.041

Arrieta, M.-C., Stiemsma, L. T., Dimitriu, P. A., Thorson, L., Russell, S., Yurist-Doutsch, S., et al. (2015). Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* 7:307ra152. doi: 10.1126/scitranslmed.aab2271

Asnicar, F., Manara, S., Zolfo, M., Truong, D. T., Scholz, M., Armanini, F., et al. (2017). Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling. *MSystems* 2, e00164–e00216. doi: 10.1128/mSystems.00164-16

Babakobi, M., Reshef, L., Gihaz, S., Belgorodsky, B., Fishman, A., Bujanover, Y., et al. (2020). Effect of maternal diet and Milk lipid composition on the infant gut and maternal milk microbiomes. *Nutrients* 12:2539. doi: 10.3390/nu12092539

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., et al. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* 17, 690–703. doi: 10.1016/j.chom.2015.04.004

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J., et al. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl. Acad. Sci.* 107, 18933–18938. doi: 10.1073/pnas.1007028107

Bokulich, N., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., et al. (2016). Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* 8:343ra82. doi: 10.1126/scitranslmed.aad7121

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

Cai, Z., Guo, Q., Yao, Z., Zheng, W., Xie, J., Bai, S., et al. (2020). Comparative genomics of Klebsiella michiganensis BD177 and related members of Klebsiella sp. reveal the symbiotic relationship with *Bactrocera dorsalis*. *BMC Genet.* 21, 1–13. doi: 10.1186/s12863-020-00945-0

Caspi, R., Billington, R., Keseler, I. M., Kothari, A., Krummenacker, M., Midford, P. E., et al. (2020). The MetaCyc database of metabolic pathways and enzymes-a 2019 update. *Nucleic Acids Res.* 48, D445–D453. doi: 10.1093/nar/gkz862

Chatzi, L., Garcia, R., Roumeliotaki, T., Basterrechea, M., Begiristain, H., Iniguez, C., et al. (2013). Mediterranean diet adherence during pregnancy and risk of wheeze and eczema in the first year of life: INMA (Spain) and RHEA (Greece) mother–child cohort studies. *Br. J. Nutr.* 110, 2058–2068. doi: 10.1017/S0007114513001426

Chatzi, L., Torrent, M., Romieu, I., Garcia-Esteban, R., Ferrer, C., Vioque, J., et al. (2008). Mediterranean diet in pregnancy is protective for wheeze and atopy in childhood. *Thorax* 63, 507–513. doi: 10.1136/thx.2007.081745

Chu, D., Antony, K., Ma, J., Prince, A., Showalter, L., Moller, M., et al. (2016). The early infant gut microbiome varies in association with a maternal high-fat diet. *Genome Med.* 8:77. doi: 10.1186/s13073-016-0330-z

Cortes-Macías, E., Selma-Royo, M., García-Mantrana, I., Calatayud, M., González, S., Martínez-Costa, C., et al. (2020). Maternal diet shapes the breast milk microbiota composition and diversity: impact of mode of delivery and antibiotic exposure. *J. Nutr.* 151, 330–340. doi: 10.1093/jn/nxaa310

Dai, Z.-L., Wu, G., and Zhu, W.-Y. (2011). Amino acid metabolism in intestinal bacteria: links between gut ecology and host health. *Front. Biosci.* 16, 1768–1786. doi: 10.2741/3820

David, L., Maurice, C., Carmody, R., Gootenberg, D., Button, J., Wolfe, B., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505, 559–563. doi: 10.1038/nature12820

De Filippis, F., Pellegrini, N., Vannini, L., Jeffery, I., La Storia, A., Laghi, L., et al. (2016). High-level adherence to a Mediterranean diet beneficially impacts the gut microbiota and associated metabolome. *Gut* 65, 1812–1821. doi: 10.1136/gutjnl-2015-309957

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J., Massart, S., et al. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci.* 107, 14691–14696. doi: 10.1073/pnas.1005963107

De Filippo, C., Di Paola, M., Ramazzotti, M., Albanese, D., Pieraccini, G., Banci, E., et al. (2017). Diet, environments, and gut microbiota. A preliminary investigation in children living in rural and urban Burkina Faso and Italy. *Front. Microbiol.* 8:1979. doi: 10.3389/fmicb.2017.01979

Derrien, M., Alvarez, A.-S., and de Vos, W. M. (2019). The gut microbiota in the first decade of life. *Trends Microbiol.* 27, 997–1010. doi: 10.1016/j.tim.2019.08.001

Dimova, L. G., Zlatkov, N., Verkade, H. J., Uhlin, B. E., and Tietge, U. J. (2017). High-cholesterol diet does not alter gut microbiota composition in mice. *Nutrit. Metabol.* 14, 1–7. doi: 10.1186/s12986-017-0170-x

Do, M. H., Lee, E., Oh, M.-J., Kim, Y., and Park, H.-Y. (2018). High-glucose or-fructose diet cause changes of the gut microbiota and metabolic disorders in mice without body weight change. *Nutrients* 10:761. doi: 10.3390/nu10060761

Duranti, S., Lugli, G., Mancabelli, L., Armanini, F., Turroni, F., James, K., et al. (2017). Maternal inheritance of bifidobacterial communities and bifidophages in infants through vertical transmission. *Microbiome* 5, 66–13. doi: 10.1186/s40168-017-0282-6

Fava, F., Gitau, R., Griffin, B., Gibson, G., Tuohy, K., and Lovegrove, J. (2012). The type and quantity of dietary fat and carbohydrate alter faecal microbiome and short-chain fatty acid excretion in a metabolic syndrome 'at-risk' population. *Int. J. Obes.* 37, 216–223. doi: 10.1038/ijo.2012.33

Fernández, L., Cárdenas, N., Arroyo, R., Manzano, S., Jiménez, E., Martín, V., et al. (2016). Prevention of infectious mastitis by oral administration of *Lactobacillus salivarius* PS2 during late pregnancy. *Clin. Infect. Dis.* 62, 568–573. doi: 10.1093/cid/civ974

Food Standards Australia New Zealand (2019). Sugar. Available at: https://www.foodstandards.gov.au/consumer/nutrition/Pages/Sugar.aspx (Accessed 10, 2021).

García-Mantrana, I., Selma-Royo, M., González, S., Parra-Llorca, A., Martínez-Costa, C., and Collado, M. C. (2020). Distinct maternal microbiota clusters are associated with diet during pregnancy: impact on neonatal microbiota and infant growth during the first 18 months of life. *Gut Microbes* 11, 962–978. doi: 10.1080/19490976.2020.1730294

Graf, D., Di Cagno, R., Fåk, F., Flint, H., Nyman, M., Saarela, M., et al. (2015). Contribution of diet to the composition of the human gut microbiota. *Microb. Ecol. Health Dis.* 26:26164. doi: 10.3402/mehd.v26.26164

Gueimonde, M., Salminen, S., and Isolauri, E. (2006). Presence of specific antibiotic (tet) resistance genes in infant faecal microbiota. *FEMS Immunol. Med. Microbiol.* 48, 21–25. doi: 10.1111/j.1574-695X.2006.00112.x

Hildebrandt, M. A., Hoffmann, C., Sherrill–Mix, S. A., Keilbaugh, S. A., Hamady, M., Chen, Y. Y., et al. (2009). High-fat diet determines the composition of the murine gut microbiome independently of obesity. *Gastroenterology* 137, 1716–1724.e2. doi: 10.1053/j.gastro.2009.08.042

Hill, C. J., Lynch, D. B., Murphy, K., Ulaszewska, M., Jeffery, I. B., O'Shea, C. A., et al. (2017). Evolution of gut microbiota composition from birth to 24 weeks in the INFANTMET cohort. *Microbiome* 5, 1–18. doi: 10.1186/s40168-017-0240-3

Ho, N. T., Li, F., Lee-Sarwar, K. A., Tun, H. M., Brown, B. P., Pannaraj, P. S., et al. (2018). Meta-analysis of effects of exclusive breastfeeding on infant gut microbiota across populations. *Nat. Commun.* 9, 1–13. doi: 10.1038/s41467-018-06473-x

Jiménez, E., Fernández, L., Maldonado, A., Martín, R., Olivares, M., Xaus, J., et al. (2008). Oral administration of Lactobacillus strains isolated from breast milk as an alternative for the treatment of infectious mastitis during lactation. *Appl. Environ. Microbiol.* 74, 4650–4655. doi: 10.1128/AEM.02599-07

Jost, T., Lacroix, C., Braegger, C., Rochat, F., and Chassard, C. (2014). Vertical mother–neonate transfer of maternal gut bacteria via breastfeeding. *Environ. Microbiol.* 16, 2891–2904. doi: 10.1111/1462-2920.12238

Khan, S., Waliullah, S., Godfrey, V., Khan, M. A. W., Ramachandran, R. A., Cantarel, B. L., et al. (2020). Dietary simple sugars alter microbial ecology in the gut and promote colitis in mice. *STM* 12:eaay6218. doi: 10.1126/scitranslmed.aay6218

Kordy, K., Gaufin, T., Mwangi, M., Li, F., Cerini, C., Lee, D., et al. (2020). Contributions to human breast milk microbiome and enteromammary transfer of *Bifidobacterium breve*. *PLoS One* 15:e0219633. doi: 10.1371/journal.pone.0219633

Korem, T., Zeevi, D., Zmora, N., Weissbrod, O., Bar, N., Lotan-Pompan, M., et al. (2017). Bread affects clinical parameters and induces gut microbiome-associated personal glycemic responses. *Cell Metab.* 25, 1243–1253.e5. doi: 10.1016/j.cmet.2017.05.002

Kovatcheva-Datchary, P., Nilsson, A., Akrami, R., Lee, Y., De Vadder, F., Arora, T., et al. (2015). Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of Prevotella. *Cell Metab.* 22, 971–982. doi: 10.1016/j.cmet.2015.10.001

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Lebeaux, R. M., Coker, M. O., Dade, E. F., Palys, T. J., Morrison, H. G., Ross, B. D., et al. (2021). The infant gut resistome is associated with E. coli and early-life exposures. *BMC Microbiol.* 21, 1–18. doi: 10.1186/s12866-021-02129-x

LeBlanc, J. G., Milani, C., de Giori, G. S., Sesma, F., van Sinderen, D., and Ventura, M. (2013). Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Curr. Opin. Biotechnol.* 24, 160–168. doi: 10.1016/j.copbio.2012.08.005

Leghi, G. E., Netting, M. J., Lai, C. T., Narayanan, A., Dymock, M., Rea, A., et al. (2021). Reduction in maternal energy intake during lactation decreased maternal body weight and concentrations of leptin, insulin and adiponectin in human milk without affecting milk production, milk macronutrient composition or infant growth. *Nutrients* 13:1892. doi: 10.3390/nu13061892

LeMay-Nedjelski, L., Asbury, M., Butcher, J., Ley, S., Hanley, A., Kiss, A., et al. (2020). Maternal diet and infant feeding practices are associated with variation in the human Milk microbiota at 3 months postpartum in a cohort of women with high rates of gestational glucose intolerance. *J. Nutr.* 151, 320–329. doi: 10.1093/jn/nxaa248

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [Preprint]. doi: 10.48550/arXiv.1303.3997

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

Lin, A., Bik, E., Costello, E., Dethlefsen, L., Haque, R., Relman, D., et al. (2013). Distinct distal gut microbiome diversity and composition in healthy children from Bangladesh and the United States. *PLoS One* 8:e53838. doi: 10.1371/journal.pone.0053838

Loo, E. X. L., Zain, A., Yap, G. C., Purbojati, R. W., Drautz-Moses, D. I., Koh, Y. Q., et al. (2020). Longitudinal assessment of antibiotic resistance gene profiles in gut microbiomes of infants at risk of eczema. *BMC Infect. Dis.* 20, 1–12. doi: 10.1186/s12879-020-05000-y

Lundgren, S., Madan, J., Emond, J., Morrison, H., Christensen, B., Karagas, M., et al. (2018). Maternal diet during pregnancy is related with the infant stool microbiome in a delivery mode-dependent manner. *Microbiome* 6, 109–111. doi: 10.1186/s40168-018-0490-8

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., et al. (2006). Comparative genomics of the lactic acid bacteria. *Proc. Natl. Acad. Sci.* 103, 15611–15616. doi: 10.1073/pnas.0607117103

Mandal, S., Godfrey, K., McDonald, D., Treuren, W., Bjørnholt, J., Midtvedt, T., et al. (2016). Fat and vitamin intakes during pregnancy have stronger relations with a pro-inflammatory maternal microbiota than does carbohydrate intake. *Microbiome* 4:55. doi: 10.1186/s40168-016-0200-3

Martín, R., Olivares, M., Marín, M. L., Fernández, L., Xaus, J., and Rodríguez, J. M. (2005). Probiotic potential of 3 lactobacilli strains isolated from breast milk. *J. Hum. Lact.* 21, 8–17. doi: 10.1177/0890334404272393

Milani, C., Mancabelli, L., Lugli, G., Duranti, S., Turroni, F., Ferrario, C., et al. (2015). Exploring vertical transmission of bifidobacteria from mother to child. *Appl. Environ. Microbiol.* 81, 7078–7087. doi: 10.1128/AEM.02037-15

Murphy, K., Curley, D., O'Callaghan, T., O'Shea, C.-A., Dempsey, E., O'Toole, P., et al. (2017). The composition of human milk and infant faecal microbiota over the first three months of life: a pilot study. *Sci. Rep.* 7:40597. doi: 10.1038/srep40597

Murphy, D. J., and Vance, J. (1999). Mechanisms of lipid-body formation. *Trends Biochem. Sci.* 24, 109–115. doi: 10.1016/S0968-0004(98)01349-8

Olin, A., Henckel, E., Chen, Y., Lakshmikanth, T., Pou, C., Mikes, J., et al. (2018). Stereotypic immune system development in newborn children. *Cell* 174, 1277–1292.e14. doi: 10.1016/j.cell.2018.06.045

Padilha, M., Danneskiold-Samsøe, N., Brejnrod, A., Hoffmann, C., Cabral, V., Iaucci, J., et al. (2019). The human milk microbiota is modulated by maternal diet. *Microorganisms* 7:502. doi: 10.3390/microorganisms7110502

Parks, D. H., Rigato, F., Vera-Wolf, P., Krause, L., Hugenholtz, P., Tyson, G. W., et al. (2021). Evaluation of the microba community profiler for taxonomic profiling of metagenomic datasets from the human gut microbiome. *Front. Microbiol.* 12:643682. doi: 10.3389/fmicb.2021.643682

Pärnänen, K., Karkman, A., Hultman, J., Lyra, C., Bengtsson-Palme, J., Larsson, D. J., et al. (2018). Maternal gut and breast milk microbiota affect infant gut antibiotic resistome and mobile genetic elements. *Nat. Commun.* 9:3891. doi: 10.1038/s41467-018-06393-w

Ponzo, V., Ferrocino, I., Zarovska, A., Amenta, M., Leone, F., Monzeglio, C., et al. (2019). The microbiota composition of the offspring of patients with gestational diabetes mellitus (GDM). *PLoS One* 14:e0226545. doi: 10.1371/journal.pone.0226545

