frontiers | Research Topics

Computational and systems biology methods for elucidating associations between cancer and microbes

Edited by
Lihong Peng, Taoyang Wu, Fei Ma and Chuan Lu

Published in
Frontiers in Microbiology

ISSN 1664-8714
ISBN 978-2-8325-4396-2
DOI 10.3389/978-2-8325-4396-2

About Frontiers

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

Frontiers journal series

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

Dedication to quality

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

What are Frontiers Research Topics?

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

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

Computational and systems biology methods for elucidating associations between cancer and microbes

Topic editors

Lihong Peng — Chinese PLA General Hospital, China
Taoyang Wu — University of East Anglia, United Kingdom
Fei Ma — Center for National Cancer, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, China
Chuan Lu — Aberystwyth University, United Kingdom

Citation

Peng, L., Wu, T., Ma, F., Lu, C., eds. (2024). *Computational and systems biology methods for elucidating associations between cancer and microbes*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4396-2

Table of contents

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 14 September 2022
DOI 10.3389/fmicb.2022.1005241

OPEN ACCESS

EDITED BY
Fei Ma,
Chinese Academy of Medical Sciences and Peking Union Medical College, China

REVIEWED BY
Guoping Cai,
Yale University, United States
Kopen Wang,
Johns Hopkins Medicine, United States

*CORRESPONDENCE
Xu-Ru Jin
wzjinxuru@163.com
Cheng-Shui Chen
wzchencs@163.com

†These authors have contributed equally to this work

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

RECEIVED 28 July 2022
ACCEPTED 29 August 2022
PUBLISHED 14 September 2022

CITATION
He J-H, Ruan J-X, Lei Y, Hua Z-D, Chen X, Huang D, Chen C-S and Jin X-R (2022) Supplementary benefits of CT-guided transthoracic lung aspiration biopsy for core needle biopsy.
Front. Microbiol. 13:1005241.
doi: 10.3389/fmicb.2022.1005241

Supplementary benefits of CT-guided transthoracic lung aspiration biopsy for core needle biopsy

Jia-Huan He[1†], Jia-Xing Ruan[2†], Ying Lei[1], Zhi-Dan Hua[1], Xiang Chen[3], Da Huang[4], Cheng-Shui Chen[3*] and Xu-Ru Jin[3*]

[1]Department of Respiratory and Critical Care Medicine, Quzhou People's Hospital (Quzhou Hospital Affiliated to Wenzhou Medical University), Quzhou, China, [2]Department of Respiratory and Critical Care Medicine Taizhou Central Hospital (Taizhou University Hospital), Taizhou, China, [3]Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China, [4]Department of Radiology, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China

Objective: This study aimed to investigate the diagnostic efficacy of computed tomography (CT)-guided transthoracic lung core needle biopsy combined with aspiration biopsy and the clinical value of this combined routine microbial detection.

Materials and methods: We retrospectively collected the electronic medical records, CT images, pathology, and other data of