Prescott, S. L. (2013). Early-life environmental determinants of allergic diseases and the wider pandemic of inflammatory noncommunicable diseases. *J. Allergy Clin. Immunol.* 131, 23–30. doi: 10.1016/j.jaci.2012.11.019

Roager, H. M., Vogt, J. K., Kristensen, M., Hansen, L. B. S., Ibrügger, S., Mærkedahl, R. B., et al. (2019). Whole grain-rich diet reduces body weight and systemic low-grade inflammation without inducing major changes of the gut microbiome: a randomised cross-over trial. *Gut* 68, 83–93. doi: 10.1136/gutjnl-2017-314786

Rodríguez, J. (2014). The origin of human milk bacteria: is there a bacterial entero-mammary pathway during late pregnancy and lactation? *Adv. Nutr.* 5, 779–784. doi: 10.3945/an.114.007229

Rose, G., Shaw, A. G., Sim, K., Wooldridge, D. J., Li, M.-S., Gharbia, S., et al. (2017). Antibiotic resistance potential of the healthy preterm infant gut microbiome. *PeerJ* 5:e2928. doi: 10.7717/peerj.2928

Rowland, I., Gibson, G., Heinken, A., Scott, K., Swann, J., Thiele, I., et al. (2018). Gut microbiota functions: metabolism of nutrients and other food components. *Eur. J. Nutr.* 57, 1–24. doi: 10.1007/s00394-017-1445-8

Röytiö, H., Mokkala, K., Vahlberg, T., and Laitinen, K. (2017). Dietary intake of fat and fibre according to reference values relates to higher gut microbiota richness in overweight pregnant women. *Br. J. Nutr.* 118, 343–352. doi: 10.1017/S0007114517002100

Saier, M. H. Jr., Reddy, V. S., Tsu, B. V., Ahmed, M. S., Li, C., and Moreno-Hagelsieb, G. (2016). The transporter classification database (TCDB): recent advances. *Nucleic Acids Res.* 44, D372–D379. doi: 10.1093/nar/gkv1103

Savage, J., Lee-Sarwar, K., Sordillo, J., Lange, N., Zhou, Y., O'Connor, G., et al. (2018). Diet during pregnancy and infancy and the infant intestinal microbiome. *J. Pediatr.* 203, 47–54.e4. doi: 10.1016/j.jpeds.2018.07.066

Schoeler, M., and Caesar, R. (2019). Dietary lipids, gut microbiota and lipid metabolism. *Rev. Endocr. Metab. Disord.* 20, 461–472. doi: 10.1007/s11154-019-09512-0

Seferovic, M. D., Mohammad, M., Pace, R. M., Engevik, M., Versalovic, J., Bode, L., et al. (2020). Maternal diet alters human milk oligosaccharide composition with implications for the milk metagenome. *Sci. Rep.* 10, 1–18. doi: 10.1038/s41598-020-79022-6

Sen, T., Cawthon, C. R., Ihde, B. T., Hajnal, A., DiLorenzo, P. M., Claire, B., et al. (2017). Diet-driven microbiota dysbiosis is associated with vagal remodeling and obesity. *Physiol. Behav.* 173, 305–317. doi: 10.1016/j.physbeh.2017.02.027

Shan, K., Qu, H., Zhou, K., Wang, L., Zhu, C., Chen, H., et al. (2019). Distinct gut microbiota induced by different fat-to-sugar-ratio high-energy diets share similar pro-obesity genetic and metabolite profiles in prediabetic mice. *Msystems* 4:e00219. doi: 10.1128/mSystems.00219-19

Shang, Y., Khafipour, E., Derakhshani, H., Sarna, L. K., Woo, C. W., Siow, Y. L., et al. (2017). Short term high fat diet induces obesity-enhancing changes in mouse gut microbiota that are partially reversed by cessation of the high fat diet. *Lipids* 52, 499–511. doi: 10.1007/s11745-017-4253-2

Shokryazdan, P., Sieo, C. C., Kalavathy, R., Liang, J. B., Alitheen, N. B., Faseleh Jahromi, M., et al. (2014). Probiotic potential of Lactobacillus strains with antimicrobial activity against some human pathogenic strains. *Biomed. Res. Int.* 2014, 1–16. doi: 10.1155/2014/927268

Shulman, R. J., Hollister, E. B., Cain, K., Czyzewski, D. I., Self, M. M., Weidler, E. M., et al. (2017). Psyllium fiber reduces abdominal pain in children with irritable bowel syndrome in a randomized, double-blind trial. *Clin. Gastroenterol. Hepatol.* 15, 712–719.e4. doi: 10.1016/j.cgh.2016.03.045

Srinivasan, S., Raipuria, M., Bahari, H., Kaakoush, N., and Morris, M. (2018). Impacts of diet and exercise on maternal gut microbiota are transferred to offspring. *Front. Endocrinol.* 9:716. doi: 10.3389/fendo.2018.00716

Steinegger, M., and Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nat. Commun.* 9, 1–8. doi: 10.1038/s41467-018-04964-5

Stewart, C., Ajami, N., O'Brien, J., Hutchinson, D., Smith, D., Wong, M., et al. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* 562, 583–588. doi: 10.1038/s41586-018-0617-x

Subar, A. F., Kirkpatrick, S. I., Mittl, B., Zimmerman, T. P., Thompson, F. E., Bingley, C., et al. (2012). The automated Self-administered 24-hour dietary recall (ASA24): a resource for researchers, clinicians, and educators from the National Cancer Institute. *J. Acad. Nutr. Diet.* 112, 1134–1137. doi: 10.1016/j.jand.2012.04.016

Suzek, B. E., Wang, Y., Huang, H., PB, M. G., Wu, C. H., and Consortium, U. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31, 926–932. doi: 10.1093/bioinformatics/btu739

Thompson, A. L., Monteagudo-Mera, A., Cadenas, M. B., Lampl, M. L., and Azcarate-Peril, M. A. (2015). Milk-and solid-feeding practices and daycare attendance are associated with differences in bacterial diversity, predominant communities, and metabolic and immune function of the infant gut microbiome. *Front. Cell. Infect. Microbiol.* 5:3. doi: 10.3389/fcimb.2015.00003

Timmerman, H. M., Rutten, N. B., Boekhorst, J., Saulnier, D. M., Kortman, G. A., Contractor, N., et al. (2017). Intestinal colonisation patterns in breastfed and formula-fed infants during the first 12 weeks of life reveal sequential microbiota signatures. *Sci. Rep.* 7:8327. doi: 10.1038/s41598-017-08268-4

Walker, W. A. (2017). The importance of appropriate initial bacterial colonization of the intestine in newborn, child, and adult health. *Pediatr. Res.* 82, 387–395. doi: 10.1038/pr.2017.111

Walsh, A. M., Crispie, F., O'Sullivan, O., Finnegan, L., Claesson, M. J., and Cotter, P. D. (2018). Species classifier choice is a key consideration when analysing low-complexity food microbiome data. *Microbiome* 6, 1–15. doi: 10.1186/s40168-018-0437-0

Wankhade, U. D., Zhong, Y., Kang, P., Alfaro, M., Chintapalli, S. V., Piccolo, B. D., et al. (2018). Maternal high-fat diet programs offspring liver steatosis in a sexually dimorphic manner in association with changes in gut microbial ecology in mice. *Sci. Rep.* 8, 1–15. doi: 10.1038/s41598-018-34453-0

Wankhade, U., Zhong, Y., Kang, P., Alfaro, M., Chintapalli, S., Thakali, K., et al. (2017). Enhanced offspring predisposition to steatohepatitis with maternal high-fat diet is associated with epigenetic and microbiome alterations. *PLoS One* 12:e0175675. doi: 10.1371/journal.pone.0175675

Williams, J., Carrothers, J., Lackey, K., Beatty, N., York, M., Brooker, S., et al. (2017). Human milk microbial community structure is relatively stable and related to variations in macronutrient and micronutrient intakes in healthy lactating women. *J. Nutr.* 147, 1739–1748. doi: 10.3945/jn.117.248864

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334, 105–108. doi: 10.1126/science.1208344

Wu, G., Zhang, C., Wang, J., Zhang, F., Wang, R., Shen, J., et al. (2016). Diminution of the gut resistome after a gut microbiota-targeted dietary intervention in obese children. *Sci. Rep.* 6, 1–9. doi: 10.1038/srep24030

Xiao, L., Sonne, S. B., Feng, Q., Chen, N., Xia, Z., Li, X., et al. (2017). High-fat feeding rather than obesity drives taxonomical and functional changes in the gut microbiota in mice. *Microbiome* 5, 1–12. doi: 10.1186/s40168-017-0258-6

Yatsunenko, T., Rey, F., Manary, M., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227. doi: 10.1038/nature11053

Yue, S., Zhao, D., Peng, C., Tan, C., Wang, Q., and Gong, J. (2019). Effects of theabrownin on serum metabolites and gut microbiome in rats with a high-sugar diet. *Food Funct.* 10, 7063–7080. doi: 10.1039/C9FO01334B

Ziegler, K., Deutzmann, R., and Lockau, W. (2002). Cyanophycin synthetase-like enzymes of non-cyanobacterial eubacteria: characterization of the polymer produced by a recombinant synthetase of *Desulfitobacterium hafniense*. *Z. Naturforsch. C* 57, 522–529. doi: 10.1515/znc-2002-5-621

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 08 September 2022
DOI 10.3389/fmicb.2022.975344

OPEN ACCESS

EDITED BY
Zheng Zhang,
Shandong University,
China

REVIEWED BY
Guotian Liu,
Northwest A&F University Hospital, China
Vijay K. Sharma,
Agricultural Research Organization, Israel
Rakshak Kumar,
Institute of Himalayan Bioresource Technology (CSIR), India

*CORRESPONDENCE
Yanfeng Wei
weiyanfeng2022@163.com
Xiangtian Yin
yxt1985@163.com

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

RECEIVED 22 June 2022
ACCEPTED 10 August 2022
PUBLISHED 08 September 2022

CITATION
Yuan L, Jiang H, Jiang X, Li T, Lu P, Yin X and Wei Y (2022) Comparative genomic and functional analyses of *Paenibacillus peoriae* ZBSF16 with biocontrol potential against grapevine diseases, provide insights into its genes related to plant growth-promoting and biocontrol mechanisms. *Front. Microbiol.* 13:975344. doi: 10.3389/fmicb.2022.975344

Comparative genomic and functional analyses of *Paenibacillus peoriae* ZBSF16 with biocontrol potential against grapevine diseases, provide insights into its genes related to plant growth-promoting and biocontrol mechanisms

Lifang Yuan[1], Hang Jiang[2], Xilong Jiang[1], Tinggang Li[1], Ping Lu[3], Xiangtian Yin[1]* and Yanfeng Wei[1]*

[1]Shandong Academy of Grape, Shandong Academy of Agricultural Sciences, Jinan, Shandong, China, [2]Institute of Plant Protection, Shandong Academy of Agricultural Sciences, Jinan, Shandong, China, [3]College of Advanced Agricultural Sciences, Zhejiang A&F University, Hangzhou, Zhejiang, China

Paenibacillus peoriae is a plant growth-promoting rhizobacteria (PGPR) widely distributed in various environments. *P. peoriae* ZBFS16 was isolated from the wheat rhizosphere and significantly suppressed grape white rot disease caused by *Coniella vitis*. Here, we present the complete genome sequence of *P. peoriae* ZBFS16, which consists of a 5.83Mb circular chromosome with an average G+C content of 45.62%. Phylogenetic analyses showed that ZBFS16 belongs to the genus *P. peoriae* and was similar to *P. peoriae* ZF390, *P. peoriae* HS311 and *P. peoriae* HJ-2. Comparative analysis with three closely related sequenced strains of *P. peoriae* identified the conservation of genes involved in indole-3-acetic acid production, phosphate solubilization, nitrogen fixation, biofilm formation, flagella and chemotaxis, quorum-sensing systems, two-component systems, antimicrobial substances and resistance inducers. Meanwhile, *in vitro* experiments were also performed to confirm these functions. In addition, the strong colonization ability of *P. peoriae* ZBFS16 was observed in soil, which provides it with great potential for use in agriculture as a PGPR. This study will be helpful for further studies of *P. peoriae* on the mechanisms of plant growth promotion and biocontrol.

KEYWORDS

***Paenibacillus peoriae*, comparative genome analysis, plant growth-promoting, biocontrol, antimicrobial substances**

Introduction

Paenibacillus peoriae (previously *Bacillus peoriae*) is a Gram-positive, facultatively anaerobic, rod-shaped bacterium with flagella and belongs to the genus *Paenibacillus* and the family *Paenibacillaceae*. Species in the genus *Paenibacillus* are either Gram-positive or variable, facultatively anaerobic or strictly aerobic, produce ellipsoidal endospores, and are nonpigmented, rod-shaped and motile (Ash et al., 1993; Siddiqi et al., 2015). Currently, the genus *Paenibacillus* contains 240 species, including the plant-beneficial species of *P. polymyxa* (Zhang et al., 2018; Timmusk et al., 2019), *P. ehimensis* (Naing et al., 2015), *P. alvei* (Emmanouil et al., 2016), *P. macerans* (Liang et al., 2014), *P. lentimorbus* (DasGupta et al., 2006) and *P. peoriae* (Von der Weid et al., 2003; Jiang et al., 2022). Previously, *P. peoriae* was reported to act as a plant growth-promoting rhizobacteria (PGPR), which can produce biofilms, stably colonize the rhizosphere of plants and compete with other microbiota (Von der Weid et al., 2003; Vejan et al., 2016; Jiang et al., 2022). Meanwhile, *P. peoriae* has the ability to act as a biological control agent against many plant pathogens, including *Fusarium* spp., *Diplodia macrospora*, *D. maydis*, *Verticillium dahlia*, *Rhizoctonia solani*, *Colletotrichum gloeosporioides*, and *C. graminicola* (Von der Weid et al., 2003; Yadav D. et al., 2021; Jiang et al., 2022), and even the antimicrobial peptide purified from *P. peoriae* could protect against *Staphylococcus aureus*, *Escherichia coli*, and *Candida albicans* (Ngashangva et al., 2021).

PGPR has been considered environmentally friendly alternatives to fertilizers or agrochemicals for improving crop yield and quality (Vejan et al., 2016; Hashem et al., 2019). Many microorganisms, such as *Bacillus*, *Pseudomonas*, *Burkholderia*, *Caulobacter*, and *Paenibacillus* spp., are PGPRs, and some have or will be successfully applied in practical applications (Ahemad, 2015; Garcia-Seco et al., 2015; Hashem et al., 2019). Production of indole-3-acetic acid (IAA), the capability of fixation of nitrogen, dissolution of phosphorus, secretion of ferriphagin and plant hormones, and antibiotic biosynthesis are important mechanisms of PGPR (Li et al., 2020; Yin et al., 2022). IAA is an important phytohormone that controls cell enlargement and tissue differentiation of plants. Nitrogen (N) and phosphorus (P) are important nutrients for plant growth and productivity. PGPRs are called diazotrophs because of their ability to fix N_2 in nonleguminous plants and form a nonobligate interaction with host plants (Ahemad, 2015). Additionally, by providing P to plants, PGPRs solubilize inorganic P in soil to low molecular weight organic acids (Zaidi et al., 2009; Yuan et al., 2020). Siderophores can form stable complexes with Fe and other heavy metals (Al, Cd, Cu, Ga, In, Pb and Zn), and most plant growth promotion occurs *via* siderophore-mediated Fe uptake (Rajkumar et al., 2010). *P. polymyxa*., which is closest to *P. peoriae*, was identified as having key genes or gene clusters related to IAA, phosphate solubilization and nitrogen fixation for plant growth promotion (Li et al., 2020; Zhou et al., 2020).

The predominant genera of PGPRs are *Pseudomonas* and *Bacillus*, which have the feature of biocontrol, as well as most species in *Paenibacillus* (Naing et al., 2015; Grady et al., 2016; Hashem et al., 2019). *Paenibacillus* helps to control phytopathogens (bacteria, fungi, nematodes and viruses) by triggering induced systemic resistance (ISR) by producing secondary metabolites (Grady et al., 2016). Antimicrobial substances produced by *Paenibacillus*, including peptides, enzymes, and volatile organic compounds, could be used to control soil-borne fungal pathogens and food-borne bacteria (Zhai et al., 2021). Paenicidin A and penisin are antimicrobial peptides produced by *P. polymyxa* NRRL B-30509 and *Paenibacillus* sp. strain A3, respectively (Baindara et al., 2015; Van Belkum et al., 2015). Paenibacillin exhibits excellent tolerance to pH and heat, with activity against a broad range of fungi and bacteria (Abriouel et al., 2011; Li Y. et al., 2019; Li L. et al., 2019). Nonribosomal peptide synthetases are large multimodular biocatalysts that utilize complex regiospecific and stereospecific reactions to assemble structurally and functionally diverse peptides that have important medicinal applications (Strieker et al., 2010).

The role of *P. peoriae* in plant growth promotion and biological control remained unexplored until very recently, and few reports revealed the mechanisms regarding the plant growth promotion and biological control of *P. peoriae*. *P. peoriae* ZBSF16 exhibit significant broad inhibitory spectra against various pathogenic fungi and bacteria on grape and possess perfect characteristics and potential for the biocontrol of grape diseases. In this study, we demonstrated the sequence and annotation of *P. peoriae* strain ZBSF16 and compared its genome with the three major representative *P. peoriae* strains (*P. peoriae* ZF390, *P. peoriae* HS311 and *P. peoriae* HJ-2) that are beneficial to plant growth. Our aim was to provide important insights into the functions of the biocontrol strains and analyze the mechanisms of plant growth promotion and biological control at the gene level, which will benefit improved application of *P. peoriae* to plants in the field.

Materials and methods

Bacterial strains, culture conditions, antagonistic assays and genomic DNA extraction

P. peoriae ZBSF16 was isolated from the wheat rhizosphere in Shandong Province, China on May 7, 2020 and was deposited as a reference strain (strain no. 24769) in the China General Microbiological Culture Collection Center. Strain ZBSF16 was cultivated in LB (Luria broth) medium at 28°C with shaking at 180 rpm for 24 h. The growth curve and the dynamic change in pH were measured every 4 h by spectrophotometer (Persee, TU-1900) and pH meter (Sartorius, PB-10) and the biochemical tests were performed as described by Yin et al. (2022). The morphology of the strains was observed scanning electron microscope (TESCAN VEGA3 SBU). Strain ZBSF16 was evaluated for its antagonistic

activities to *Coniella vitis*, *Gloeosporium fructigrum, Pestalotiopsis clavispora*, *Alternaria viticola*, *Diaporthe eres*, *F. oxysporum*, *Botrytis cinerea*, *Botryosphaeria dothidea*, *Aspergillus niger*, *F. graminearum*, *F. pseudograminearum* and *Allorhizobium vitis* by plate bioassays inoculated with 2 μl of bacterial suspension (Li Y. et al., 2019). The inoculation concentration of strain ZBSF16 was determined by the optical density at 600 nm (OD_{600} = 0.8). Genomic DNA was extracted from cultured ZBSF16 cells (OD_{600} = 0.8) using a QIAamp® DNA Mini Kit (Qiagen, Valencia, CA, United States) according to the manufacturer's instructions.

Whole-genome sequencing and assembly

The genomic DNA of *P. peoriae* ZBSF16 was sequenced at Biomarker Technologies with the Pacific Biosciences (PacBio) RSII Single Molecule Real Time (SMRT) sequencing platform (Li Y. et al., 2019). For genome assembly, the filtered subreads were assembled by Canu v1.5 software, and then, circlator v1.5.5 was used to cyclize the assembled genome. A 10-kb insert size template library was prepared according to the PacBio Sequel gDNA protocol and sequenced using the PacBio Sequel instrument. Circular genome views of the alignments were generated by CGView (Yuan et al., 2020).

Gene prediction and functional annotation

Genes and components of the genome were predicted by using Prodigal v2.6.3, and functional annotation was performed by comparisons against multiple databases, including NR (nonredundant) protein databases, SwissProt and the enhanced COG database, KEGG database, TrEMBL, and the Eggnog database. Transfer RNA (tRNA) genes were predicted with tRNAscan-SE v2.0, and ribosome RNA (rRNA) genes were predicted with Infernal v1.1.3. antiSMASH v5.0.0 was used to predict secondary metabolic gene clusters, and CRT v1.2 was used for CRISPR identification. Furthermore, pathogenicity and drug resistance can be researched by BLAST against the CAZy, TCDB, CARD, PHI, and VFDB databases.

Phylogenetic tree construction

The evolutionary position of *P. peoriae* ZBSF16 was determined by 16S rDNA gene sequence analysis, multilocus sequence analysis (MLSA) and PhyloPhlAn method (Segata et al., 2013; Asnicar et al., 2020; Yin et al., 2022). 22 strains belonging to *Paenibacillus* were selected for constructing phylogenetic trees to investigate the evolution of strain ZBSF16 (Supplementary Table 1). Five housekeeping genes (16S rRNA, *gyrB*, *rpoD*, *rho*, and *pgk*) were selected for MLSA, sequence alignments of ZBSF16 with other *Paenibacillus* strains were carried out using the maximum likelihood clustering method, which was performed in MEGA6 with a bootstrapping test of 1,000 replications to generate phylogenetic trees.

Comparative genomics analysis and mining for genes related to plant-beneficial traits

For the comparative genomic analysis, the genome sequences of *P. peoriae* ZBSF16 were compared to *P. peoriae* ZF390, *P. peoriae* HS311 and *P. peoriae* HJ-2 by MAUVE comparison software (Darling et al., 2004). Additionally, a circular chromosomal map of all the genomes used in the pan-genome analysis was prepared by using BLAST Ring Image Generator (BRIG) v 0.95, taking strain ZBSF16 as a reference genome (Alikhan et al., 2011; Mukhia et al., 2022). Furthermore, average nucleotide identity (ANI) was conducted by using the orthologous average nucleotide identity (OrthoANI) tool, and *in silico* DNA–DNA hybridization (DDH) was calculated by using the Genome-to-Genome Distance Calculator (GGDC) (Goris et al., 2007). Functional genes involved in plant growth promotion, such as genes responsible for IAA production, phosphate solubilization, nitrogen fixation, biofilm formation and synthesis resistance inducers, were searched in the NCBI databases as described by Kumar et al. (2019). The blast search was performed against the locally constructed database of the publically available genomes of *P. peoriae*, with the genome of *P. peoriae* ZBSF16 as a query. The identities of different functional genes at the amino acid level were compared among the strains by using BLAST, with an E-value cut off of 1e-15 was used for the BLAST search (Kumar et al., 2019). Secondary metabolite gene clusters were predicted by antiSMASH 4.0.2 (Jiang et al., 2022).

Measurement of IAA production, phosphate solubilization, siderophores and ammonia production

To determine the production of IAA, strain ZBSF16 was cultured in DF (peptone, 5.0 g; yeast extract, 1.5 g; beef extract, 1.5 g/l; NaCl, 5.0 g/l; tryptophan, 0.5 g/l) salt minimal medium, with a concentration of L-tryptophan of 1.02 g/l. After incubation for 24 h at 28°C, the IAA concentration was estimated as the method described by Yuan et al., (2020). The capability of strain ZBSF16 to solubilize phosphate was estimated *via* National Botanical Research Institute Phosphate (NBRIP) solid medium as described by Yin et al., (2022), and the clear zone around the colony was measured after 7 days at 28°C. A CAS agar plate was used for qualitative analysis of siderophores, and yellow circles that appeared around the colonies were measured after 7 days at 28°C. The capability of strain ZBSF16 to produce ammonia was detected by the method described by Przemienieck, and Nessler's reagent was used to determine its ability to produce ammonium (Przemieniecki et al., 2019; Elhaissoufi et al., 2020).

Analyses of antibiotic resistance and hemolysis

The characteristics of antibiotic resistance of strain ZBSF16 were tested on nine antibiotics, including ampicillin (200 μg/ml), kanamycin (50 μg/ml), rifampicin (50 μg/ml), vancomycin (50 μg/ml), streptomycin (10 μg/ml), spectinomycin (50 μg/ml), gentamycin (10 μg/ml), tetracycline (5 μg/ml), and chloramphenicol (20 μg/ml). The minimum inhibitory concentration (MIC) and minimum bactericidal concentration (MBC) of spectinomycin for strain ZBSF16 were determined as previously described. *P. peoriae* ZBSF16 was grown in LB broth at 28°C for 24 h, Wagstsuma Blood Agar Base (Hopebio, China) was used to determine hemolysis as described previously (Brillard et al., 2001; Yuan et al., 2020).

Plant growth promotion, colonization and biocontrol assays

To determine the plant growth promotion capability of ZBSF16, ten *Vitis vinifera* seedlings (cv. Red globe) were treated with 50 ml of ZBSF16 culture (10^8 CFU/ml) by irrigation every 15 days for 2 months. Another ten *V. vinifera* seedlings used as controls were treated with sterile water. All treated grape plants were placed in a greenhouse maintained at temperature 28°C and 90% relative humidity (RH). At 60 days after inoculation, the root length, shoot length, fresh weight, and dry weight of the seedlings were measured. Meanwhile, the infection rate and disease index of grape white rot on *Vitis vinifera* seedlings (cv. Red globe) were calculated after inoculating *C. vitis* conidial suspension (10^6 conidial/ml) two month later at 28°C and 70–80% RH (Chethana et al., 2017; Ji et al., 2021).

To observe the population dynamics of the ZBSF16 strain in the rhizosphere soil, *Vitis vinifera* seedlings (cv. Red globe) were transplanted into nursery pots containing sterile soil, and each seedling was irrigated with 50 ml of *P. peoriae* ZBSF16 bacterial suspension at a concentration of 10^8 CFU ml^{-1}. Rhizosphere soil was collected at different time points (0, 7, 14, 21, 28, 35, 42, 49 and 56 days after inoculation), and the number of ZBSF16 in the rhizosphere soil was determined by the plating counting method with LB medium containing spectinomycin and streptomycin.

Grape white rot caused by *Coniella vitis* was used as the pathosystem to determine the biocontrol potential of ZBSF16. Leaves and fruit of *V. vinifera* (cv. Red globe) were used to assess the preventive effect and control effect of strain ZBSF16 as described by Yin et al. (2022). Ten biological replicates were performed for each treatment, and the experiments were independently repeated three times. All the leaves and fruit were maintained at 28°C and 90% RH.

Statistical analysis

All experimental data were analyzed by SPSS 22.0 software, and all the values are presented as the mean ± standard error of at least three replications. Significant differences ($p < 0.05$) were determined by one-way analysis (ANOVA) of variance and Duncan's multiple range test (Yuan et al., 2020; Yin et al., 2022).

Results

Organism information and antagonistic characteristics

As a gram-positive, anaerobic, rod-shaped bacterium with a length of 3–5 μm and a diameter of 0.8–1.2 μm, ZBSF16 can utilize diverse carbon sources and belongs to the *Paenibacillus* genus (Supplementary Figures 1A,B; Supplementary Table 2). The growth curve showed that the strain was in the exponential growth phase between 4 and 20 h after inoculation, with the pH value increasing to 7.77 (Supplementary Figure 1C). Additionally, the strain grew best when the pH value was between 6 and 8 and could endure 2% NaCl (Supplementary Figures 1G,H).

P. peoria ZBSF16 was isolated as a biocontrol agent for use against *Coniella vitis*, which exhibited the highest inhibitory rate of 64.44% (Supplementary Figure 2A). Antagonistic spectrum assays showed that strain ZBSF16 presented broad, strong antipathogenic activities against various fungi on grape, including *Gloeosporium fructigrum*, *Botrytis cinerea*, *Diaporthe eres*, *Alternaria viticola*, *F. oxysporum*, *Aspergillus niger*, *Pestalotiopsis clavispora*, and *Allorhizobium vitis* (Figure 1A). In addition, ZBSF16 is considered a biocontrol agent for its extracellular enzyme activity, and it can produce protease, cellulase and lipoidase, which is an important mechanism for inhibiting pathogens (Supplementary Figures 1D–F).

Plant growth promotion, colonization and biocontrol assays

The ability of ZBSF16 to promote growth was verified by inoculating the rhizosphere of plants of *V. vinifera* (cv. Red globe) with the suspension in the greenhouse. *P. peoriae* ZBSF16 produced siderophores and was considered an excellent PGRP (Supplementary Figure 3E). The rate of growth promotion for the length (weight) of the aboveground parts and the root length (fresh weight, dry weight) were 46.56% (60.20, 183.75%) and 60.78% (137.25, 454.54%), respectively (Figure 2C). In addition, the bacterial counts of ZBSF16 on the root surface were maintained at 10^5 CFU/g after 1 month of inoculation (Supplementary Figure 3F). Further study showed that the infection rate and disease index of grape white rot on *V. vinifera* caused by *C. vitis* were decreased 70% and 62.97, inoculating with strain ZBSF16 compared to the control plants (Supplementary Figures 2D,E).

Two treatments were performed to determine the preventive effect and control effect of strain ZBSF16. The results demonstrated that strain ZBSF16 displayed excellent biocontrol traits for grape white rot disease (Figures 1B,C), with the preventive effects for

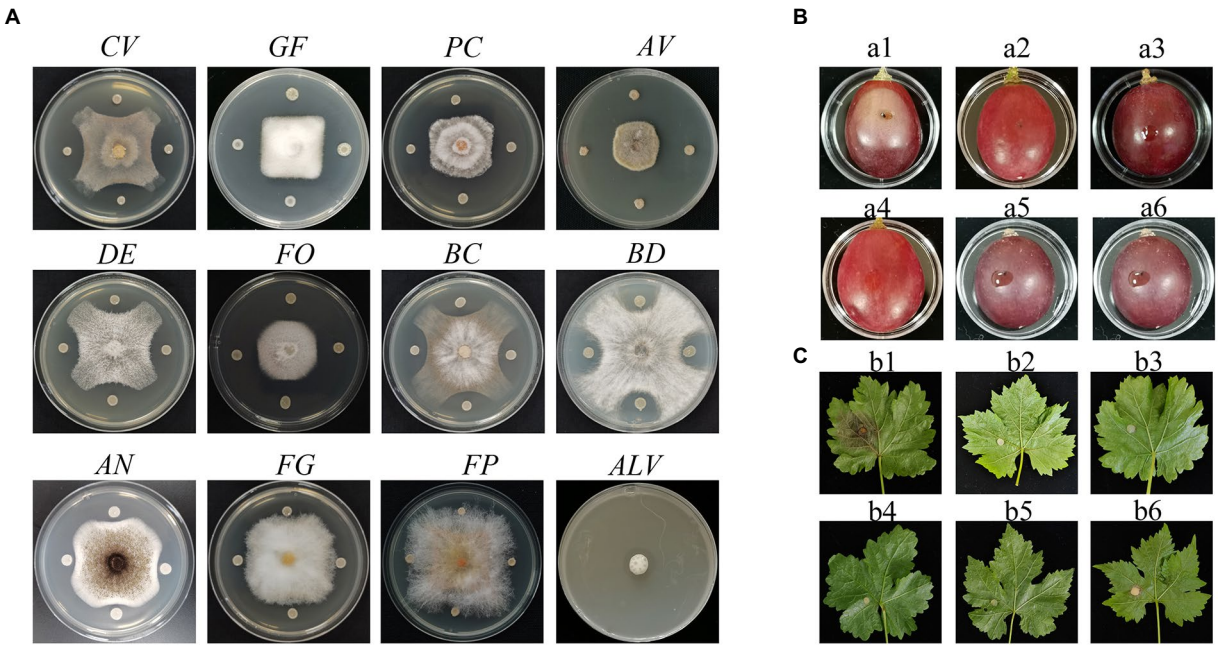

FIGURE 1
Antagonistic assay of *Paenibacillus peoriae* ZBSF16 against eleven pathogenic fungi and one pathogenic bacterium. **(A)** Antagonistic assay of *P. peoriae* ZBSF16. *Coniella vitis* (CV). *Gloeosporium fructigrum* (GF). *Pestalotiopsis clavispora* (Pc). *Alternaria viticola* (Av). *Diaporthe eres* (DE). *Fusarium oxysporum* (FO). *Botrytis cinerea* (BC). *Botryosphaeria dothidea* (BD). *Aspergillus niger* (AN). *Fusarium graminearum* (FG). *Fusarium pseudograminearum* (FP). *Allorhizobium vitis* (ALV). **(B,C)** Biocontrol efficiency of *P. peoriae* ZBSF16 on grape white rot caused by *Coniella vitis*. (a1, b1) Inoculated with *C. vitis*; (a2, b2) LB broth; (a3, b3) sterile water; (a4, b4) culture of ZBSF16; (a5, b5) inoculated with *C. vitis* 24h after inoculation with the culture of ZBSF16; (a6, b6) inoculated culture of ZBSF16 24h after inoculation with *C. vitis*.

detached leaf and detached fruit being 90.59 and 94.52%, respectively. The control effects for detached leaves and detached fruit were 94.52 and 84.70%, respectively (Supplementary Figures 2B,C).

Analyses of antibiotic resistance and hemolysis

The strain ZBSF16 exhibited resistance to ampicillin, chloramphenicol, tetracycline, gentamycin, rifampicin, kanamycin and vancomycin but not to streptomycin or spectinomycin. In addition, strain ZBSF16 showed an MIC of spectinomycin of 216 μg/ml and an MBC of 1,024 μg/ml (Supplementary Figures 3A–C). Meanwhile, the strain was unable to produce hemolysin activity on plates according to the blood agar hemolysis assay (Supplementary Figure 3D).

General genomic features of *Paenibacillus peoriae* ZBSF16

The completed genome of the rod-shaped bacterium *P. peoria* ZBSF16[1] has been shown to be composed of one circular chromosome of 5,839,239 bp in size, with an average G + C content of 45.62% (Figure 3). The details of the assembly information and genomic features are summarized in Supplementary Tables 3, 4. A total of 5,188 predicted genes were identified in the genome, including 4,944 protein-coding sequences, 39 ribosomal RNA operons, 109 tRNAs, and 4 other RNAs. Genes associated with carbohydrate transport and metabolism (7.98%) were the highest density, followed by transcription (7.51%), amino acid transport and metabolism (5.59%), inorganic ion transport and metabolism (4.81%), signal transduction mechanisms (3.82%), replication, cell wall/membrane/envelope biogenesis (3.92%), replication, recombination, and repair (3.61%) and energy production and conversion (3.31%) (Figure 3). In addition, four crisprs were involved in ZBSF16, and the length of the repeated sequences ranged from 19 to 30 bp (Supplementary Table 4).

Comparison of the *Paenibacillus peoriae* ZBSF 16 genome with other completely sequenced *Paenibacillus peoriae* strains

Phylogenetic tree

To determine the relationships of *P. peoria* ZBSF16 with *Paenibacillus* spp. strains, phylogenetic trees based on the 16S rRNA gene sequences were built. The result indicated that ZBSF16 was close to the strain *P. peoria* ZF390; however, *P. kribbensis* AM49

1 https://www.ncbi.nlm.nih.gov/nuccore/CP092831.1

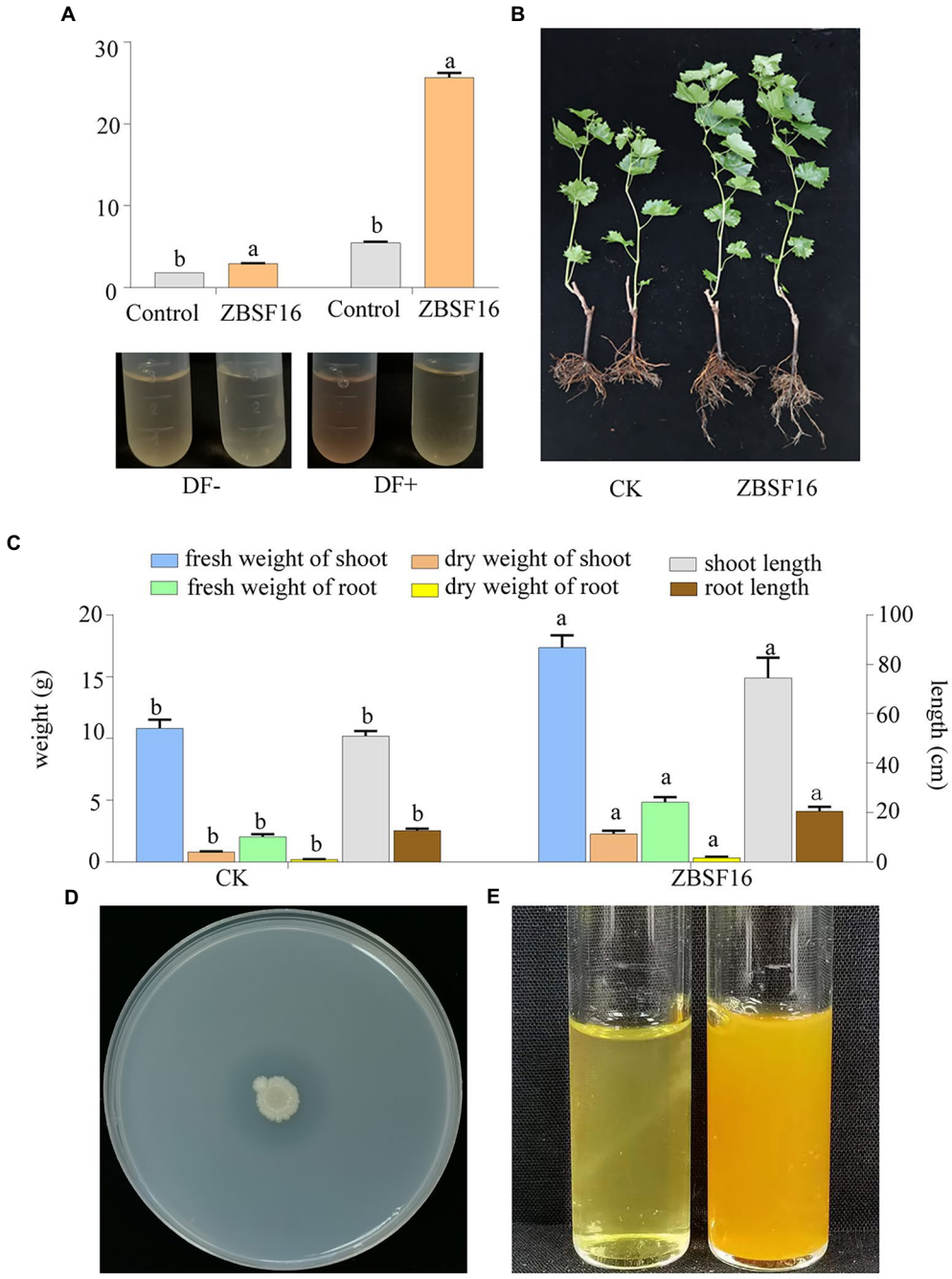

FIGURE 2
Determination of the plant growth-promoting properties of *Paenibacillus peoriae* ZBSF16. **(A)** IAA production of *P. peoriae* ZBSF16. DF–, DF medium without L-tryptophan; DF+, DF medium containing L-tryptophan. **(B,C)** The growth-promoting effect of *Paenibacillus peoriae* ZBSF16 on grape; **(D)** mineral phosphate solubilization of *P. peoriae* ZBSF16; **(E)** ammonia production of *P. peoriae* ZBSF16.

and *P. peoria* ZF390 were in a clade (Supplementary Figure 4A). Additionally, strain ZBSF16 was clearly classified as *P. peoria* in the phylogenetic tree based on the MLSA, and *P. peoria* ZBSF16 was most closely related to strains *P. peoria* ZF390, *P. peoria* HS311 and *P. peoria* HJ-2 (Supplementary Figure 4B). PhyloPhlAn method was performed to verify the evolutionary position. As expected, *P. peoria* ZBSF16 was most closely related to strains *P. peoria* ZF390, *P. peoria* HS311 (Figure 4).

ANI and DDH analysis

Average nucleotide identity (ANI) and DNA–DNA hybridization (DDH) are powerful approaches for evolutionary distance assessment between bacteria at the genomic level, and compared strains usually with ANI values >96% and DDH values ≥70% are regarded as the same species (Richter and Rosselló-Móra, 2009; Jiang et al., 2022). ANI values showed that ZBSF16

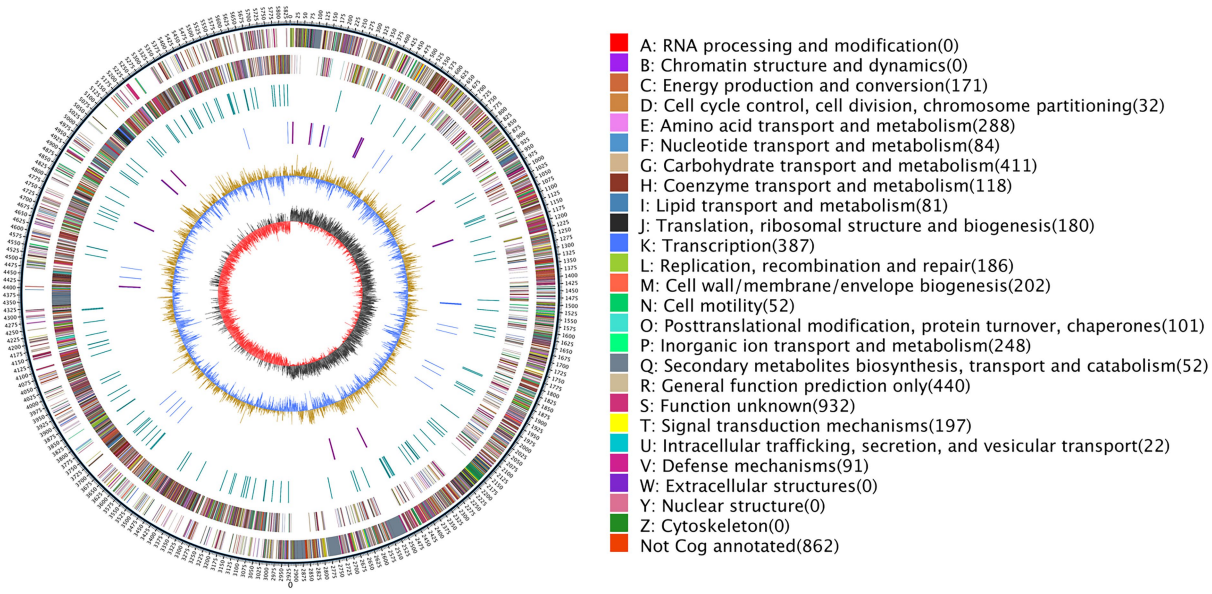

FIGURE 3
Genome map of *Paenibacillus peoriae* ZBSF16. The distribution of the circle from the outside indicates the genome size, forward CDS, reverse CDS, repeat sequence, tRNA (blue), rRNA (purple), GC ratio (yellow and blue indicate regions where the GC ratio is higher than average and lower than average, respectively), and CG skew positive (dark) and negative (red).

between *P. peoria* ZF390, HS311 and HJ-2 were 95.22, 95.23 and 95.24%, respectively. However, the DDH value between ZBSF16 and *P. peoria* HS311 was >70% (Supplementary Figure 4). Obviously, ZBSF16 did not belong to *P. polymyxa* and *P. kribbensis*, according to the lower ANI values (<91%) and DDH values (<50%; Supplementary Figure 5).

Comparison of ZBSF16 with *Paenibacillus peoriae* strains

In comparison, the entire genome size of the four *P. peoriae* strains ranged from 5.84 to 6.19 Mb, the G+C content ranged from 44.99 to 45.62%, and the predicted coding genes ranged from 5,188 to 5,894. Furthermore, the genomes of strains ZF390 and HS311 contained three and one plasmids, respectively. ZBSF16 and HJ-2 both contained one circular chromosome, and the additional genomic features of the six strains are described in Table 1.

To evaluate the evolutionary distance among these sequenced strains in relation to several *Paenibacillus* strains, the genome sequence of ZBSF16 was compared to three sequenced *P. peoriae* strains (ZF390, HS311 and HJ-2), two *P. polymyxa* strains (HY96-2 and SQR-21) and one *P. kribbensis* (AM49) by mauve. The alignments among *Paenibacillus* strains are presented in Figure 5A. Horizontal gene transfer was obviously observed among *Paenibacillus* strains, and the ZBSF16 genome is much more similar to HS311 than to ZF390 within *P. peoriae* strains based on comparative analysis. There were 3,479 conserved genes shared by the seven sequenced strains of the *Paenibacillus* strains, and 3,960 genes were shared within the four sequenced *P. peoriae* strains, including ZBSF16, ZF390, HS311 and HJ-2. In detail, ZBSF16 shared 4,152, 4,143 and 4,135 genes with ZF390, HS311 and HJ-2, respectively. Furthermore, 357 unique genes were present in the genome of *P. peoriae* ZBSF16, genomes with their unique regions are presented in circular images (Figure 5B), and the functions of most unique genes are still unknown. Notably, only 3,772 genes were shared by ZBSF16 and *P. kribbensis* AM49, which is less than those in The *P. polymyxa* strains (Figures 5B,C; Supplementary Figure 6).

Genetic basis for promoting plant growth

IAA is an important phytohormone that controls cell enlargement and tissue differentiation in plants. In this study, ZBSF16 showed a higher IAA biosynthetic capacity (28.67 μg/ml; Figure 2A), and 12 genes related to IAA biosynthesis were identified in strain ZBSF16. Nine genes in the IAA biosynthesis pathway were shared among the four *P. peoriae* strains with homology higher than 90%, except for three genes (*trpE*, *trpG* and *trpCF*) that were not found in strain HJ-2 (Table 2). As a major essential nutrient, phosphorus and nitrogen are necessary for the growth and development of plants, and ZBSF16 exhibits the capability of phosphate solubilization and nitrogen fixation (Figure 2D). Additionally, comparative genome analysis showed 14 genes related to phosphate solubilization in ZBSF16, which was highly similar to

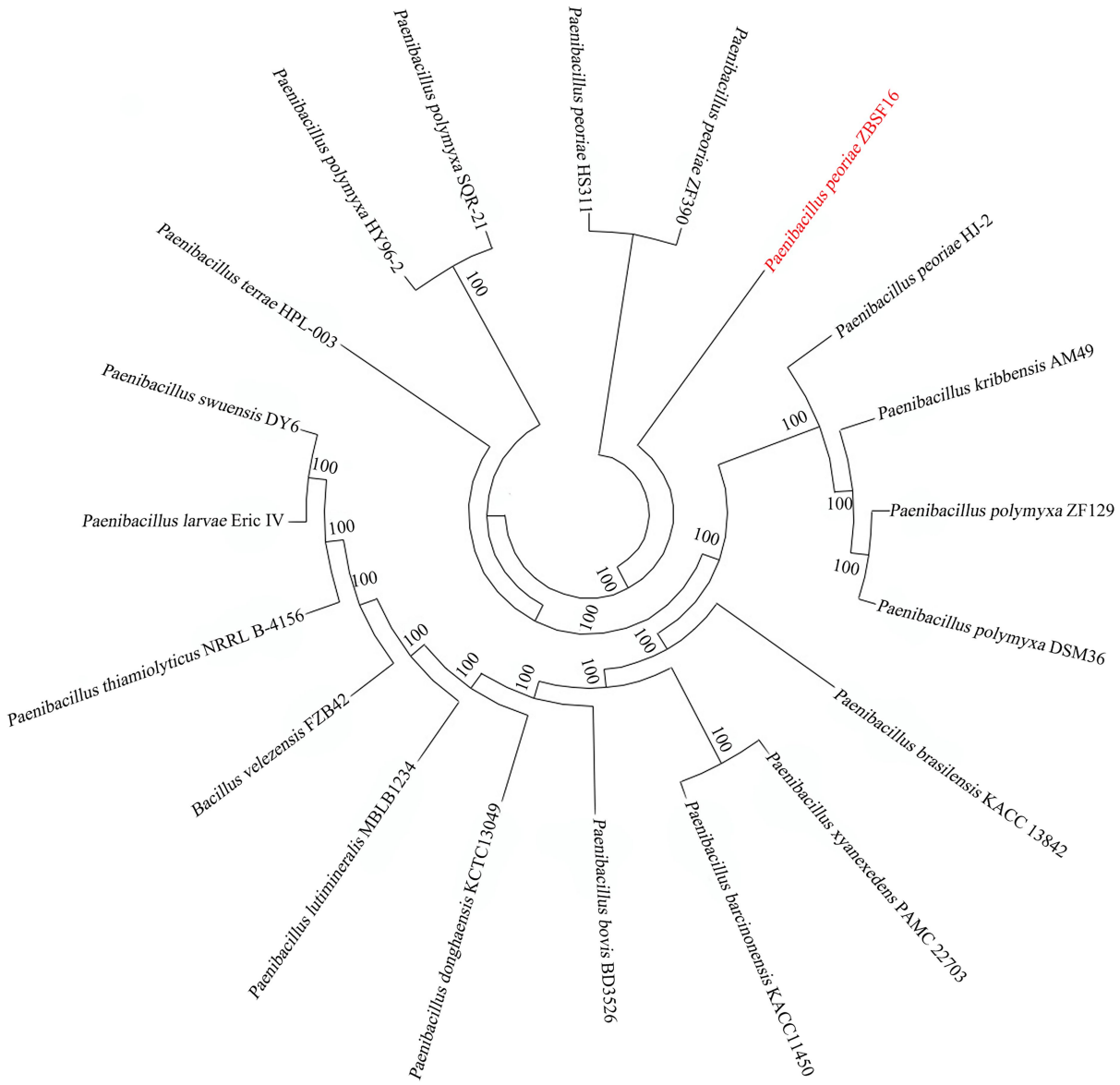

FIGURE 4
Phylogenetic analysis of *Paenibacillus peoriae* ZBSF16 against six other *Paenibacillus* from genomes using PhyloPhlAn 3.0.2.

ZF390, HS311 and HJ-2, and the gene *iap*, which is shared by strain ZF390 and HS311 (Table 2). Furthermore, 15 genes responsible for nitrogen fixation were all found in the genomes of ZBSF16, HS311 and HJ-2, most of which were highly conserved, with sequence identities ranging from 93 to 100%. However, *nifH*, *nifN*, *nifB*, *nifD*, *nifE*, *nifK*, *nifX* and *hesA* were absent in strain ZF390 (Table 2). Meanwhile, 30 genes involved in flagella and 12 genes related to biofilm formation were discovered in strain ZBSF16, and 40 genes involved in flagella (except for *fliD* and *fliS*) and biofilm formation exhibited high conservation (>88%) in ZF390, HS311, HJ-2 and ZBSF16 (Supplementary Tables 5, 6). Quorum sensing (QS) relegated many traits of bacteria, including biofilm formation and colonization. QS is conserved across hundreds of species belonging to the *Paenibacillaceae* family, and seven genes related to QS were identified in *P. peoriae* strains in this study (Supplementary Table 7). Additionally, 11 genes associated with the chemtaxis and two-component systems (TCS), except *CitG* and *DcuS*, were conservative in different strains of *P. peoriae* (Supplementary Table 8).

Genes/gene cluster for antibiotic synthesis and induction of resistance

P. peoriae ZBSF16 showed potent broad-spectrum antifungal activities. Based on the antiSMASH database, 14 clusters related to secondary metabolite synthesis were identified in ZBSF16.

TABLE 1 Genomic features of *Paenibacillus peoriae* ZBSF16 and other *P. peoriae* strains.

Features	*P. peoriae* ZBSF16	*P. peoriae* ZF390	*P. peoriae* HJ-2	*P. peoriae* HS311	*P. polymyxa* HY96-2	*P. polymyxa* SQR21	*P. kribbensis* AM49
Size (bp)	5,839,239	6,193,667	6,001,192	6,006,533	5,745,779	5,828,436	5,778,702
GC content (%)	45.62	44.99	45	45.47	45.60	45.60	46.80
Replicons	Chromosome	Chromosome; Plasmid pPlas1; plasmid pPlas2; plasmid pPlas3	Chromosome	Chromosome; plasmid unnamed	chromosome	chromosome	chromosome
Total genes	5,188	5,894	5,439	5,408	4,955	5,128	5,149
Predicted no. of CDS	4,944	5,749	5,237	5,131	4,799	4,974	5,023
Ribosomal RNA	39	40	39	39	42	39	30
Transfer RNA	109	101	108	99	110	111	92
Other RNA	4	4	N/A	1	4	4	4
CRISPR	4	N/A	9	1	N/A	2	4
Pseudogene	92	115	N/A	138	136	78	184

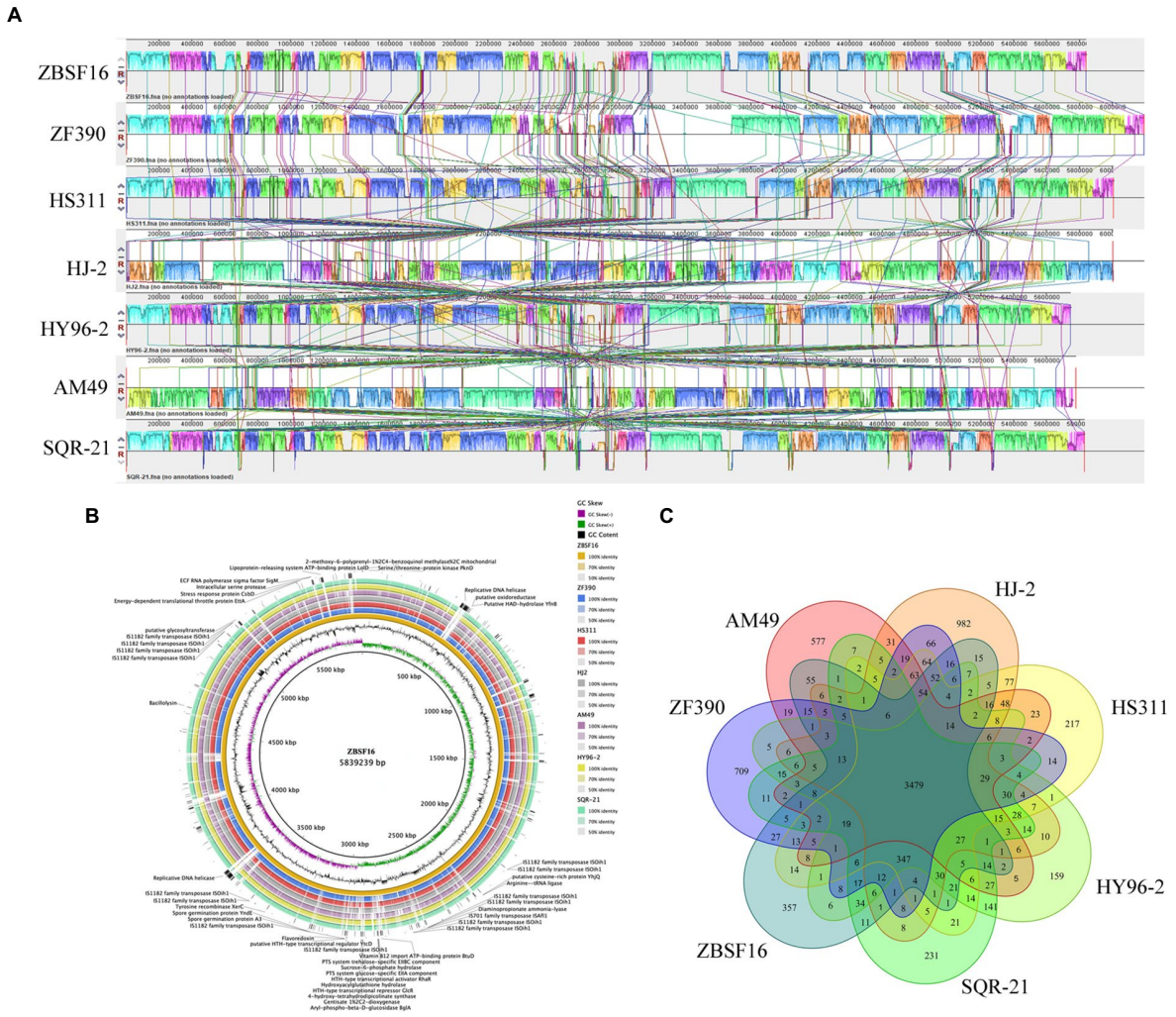

FIGURE 5
Comparison of *Paenibacillus peoriae* ZBSF16 genome sequences against six other *Paenibacillus* genome sequences. **(A)** Synteny analysis of *P. peoriae* ZBSF16 with the *P. peoriae* ZF390, *P. peoriae* HS311, *P. peoriae* HJ-2, *P. polymyxa* HY96-2 *P. polymyxa* SQR-21 and *P. kribbensis* AM49 genomes. Pairwise alignments of the genomes were generated using MAUVE. The genome of strain ZBSF16 was used as the reference genome. Boxes with the same color indicate syntenic regions. Boxes below the horizontal strain line indicate inverted regions. Rearrangements are shown by colored lines. Scale is in nucleotides. **(B)** Pangenome analysis with closely related strains identified the unique genes present in the query genomes that are highlighted in the outermost circle, the strain ZBSF16 as the query genome is placed in the innermost circle. **(C)** Venn diagram showing the number of clusters of orthologous genes shared and unique genes.

TABLE 2 Homolog analysis of genes involved in plant growth promotion in *Paenibacillus peoriae* ZBSF16 and other *P. peoriae* strains.

Genes	Product definition	*P. peoriae* ZBSF16		*P. peoriae* ZF390		*P. peoriae* HS311		*P. peoriae* HJ-2	
		Locus Tag	Protein ID	Protein ID	Homology (%)	Protein ID	Homology (%)	Protein ID	Homology (%)
Indole-3-acetic acid biosynthesis genes									
acoc	Chorismite synthase	MLD56_14630	UMY52826.1	WP_007430826.1	99.49	WP_007430826.1	99.49	NA	100.00
pheB	Chorismite mutase	MLD56_05715	UMY55941.1	WP_013369718.1	99.17	WP_013369718.1	99.17	NA	99.72
aroF	3-Deoxy-7-phosphoheptulonate synthase/chorismite mutase	MLD56_07785	UMY56321.1	WP_017426658.1	97.97	WP_013309435.1	99.71	NA	99.71
	Tryptophan-rich sensory protein	MLD56_02865	UMY55416.1	WP_014279544.1	89.64	WP_014279544.1	89.64	NA	96.02
trpA	Tryptophan synthase subunit alpha	MLD56_14590	UMY52818.1	WP_013310712.1	98.13	WP_013310712.1	98.13	NA	98.51
trpB	Tryptophan synthase subunit beta	MLD56_14595	UMY52819.1	WP_014282083.1	96.73	WP_014282083.1	96.73	NA	99.25
trpS	Tryptophan--tRNA ligase	MLD56_20860	UMY53981.1	WP_016819987.1	98.18	WP_013311859.1	97.26	NA	96.66
trpC	Indole-3-glycerol phosphate synthase TrpC	MLD56_14605	UMY52821.1	WP_017427551.1	94.30	WP_013371667.1	95.06	NA	97.35
trpD	Anthranilate phosphoribosyltransferase	MLD56_14610	UMY52822.1	WP_013371668.1	95.98	WP_013371668.1	95.98	NA	97.70
trpE	Anthranilate synthase component I	MLD56_14615	UMY52823.1	WP_007430823.1	97.87	WP_007430823.1	97.87	NA	NA
trpG	Glutamine amidotransferase	MLD56_18435	UMY53538.1	WP_007431477.1	91.83	WP_007431477.1	91.83	NA	NA
trpCF	Phosphoribosylanthranilate isomerase	MLD56_14600	UMY52820.1	WP_019687860.1	91.23	WP_019687860.1	91.23	NA	NA
ipdC	Thiamine pyrophosphate-binding protein	MLD56_00395	UMY54998.1	WP_007428062.1	100.00	WP_007428062.1	100.00	NA	98.75
Phosphate solubilization genes									
phoN	Phosphatase PAP2 family protein	MLD56_05880	UMY55971.1	WP_010347599.1	87.07	WP_013309103.1	97.79	NA	99.97
iap	Aminopeptidase	NA	NA	WP_013309451.1	NA	WP_013309451.1	NA	NA	NA
phoA	Alkaline phosphatase	MLD56_07175	UMY56211.1	WP_019686611.	93.21	WP_013309329.1	97.48	NA	97.03
phnE	Phosphonate ABC transporter, permease protein PhnE	MLD56_21880	UMY54161.1	WP_016324733.1	99.30	WP_016324733.1	99.30	NA	99.30
phnE	Phosphonate ABC transporter, permease protein PhnE	MLD56_21885	UMY54162.1	WP_016820374.1	98.87	WP_016820374.1	98.87	NA	99.25
phnD	Phosphonate ABC transporter substrate-binding protein	MLD56_21870	UMY54159.1	WP_010344588.1	96.89	WP_010344588.1	96.89	NA	99.69
phnC	Phosphonate ABC transporter ATP-binding protein	MLD56_21875	UMY54160.1	WP_020723499.1	98.83	WP_020723499.1	98.83	NA	97.66
pstS	Phosphate ABC transporter substrate-binding protein PstS	MLD56_08410	UMY56438.1	WP_016819622.1	98.70	WP_016819622.1	98.70	NA	99.35
pstC	Phosphate ABC transporter permease PstC	MLD56_08415	UMY56439.1	WP_013370343.1	99.66	WP_053325097.1	99.33	NA	99.68
pstA	Phosphate ABC transporter permease PstA	MLD56_08420	UMY57301.1	WP_013309592.1	99.66	WP_013309592.1	99.66	NA	100
pstB	Phosphate ABC transporter ATP-binding protein PstB	MLD56_08425	UMY56440.1	WP_013370344.1	97.86	WP_013370344.1	97.86	NA	99.29

(Continued)

TABLE 2 Continued

Genes	Product definition	*P. peoriae* ZBSF16		*P. peoriae* ZF390		*P. peoriae* HS311		*P. peoriae* HJ-2	
		Locus Tag	Protein ID	Protein ID	Homology (%)	Protein ID	Homology (%)	Protein ID	Homology (%)
pstB	Phosphate ABC transporter ATP-binding protein PstB	MLD56_08490	UMY56453.1	WP_007429703.1	96.83	WP_007429703.1	96.83	NA	99.21
phoU	Phosphate signaling complex protein PhoU	MLD56_08495	UMY56454.1	WP_016819636.1	96.80	WP_016819636.1	96.80	NA	100
phoN	Phosphatase PAP2 family protein	MLD56_05880	UMY55971.1	WP_010347599.1	87.07	WP_013309103.1	97.79	NA	99.97
Nitrate transport and nitrate/nitrite reduction									
narI	Nitrate reductase gamma subunit	MLD56_17955	UMY57358.1	WP_013372381.1	95.59	WP_013311337.1	96.93	NA	96.37
narJ	Nitrate reductase molybdenum cofactor assembly chaperone	MLD56_17960	UMY57359.1	WP_010345152.1	95.72	WP_010345152.1	95.72	NA	98.40
narH	Nitrate reductase beta subunit	MLD56_17965	UMY53450.1	WP_016324613.1	92.25	WP_014282714.1	99.43	NA	99.62
narG	Nitrate reductase alpha subunit	MLD56_17970	UMY53451.1	WP_007431447.1	93.95	WP_007431447.1	93.95	NA	98.62
narK	MFS transporter NNP family nitrate/nitrite transporter	MLD56_17930	UMY53446.1	WP_013311332.1	98.18	WP_013311332.1	98.18	NA	98.63
Niterate transport and reduction									
nirD	Nitrite reductase small subunit NirD	MLD56_03440	UMY55525.1	WP_017428677.1	93.58	WP_017428677.1	93.58	NA	96.33
nirC	Nitrite transporter NirC	MLD56_04985	UMY55818.1	WP_016819917.1	98.47	WP_016819917.1	98.47	NA	90.46
nirB	Nitrite reductase large subunit NirB	MLD56_03435	UMY55524.1	WP_016818403.1	97.65	WP_016818403.1	97.65	NA	98.27
amtB	Ammonium transporter Amt family	MLD56_09035	UMY56554.1	WP_010348916.1	96.79	WP_007429827.1	97.00	NA	99.79
nifH	Nitrogenase iron protein NifH	MLD56_05440	UMY55888.1	NA	NA	WP_007429042.1	98.26	NA	100
nifN	Nitrogenase molybdenum-iron protein NifN	MLD56_05460	UMY55892.1	NA	NA	WP_014280100.1	98.16	NA	93.08
nifB	Nitrogenase fixation protein NifB	MLD56_05435	UMY55887.1	NA	NA	WP_014280095.1	95.79	NA	97.35
nifD	Nitrogenase fixation protein NifD	MLD56_05445	UMY55889.1	NA	NA	WP_007429043.1	97.93	NA	95.69
nifU	Nitrogenase fixation protein NifU	MLD56_21125	UMY54031.1	WP_013373004.1	100	WP_013373004.1	100	NA	100
nifE	Nitrogenase molybdenum-cofator synthesis protein NifE	MLD56_05455	UMY55891.1	NA	NA	WP_014280099.1	96.91	NA	99.56
nifK	Nitrogenase molybdenum-iron protein subunit beta	MLD56_RS05450	UMY55890.1	NA	NA	WP_007429044.1	97.45	NA	97.45
nifX	Nitrogen fixation protein NifX	MLD56_RS05465	UMY55893.1	NA	NA	WP_014280101.1	97.67	NA	96.90
hesA	HesA/MoeB/ThiF family protein	MLD56_RS05470	UMY55894.1	NA	NA	WP_014280102.1	100.00	NA	97.24

NA, not available.

Among these gene clusters, three clusters (Cluster 1 related to fusaricidinB, Cluster 8 related to cyclic-lactone-autoinducer, and Cluster 9 related to tridecaptin) were shared among the four *P. peoriae* strains, the two *P. polymyxa* and *P. kribbensis*; the functions of fusaricidin B and tridecaptin were antifungal and antibacterial, respectively. Cluster 3 related to paenibacillin was specific and only found in strain ZBSF16, which was a kind of lantibiotic. In addition, polymyxin and paenilan did not appear in *P. kribbensis*, paeninodin could not be detected in *P. polymyxa*, and genes related to Cluster 17 encoding the biosynthesis of paenilan, pelgipeptin, aurantinin and so on were not found in ZBSF16 (Figure 6; Supplementary Table 9).

The resistance inducer biosynthesis gene cluster, including 11 genes related to ISR and 3 genes involved in PAMP-triggered immunity (PTI), was analyzed in strain ZBSF16, which is highly conserved in the selected *P. peoriae* strains (>79% identity). The genes *alsS* and *budA* were identified in strain ZBSF16, which showed a lower similarity to ZF390. The gene *flgL* involved in PTI of plants showed higher similarity to ZF390, and it could not be identified in strains HS311 and HJ-2 (Table 3).

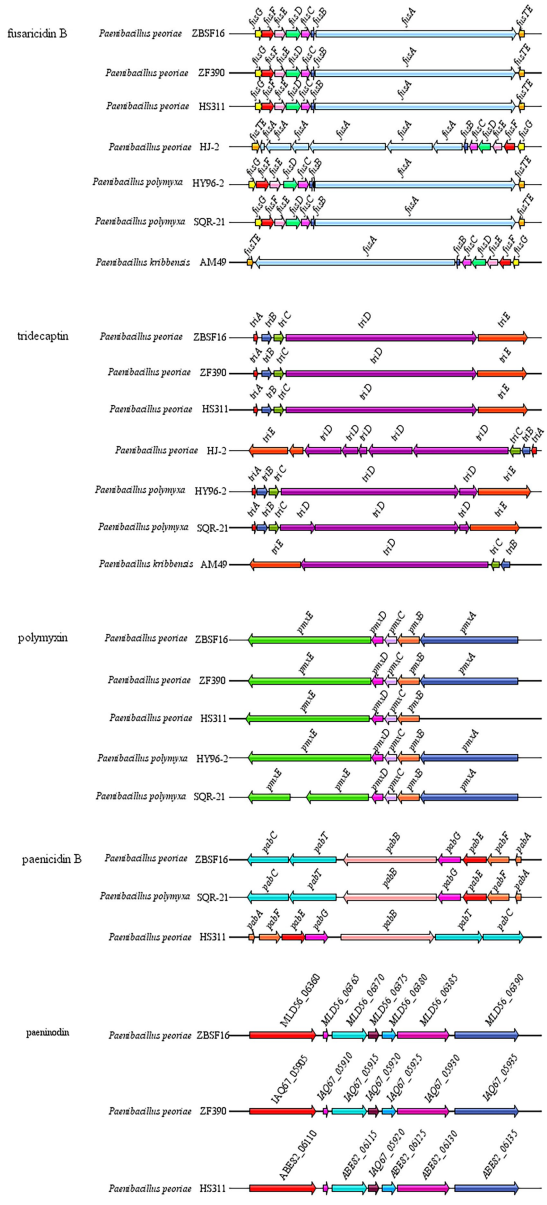

FIGURE 6
Comparison of antibiotic synthesis clusters of *Paenibacillus* strains. Antibiotic synthesis clusters were identified using antiSMASH, and gene cluster intraspecific genes were compared.

Discussion

Paenibacillus is widely distributed in a variety of environments, including wetlands, meadow soil, desert sand, oceans, wheat soil rhizospheres, cucumber greenhouses and infected honeybees (Jeon et al., 2009; Wang et al., 2013; Ahn et al., 2014). The genus *Paenibacillus* is reported to have the ability to promote the growth of many plants, such as maize, wheat, tomato, and pumpkin (Hao and Chen, 2017; Dixit et al., 2018). The genome size of *Paenibacillus* species ranges from 3.02 Mbp to 8.82 Mbp. As a member of 200 species in *Paenibacillus*, *P. peoriae* was described to play a role in promoting the growth of plants by some studies in the past and was confirmed in this study (Figure 2), with a genome size of 5.74–6.19 Mbp and GC content of 44.99–45.62% (Table 1). *P. peoriae* was close to *P. polymyxa* and *P. kribbensis* in terms of evolutionary status, and ZBSF16 was identified and confirmed to belong to *P. peoriae* by ANI and DDH. Compared to *P. peoriae* HJ-2, which presented antagonistic activity against *Fusarium* spp., ZBSF16 had a broad antifungal and antibacterial spectrum, which could protect against 10 species of fungi and 2 species of bacteria.

Many PGPRs, including *Bacillus*, *Rahnella*, *Pseudomonas*, *Klebsiella*, *Agrobacterium* and *Paenibacillus* sp. can produce IAA to stimulate the growth of plants, and *Paenibacillus* nonsymbiotic bacteria yielded high concentrations of IAA (in the range of 4.90–0.19 IAA/mg biomass; Shokri and Emtiazi, 2010; Trinh et al., 2018). *P. polymyxa*, *P. borealis*, and *P. terrae* showed the secretion of a significant amount of IAA, but no *P. graminis* had the ability to produce IAA (Navarro-Noya et al., 2012; Kim et al., 2017). *P. peoriae* HJ-2 isolated from soil significantly promoted the growth of *P. polyphylla*, and *P. peoriae* ZBSF16 for the first time was used to describe the ability to synthesize IAA and promote the growth of grape, with IAA production of 28.67 μg ml^{-1}. The various pathways for IAA biosynthesis include tryptophan (Trp), tryptamine (Tam), indole-3-pyruvic acid (IPyA) and indole-3-acetamide (IAAm) pathways, and the IPyA pathway was suggested in *Paenibacillus* because of the absence of tryptophan monooxygenase or indole-3-acetamide hydrolase (Mano and Nemoto, 2012; Xie et al., 2016). In addition, the *ipdC* gene, encoding a key enzyme in the IPyA pathway, is shared in all *Paenibacillus* (Xie et al., 2016). In this study, *ipdC* homologies were present in all sequenced *P. peoriae*, which demonstrated that *P. peoriae* may rely on the IPyA pathway for IAA synthesis.

TABLE 3 Genes related to synthesis resistance inducer in *Paenibacillus peoriae* ZBSF16 and other *P. peoriae* strains.

Genes	Resistance inducers	Plant resistance type	Product definition	*P. peoriae* ZBSF16		*P. peoriae* ZF390		*P. peoriae* HS311		*P. peoriae* HJ-2	
				Locus tag	Protein ID	Protein ID	Homology (%)	Protein ID	Homology (%)	Protein ID	Homology (%)
alsS	2,3-Butanediol	ISR	Acetolactate synthase	MLD56_10755	UMY56883.1	WP_003206007.1	75.92	WP_013310040.1	95.43	NA	95.96
budA/ alsD	2,3-Butanediol	ISR	Acetolactate decarboxlase	MLD56_10750	UMY56882.1	WP_000215036.1	68.25	WP_016821069.1	97.18	NA	97.58
bdh	2,3-Butanediol	ISR	2，3-Butanediol dehydrogenase	MLD56_18150	UMY53485.1	WP_019688213.1	98.29	WP_013311373.1	99.43	NA	96.00
ilvN	2,3-Butanediol	ISR	Acetolactate synthase small subunit	MLD56_07545	UMY56280.1	WP_007429525.1	98.76	WP_013309386.1	99.38	NA	99.38
metH	Methanethio	ISR	Methionine synthase	MLD56_13735	UMY52659.1	WP_010345928.1	96.68	WP_010345928.1	96.68	NA	80.94
metE	Methanethio	ISR	5-Methyltetrahydro-pteroyltriglutamate-homocysteine S-methyltransferase	MLD56_24010	UMY54554.1	WP_013373554.1	93.47	WP_013312443.1	97.73	NA	96.50
ispF	Isoprene	ISR	2-C-methyl-D-erythritol 2,4-cyclodiphosp- hata	MLD56_22685	UMY54300.1	WP_000488386.1	100	WP_007432605.1	98.10	NA	98.73
ispE	Isoprene	ISR	4-(cytidine 5′-diphospho)-2-C-methyl-D-erythritol kinase	MLD56_00170	UMY54955.1	WP_013308121.1	99.65	WP_013308121.1	99.65	NA	98.94
gcpE	Isoprene	ISR	Flavodoxin- dependent (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthae	MLD56_19660	UMY53756.1	WP_010348073.1	98.92	WP_010348073.1	98.92	NA	100
lytB	Isoprene	ISR	4-hydroxy-3 -methylbut-2-enyl diphosphate reductase	MLD56_07780	UMY56320.1	WP_013309434.1	99.37	WP_013309434.1	99.37	NA	99.00
fni	Isoprene	ISR	Type 2 isopentenyl-diphosphate Delta-isomerase	MLD56_23495	UMY54455.1	WP_017427145.1	91.80	WP_013312347.1	96.45	NA	96.72
guaB	Peptidoglycan	PTI	carboxypeptidase	MLD56_00435	UMY55002.1	WP_017427215.1	97.94	WP_017427215.1	97.94	NA	97.94
flgL	*Flagenllin*	PTI	flagellin	MLD56_23175	UMY54392.1	WP_016822919.1	96.44	N/A	N/A	NA	NA
tuf	*EF-Tu*	PTI	Elongation factor Tu	MLD56_22575	UMY54278.1	WP_017815361.1	96.21	WP_017815361.1	96.21	NA	98.99

NA, not available.

10.3389/fmicb.2022.975344

P. polymyxa strains have long been known to solubilize phosphate, which carries the *phn* genes (*phnABCDEWXM*) responsible for solubilizing organic phosphate (Zhou et al., 2020; Soni et al., 2021). The *phnB* gene was absent in some species of *Paenibacillus*, including *P. beijingensis* 1–18, *P. peoriae* KCTC 3763 and *P. terrae* HPL-003 (Jeong et al., 2012; Shin et al., 2012; Li L. et al., 2019). In this study, phnA and phnB were not found in the genomes of *P. peoriae*. The Pst (phosphate-specific transport) system is a major transport system for Pi. The *pst* operon of *Paenibacillu* is composed of *pstS*, *pstC*, *pstA* and *pstB* (Li et al., 2020), and the four *pst* genes were all present in *P. peoriae* ZBSF16, which contribute to the solubilization of phosphate. It has been reported that *Rahnella aquatilis* ZF7 can produce acid, which may have high activity for solubilizing organic phosphate (Yuan et al., 2020). A higher phosphate solubilization ability of *P. peoriae* ZBSF16 was observed, although the pH value of ZBSF16 remained alkaline when cultured.

Nitrogen fixation is one characteristic of the genus *Paenibacillus*, and more than 20 species of the genus *Paenibacillus* can fix nitrogen (Grady et al., 2016; He et al., 2021). Nitrogen fixation is mainly catalyzed by Mo-nitrogenase, and the *nif* gene cluster (*nif*B, *nif*H, *nif*D, *nif*K, *nif*E, *nif*N, *nif*X, *hes*A and *nif*V) encoding Mo-nitrogenase is shared in N_2-fixing *Paenibacillus* strains (Xie et al., 2014). When the nif gene cluster is lost, non-N_2-fixing strains are produced, such as *P. peoriae* KTCT 3763, *P. polymyxa* SC2 and *P. polymyxa* E681 (Kim et al., 2010; Ma et al., 2011). When acquiring the *vnf* and *anf* genes, strains of *vnf*HDGKEN encoding V-nitrogenase and *anf*HDGK encoding Fe-nitrogenase appeared, such as *P. azotofixans* ATCC 35681 and *P. forsythia* T98 (Xie et al., 2014, 2016). Most likely due to gene loss, the *nif*V gene was absent in the gene cluster in *P. peoriae* ZBSF16, but ZBSF16 retained its nitrogen-fixing capacity.

The genus *Paenibacillus* is known for its ability to produce antibacterial metabolites, including fusaricidins, pelgipeptin, surfactins and polymyxins (Grady et al., 2016). The antibacterial metabolites of *P. polymyxa* ZF129 and *P. polymyxa* ZF197 were significantly different, but paeninodin, fusaricidin, paenibacterin and tridecaptin were shared by the two strains (Li et al., 2020). In our study, fusaricidin B, tridecaptin, polymyxin and paenicidin B were found in *P. peoriae* ZBSF16, which contribute to its strong antipathogenic activities. In addition, fusaricidin B, tridecaptin and polymyxin were conserved in *P. peoriae*, *P. polymyxa* and *P. kibbensis*, which were also shared in *P. polymyxa* ZF129 and *P. polymyxa* ZF197. The antifungal mechanism of fusaricidin is permeabilization and disruption of cell membranes (Jiang et al., 2022), which may be one of the reasons why *P. peoriae* ZBSF16 showed a broad antifungal spectrum.

ISR is the form of induced resistance wherein plant defenses are preconditioned by prior treatment that results in resistance against subsequent challenge by a pathogen or parasite (Choudhary et al., 2007). ISR can increase systemic levels of the plant hormone salicylic acid (SA) and trigger the jasmonic acid/ethylene pathway. *Paenibacillus*-mediated ISR has been demonstrated against fungi (e.g., *C. truncatum*, *C. orbiculare* and *F. oxysporum*) and bacteria (e.g., *Xanthomonas axonopodis* pv. *vesicatoria*, *Erwinia carotovora* subsp. *carotovora*) in pepper, cucumber, banana, and *Arabidopsis thaliana* (Sang et al., 2014; Nakkeeran et al., 2021; Yadav M. et al., 2021). Nine genes involved in ISR were explored in *P. polymyxa*, with higher sequence identity (>95%) in different strains, while key genes associated with volatile organic compounds (2,3-butanediol, methanethiol and isoprene) were contained (Li et al., 2020). A total of 12 genes related to ISR were found in *P. peoriae* ZBSF16, which were highly similar to those in *P. polymyxa* (homology >99%). The results demonstrated that *P. peoriae* and *P. polymyxa* could induce similar systemic resistance in plants.

Conclusion

P. peoriae ZBSF16 showed broad-spectrum antagonistic activities against 12 plant pathogens and exhibited obvious biocontrol effects against grape white rot disease. The aim of this study was to reveal the plant growth-promoting and biocontrol mechanisms of *P. peoriae*. Whole-genome analysis and phylogenetic analysis revealed that ZBSF16 belongs to *P. peoriae* and is closely related to *P. peoriae* ZF390. Comparative analysis of the genome of *P. peoriae* ZBSF16 with other *Paenibacillus* spp. indicated that ZBSF16 harbored many genes related to IAA production, nitrogen fixation, phosphate solubilization, biofilms and flagella, which have been proven to be beneficial to plant growth. In addition, genes associated with antibiotic synthesis and induction of resistance were identified. Overall, the features of *P. peoriae* ZBSF16 make it a high-probability biocontrol agent and biofertilizer, and these results will contribute to in-depth research on the mechanisms of plant growth promotion and biocontrol.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: NCBI GenBank - CP092831.1.

Author contributions

LY, YW, and XY conceived and designed the experiments. XY, HJ, TL, PL, and XJ performed the experiments and analyzed the data. LY and YW wrote the manuscript. TL, XJ, PL, and HJ revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by Shandong Provincial Natural Science Foundation (ZR2021QC131), Innovation Project of Shandong Academy of Agricultural Sciences (CXGC2022E15), and Shandong Academy of Grape Guide Fund (SDAG2021B06, SDAG2021B10, and SDAG2021B02).

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.975344/full#supplementary-material

SUPPLEMENTARY FIGURE 1
General characteristics of *Paenibacillus peoriae* ZBSF16. **(A)** Image of ZBSF16 colony morphology. **(B)** Image of ZBSF16 cells *via* scanning electron microscopy. **(C)** Growth dynamics and pH change of *P. peoriae* ZBSF16. Bars plot the means ± standard deviation of three replicate experiments. P **(D)** Production of protease. **(E)** Cellulose degradation. **(F)** Production of lipase. Determination of NaCl **(G)** and pH **(H)** tolerance capabilities of *P. peoriae* ZBSF16.

SUPPLEMENTARY FIGURE 2
Antagonistic assay and biocontrol effect of *Paenibacillus peoriae* ZBSF16. **(A)** Colony radius and inhibition rate of each microorganism. Bars plot the means ± standard deviation of three replicate experiments. *Coniella vitis* (CV). *Gloeosporium fructigrum* (GF). *Pestalotiopsis clavispora* (Pc). *Alternaria viticola* (Av). *Diaporthe eres* (DE). *Fusarium oxysporum* (Fo). *Botrytis cinerea* (BC). *Botryosphaeria dothidea* (BD). *Aspergillus niger* (AN). *Fusarium graminearum* (FG). *Fusarium pseudograminearum* (FP). *Allorhizobium vitis* (ALV). **(B,C)** Incidence, disease index and control efficiency of *P. peoriae* ZBSF16. (a1, b1) Inoculated with *C. vitis;* (a2, b2) LB broth; (a3, b3) sterile water; (a4, b4) culture of ZBSF16; (a5, b5) inoculated with *C. vitis* 24h after inoculation with the culture of ZBSF16; (a6, b6) inoculated culture of ZBSF16 24h after inoculation with *C. vitis*. **(D)** Disease symptoms and growth state of *Vitis vinifera* (cv. Red globe) inoculated with strain ZBSF16. **(E)** The infection rate and disease index of grape white rot on *Vitis vinifera* (cv. Red globe) inoculated with strain ZBSF16. CK plants were treated with sterile water. Different letters above the bars denote a significant difference at $p<0.05$ according to Duncan's multi-range test.

SUPPLEMENTARY FIGURE 3
Determination of antibiotic resistance of *Paenibacillus peoriae* ZBSF16. **(A)** Survival of *P. peoriae* ZBSF16 treated with different antibiotics. Spectinomycin (Spe), streptomycin (Str), ampicillin (Amp), vancomycin (Van), kanamycin (Kan), gentamycin (Gen), chloramphenicol (Chl), tetracycline (Tet) and rifampicin (Rif). **(B)** Minimum inhibitory concentration (MIC) of spectinomycin for strain ZBSF16. **(C)** Minimum bactericidal concentration (MBC) of spectinomycin for strain ZBSF16. **(D)** Hemolysis assay of ZBSF16. **(E)** Siderophores production of *P. peoriae* ZBSF16. **(F)** Population dynamics of *P. peoriae* ZBSF16 in the rhizosphere soil of grape.

SUPPLEMENTARY FIGURE 4
(A) Phylogenetic tree for *P. peoriae* ZBSF16 and the genus *Paenibacillus* based on 16S rRNA (*Bacillus velezensis* FZB42 was used as an outgroup). **(B)** Phylogenetic tree of *Paenibacillus peoriae* ZBSF16 among other *Paenibacillus* species. The phylogenetic tree was constructed based on five housekeeping genes (16S rRNA, *gyrB*, *rpoD*, *rho*, and *pgk*) according to the aligned gene sequences using the maximum likelihood method in MEGA 6.0. Bootstrap values (1,000 replicates) are shown at the branch points. The scale bar indicates 0.05 nucleotide substitutions per nucleotide position. GenBank accession numbers associated with the housekeeping loci of all strains can be found in **Supplementary Table 1**.

SUPPLEMENTARY FIGURE 5
ANI **(A)** and DDH **(B)** value matrix heatmap between *Paenibacillus peoriae* ZBSF16 and six other *Paenibacillus* genome sequences.

SUPPLEMENTARY FIGURE 6
Venn diagram showing the number of clusters of orthologous genes shared and unique genes.

References

Abriouel, H., Franz, C. M., Ben Omar, N., and Gálvez, A. (2011). Diversity and applications of *Bacillus* bacteriocins. *FEMS Microbiol. Rev.* 35, 201–232. doi: 10.1111/j.1574-6976.2010.00244.x

Ahemad, M. (2015). Phosphate-solubilizing bacteria-assisted phytore mediation of metalliferous soils: a review. *3 Biotech* 5, 111–121. doi: 10.1007/s13205-014-0206-0

Ahn, J. H., Kim, B. C., Kim, B. Y., Kim, S. J., Song, J., Kwon, S. W., et al. (2014). *Paenibacillus cucumis* sp. nov. isolated from greenhouse soil. *J. Microbiol.* 52, 460–464. doi: 10.1007/s12275-014-4071-7

Alikhan, N. F., Petty, N. K., Zakour, N. L., and Ben Beatson, S. A. (2011). BLAST ring image generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12, 1–10. doi: 10.1186/1471-2164-12-402

Ash, C., Priest, F. G., and Collins, M. D. (1993). Molecular identification of rRNA group 3 bacilli (Ash, Farrow, Wallbanks and Collins) using a PCR probe test. Proposal for the creation of a new genus *Paenibacillus*. *Antonie Van Leeuwenhoek* 64, 253–260. doi: 10.1007/BF00873085

Asnicar, F., Thomas, A. M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., et al. (2020). Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using Phylo PhlAn 3.0. *Nat. Commun.* 11, 2500. doi: 10.1038/s41467-020-16366-7

Baindara, P., Chaudhry, V., Mittal, G., Liao, L. M., Matos, C. O., Khatri, N., et al. (2015). Characterization of the antimicrobial peptide penisin, a class Ia novel lantibiotic from *Paenibacillus* sp. strain A3. *Antimicrob. Agents Chemother.* 60, 580–591. doi: 10.1128/AAC.01813-15

Brillard, J., Ribeiro, C., Boemare, N., Brehelin, M., and Givaudan, A. (2001). Two distinct hemolytic activities in *Xenorhabdus nematophila* are active against immunocompetent insect cells. *Appl. Environ. Microb.* 67, 2515–2525. doi: 10.1128/AEM.67.6.2515-2525.2001

Chethana, K. W. T., Zhou, Y., Zhang, W., Liu, M., Xing, Q. K., Li, X. H., et al. (2017). *Coniella vitis* sp. nov. is the common pathogen of white rot in Chinese vineyards. *Plant Dis.* 101, 2123–2136. doi: 10.1094/pdis-12-16-1741-re

Choudhary, D. K., Prakash, A., and Johri, B. N. (2007). Induced systemic resistance (ISR) in plants: mechanism of action. *Indian J. Microbiol.* 47, 289–297. doi: 10.1007/s12088-007-0054-2

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1403. doi: 10.1101/gr.2289704

DasGupta, S. M., Khan, N., and Nautiyal, C. S. (2006). Biologic control ability of plant growth-promoting *Paenibacillus lentimorbus* NRRL B-30488 isolated from milk. *Curr. Microbiol.* 53, 502–505. doi: 10.1007/s00284-006-0261-9

Dixit, R., Agrawal, L., Singh, S. P., Prateeksha, Singh, P. C., Prasad, V., et al. (2018). *Paenibacillus lentimorbus* induces autophagy for protecting tomato from *Sclerotium rolfsii* infection. *Microbiol. Res.* 215, 164–174. doi: 10.1016/j.micres.2018.07.008

Elhaissoufi, W., Khourchi, S., Ibnyasser, A., Ghoulam, C., Rchiad, Z., Zeroual, Y., et al. (2020). Phosphate solubilizing rhizobacteria could have a stronger influence on wheat root traits and aboveground physiology than rhizosphere *P. solubilization*. *Front. Plant Sci.* 11, 979. doi: 10.3389/fpls.2020.00979

Emmanouil, A., Markakis Sotirios, E., Tjamos Polymnia, P., Antoniou Epameinondas, J., and Paplomatas Eleftherios, C. T. (2016). Biological control of *Verticillium* wilt of olive by *Paenibacillus alvei*, strain K165. *Biol. Control* 61, 293–303. doi: 10.1007/s10526-015-9669-0

Garcia-Seco, D., Zhang, Y., Gutierrez-Mañero, F. J., Martin, C., and Ramos-Solano, B. (2015). Application of *Pseudomonas fluorescens* to blackberry under field conditions improves fruit quality by modifying flavonoid metabolism. *PLoS One* 10:e0142639. doi: 10.1371/journal.pone.0142639

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* 57, 81–91. doi: 10.1099/ijs.0.64483-0

Grady, E. N., MacDonald, J., Liu, L., Richman, A., and Yuan, Z. C. (2016). Current knowledge and perspectives of *Paenibacillus*: a review. *Microb. Cell Factories* 15, 203. doi: 10.1186/s12934-016-0603-7

Hao, T., and Chen, S. (2017). Colonization of wheat, maize and cucumber by *Paenibacillus polymyxa* WLY78. *PLoS One* 12:e0169980. doi: 10.1371/journal.pone.0169980

Hashem, A., Tabassum, B., and Fathi Abd Allah, E. (2019). *Bacillus subtilis*: A plant-growth promoting rhizobacterium that also impacts biotic stress. *Saudi. J. Biol. Sci.* 26, 1291–1297. doi: 10.1016/j.sjbs.2019.05.004

He, X., Li, Q., Wang, N., and Chen, S. (2021). Effects of an EPS biosynthesis gene cluster of *Paenibacillus polymyxa* WLY78 on biofilm formation and nitrogen fixation under aerobic conditions. *Microorganisms*. 9, 289. doi: 10.3390/microorganisms9020289

Jeon, C. O., Lim, J. M., Lee, S. S., Chung, B. S., Park, D. J., Xu, L. H., et al. (2009). *Paenibacillus harenae* sp. nov., isolated from desert sand in China. *Int. J. Syst. Evol. Microbiol.* 59, 13–17. doi: 10.1099/ijs.0.65664-0

Jeong, H., Choi, S. K., Park, S. Y., Kim, S. H., and Park, S. H. (2012). Draft genome sequence of *Paenibacillus peoriae* strain KCTC 3763t. *J. Bacterio.* 194, 1237–1238. doi: 10.1128/JB.06577-11

Ji, T., Languasco, L., Li, M., and Rossi, V. (2021). Effects of temperature and wetness duration on infection by *Coniella diplodiella*, the fungus causing white rot of grape berries. *Plan. Theory* 10, 1696. doi: 10.3390/plants10081696

Jiang, A., Zou, C., Xu, X., Ke, Z., Hou, J., Jiang, G., et al. (2022). Complete genome sequence of biocontrol strain *Paenibacillus peoriae* HJ-2 and further analysis of its biocontrol mechanism. *BMC Genomics* 23, 161. doi: 10.1186/s12864-022-08330-0

Kim, J. F., Jeong, H., Park, S. Y., Kim, S. B., Park, Y. K., Choi, S. K., et al. (2010). Genome sequence of the polymyxin-producing plant-probiotic rhizobacterium *Paenibacillus polymyxa* E681. *J. Bacteriol.* 192, 6103–6104. doi: 10.1128/JB.00983-10

Kim, A. Y., Shahzad, R., Kang, S. M., Khan, A. L., Lee, S. M., Park, Y. G., et al. (2017). *Paenibacillus terrae* AY-38 resistance against *Botrytis cinerea* in *Solanum lycopersicum* L. plants through defence hormones regulation. *J. Plant Interact.* 12, 244–253. doi: 10.1080/17429145.2017.1319502

Kumar, R., Acharya, V., Mukhia, S., Singh, D., and Kumar, S. (2019). Complete genome sequence of *Pseudomonas frederiksbergensis* ERDD5: 01 revealed genetic bases for survivability at high altitude ecosystem and bioprospection potential. *Genomics* 111, 492–499. doi: 10.1016/j.ygeno.2018.03.008

Li, J. Y., Gao, T. T., and Wang, Q. (2020). Comparative and functional analyses of two sequenced *Paenibacillus polymyxa* genomes provides insights into their potential genes related to plant growth-promoting features and biocontrol mechanisms. *Front. Genet.* 11:564939. doi: 10.3389/fgene.2020.564939

Li, Y., Li, Y., Zhang, H., Wang, M., and Chen, S. (2019). Diazotrophic *Paenibacillus beijingensis* BJ-18 provides nitrogen for plant and promotes plant growth, nitrogen uptake and metabolism. *Front. Microbiol.* 10, 1119. doi: 10.3389/fmicb.2019.01119

Li, L., Yuan, L., Shi, Y., Xie, X., Chai, A., Wang, Q., et al. (2019). Comparative genomic analysis of *pseudomonas amygdali* pv. *Lachrymans* NM002: insights into its potential virulence genes and putative invasion determinants. *Genomics* 111, 1493–1503. doi: 10.1016/j.ygeno.2018.10.004

Liang, T. W., Wu, C. C., Cheng, W. T., Chen, Y. C., Wang, C. L., Wang, I. L., et al. (2014). Exopolysaccharides and antimicrobial biosurfactants produced by *Paenibacillus macerans* TKU029. *Appl. Biochem. Biotechnol.* 172, 933–950. doi: 10.1007/s12010-013-0568-5

Ma, M., Wang, C., Ding, Y., Li, L., Shen, D., Jiang, X., et al. (2011). Complete genome sequence of *Paenibacillus polymyxa* SC2, a strain of plant growth-promoting rhizobacterium with broad-spectrum antimicrobial activity. *J. Bacteriol.* 193, 311–312. doi: 10.1128/JB.01234-10

Mano, Y., and Nemoto, K. (2012). The pathway of auxin biosynthesis in plants. *J. Exp. Bot.* 63, 2853–2872. doi: 10.1093/jxb/ers091

Mukhia, S., Kumar, A., Kumari, P., and Kumar, R. (2022). Psychrotrophic plant beneficial bacteria from the glacial ecosystem of Sikkim Himalaya: genomic evidence for the cold adaptation and plant growth promotion. *Microbiol. Res.* 260:127049. doi: 10.1016/j.micres.2022.127049

Naing, K. W., Nguyen, X. H., Anees, M., Lee, Y. S., Kim, Y. C., Kim, S. J., et al. (2015). Biocontrol of *Fusarium* wilt disease in tomato by *Paenibacillus ehimensis* KWN38. *World J. Microbiol. Biotechnol.* 31, 165–174. doi: 10.1007/s11274-014-1771-4

Nakkeeran, S., Rajamanickam, S., Saravanan, R., Vanthana, M., and Soorianathasundaram, K. (2021). Bacterial endophytome-mediated resistance in banana for the management of *Fusarium* wilt. *3 Biotech* 11, 267. doi: 10.1007/s13205-021-02833-5

Navarro-Noya, Y. E., Hernández-Mendoza, E., Morales-Jiménez, J. M., Jan-Roblero, J., Martínez-Romero, E., and Hernández-Rodríguez, C. (2012). Isolation and characterization of nitrogen fixing heterotrophic bacteria from the rhizosphere of pioneer plants growing on mine tailings. *Appl. Soil Ecol.* 62, 52–60. doi: 10.1016/j.apsoil.2012.07.011

Ngashangva, N., Mukherjee, P., Sharma, K. C., Kalita, M. C., and Indira, S. (2021). Analysis of antimicrobial peptide metabolome of bacterial endophyte isolated from traditionally used medicinal plant millettia pachycarpa benth. *Front. Microbiol.* 12:656896. doi: 10.3389/fmicb.2021.656896

Przemieniecki, S. W., Kurowski, T. P., Kotlarz, K., Krawczyk, K., Damszel, M., Pszczółkowska, A., et al. (2019). Bacteria isolated from treated wastewater for biofertilization and crop protection against *Fusarium* spp. pathogens. *J. Soil Sci Plant Nut.* 19, 1–11. doi: 10.1007/s42729-018-0001-9

Rajkumar, M., Ae, N., Prasad, M. N. V., and Freitas, H. (2010). Potential of siderophore-producing bacteria for improving heavy metal phytoextraction. *Trends Biotechnol.* 28, 142–149. doi: 10.1016/j.tibtech.2009.12.002

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U. S. A.* 106, 19126–19131. doi: 10.1073/pnas.0906412106

Sang, M. K., Kim, E. N., Han, G. D., Kwack, M. S., Jeun, Y. C., and Kim, K. D. (2014). Priming-mediated systemic resistance in cucumber induced by pseudomonas azotoformans GC-B19 and *Paenibacillus elgii* MM-B22 against *Colletotrichum orbiculare*. *Phytopathology* 104, 834–842. doi: 10.1094/PHYTO-11-13-0305-R

Segata, N., Börnigen, D., Morgan, X. C., and Huttenhower, C. (2013). Phylo PhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* 4, 2304. doi: 10.1038/ncomms3304

Shin, S. H., Kim, S., Kim, J. Y., Song, H. Y., Cho, S. J., Kim, D. R., et al. (2012). Genome sequence of *Paenibacillus terrae* HPL-003, a xylanase-producing bacterium isolated from soil found in forest residue. *J. Bacteriol.* 194, 1266. doi: 10.1128/JB.06668-11

Shokri, D., and Emtiazi, G. (2010). Indole-3-acetic acid (IAA) production in symbiotic and non-symbiotic nitrogen-fixing bacteria and its optimization by Taguchi design. *Curr. Microbiol.* 61, 217–225. doi: 10.1007/s00284-010-9600-y

Siddiqi, M. Z., Siddiqi, M. H., Im, W. T., Kim, Y. J., and Yang, D. C. (2015). *Paenibacillus kyungheensis* sp. nov., isolated from flowers of magnolia. *Int. J. Syst. Evol. Microbiol.* 65, 3959–3964. doi: 10.1099/ijsem.0.000521

Soni, R., Rawal, K., and Keharia, H. (2021). Genomics assisted functional characterization of *Paenibacillus polymyxa* HK4 as a biocontrol and plant growth promoting bacterium. *Microbiol. Res.* 248:126734. doi: 10.1016/j.micres.2021

Strieker, M., Tanović, A., and Marahiel, M. A. (2010). Nonribosomal peptide synthetases: structures and dynamics. *Curr. Opin. Struct. Biol.* 20, 234–240. doi: 10.1016/j.sbi.2010.01.009

Timmusk, S., Copolovici, D., Copolovici, L., Teder, T., Nevo, E., and Behers, L. (2019). P*aenibacillus polymyxa* biofilm polysaccharides antagonise *Fusarium graminearum*. *Sci. Rep.* 9, 662. doi: 10.1038/s41598-018-37718-w

Trinh, C. S., Jeong, C. Y., Lee, W. J., Truong, H. A., Chung, N., Han, J., et al. (2018). *Paenibacillus pabuli* strain P7S promotes plant growth and induces anthocyanin accumulation in *Arabidopsis thaliana*. *Plant Physiol. Biochem.* 129, 264–272. doi: 10.1016/j.plaphy.2018.06.001

Van Belkum, M. J., Lohans, C. T., and Vederas, J. C. (2015). Draft genome sequences of *Paenibacillus polymyxa* NRRL B-30509 and *Paenibacillus terrae* NRRL B-30644, strains from a poultry environment that produce tridecaptin A and paenicidins. *Genome Announc.* 3, e00372-15. doi: 10.1128/genomeA.00372-15

Vejan, P., Abdullah, R., Khadiran, T., Ismail, S., and Nasrulhaq Boyce, A. (2016). Role of plant growth promoting rhizobacteria in agricultural sustainability-a review. *Molecules* 21, 573. doi: 10.3390/molecules21050573

Von der Weid, I., Alviano, D. S., Santos, A. L. S., Soares, R. M. A., Alviano, C. S., and Seldin, L. (2003). Antimicrobial activity of *Paenibacillus peoriae* strain NRRL BD-62 against a broad spectrum of phytopathogenic bacteria and fungi. *J. Appl. Microbiol.* 95, 1143–1151. doi: 10.1046/j.1365-2672.2003.02097.x

Wang, L. Y., Li, J., Li, Q. X., and Chen, S. F. (2013). *Paenibacillus beijingensis* sp. nov., a nitrogen-fixing species isolated from wheat rhizosphere soil. *Antonie Van Leeuwenhoek* 104, 675–683. doi: 10.1007/s10482-013-9974-5

Xie, J. B., Du, Z., Bai, L., Tian, C., Zhang, Y., and Xie, J. Y. (2014). Comparative genomic analysis of N_2-fixing and non-N_2-fixing *Paenibacillus* spp.: organization, evolution and expression of the nitrogen fixation genes. *PLoS Genet.* 10:e1004231. doi: 10.1371/journal.pgen.1004231

Xie, J., Shi, H., Du, Z., Wang, T., Liu, X., and Chen, S. (2016). Comparative genomic and functional analysis reveal conservation of plant growth promoting traits in *Paenibacillus polymyxa* and its closely related species. *Sci. Rep.* 6, 21329. doi: 10.1038/srep21329

Yadav, D. R., Adhikari, M., Kim, S. W., Kim, H. S., and Lee, Y. S. (2021). Suppression of *Fusarium* wilt caused by *Fusarium oxysporum* f. sp. *lactucae* and growth promotion on lettuce using bacterial isolates. *J. Microbiol. Biotechnol.* 31, 1241–1255. doi: 10.4014/jmb.2104.04026

Yadav, M., Dubey, M. K., and Upadhyay, R. S. (2021). Systemic resistance in chilli pepper against anthracnose (caused by *Colletotrichum truncatum*) induced by *Trichoderma harzianum*, *Trichoderma asperellum* and *Paenibacillus dendritiformis*. *J. Fungi (Basel).* 7, 307. doi: 10.3390/jof7040307

Yin, X. T., Li, T. G., Jiang, X. L., Tang, X. N., Zhang, J. K., Yuan, L. F., et al. (2022). Suppression of grape white rot caused by *Coniella vitis* using the potential biocontrol agent *Bacillus velezensis* GSBZ09. *Pathogens.* 11, 248. doi: 10.3390/pathogens11020248

Yuan, L., Li, L., Zheng, F., Shi, Y., Xie, X., Chai, A., et al. (2020). The complete genome sequence of *Rahnella aquatilis* ZF7 reveals potential beneficial properties and stress tolerance capabilities. *Arch. Microbiol.* 202, 483–499. doi: 10.1007/s00203-019-01758-1

Zaidi, A., Khan, M. S., Ahemad, M., and Oves, M. (2009). Plant growth promotion by phosphate solubilizing bacteria. *Acta Microbiol. Immunol. Hung.* 56, 263–284. doi: 10.1556/AMicr.56.2009.3.6

Zhai, Y., Zhu, J. X., Tan, T. M., Xu, J. P., Shen, A. R., Yang, X. B., et al. (2021). Isolation and characterization of antagonistic *Paenibacillus polymyxa* HX-140 and its biocontrol potential against *Fusarium* wilt of cucumber seedlings. *BMC Microbiol.* 21, 75. doi: 10.1186/s12866-021-02131-3

Zhang, F., Li, X. L., Zhu, S. J., Ojaghian, M. R., and Zhang, J. Z. (2018). Data on the ultrastructural characteristics of *Paenibacillus polymyxa* isolates and biocontrol efficacy of *P. polymyxa* ShX301. *Data Brief* 21, 259–262. doi: 10.1016/j.dib.2018.09.058

Zhou, L., Zhang, T., Tang, S., Fu, X., and Yu, S. (2020). Pan-genome analysis of *Paenibacillus polymyxa* strains reveals the mechanism of plant growth promotion and biocontrol. *Antonie Van Leeuwenhoek* 113, 1539–1558. doi: 10.1007/s10482-020-01461-y

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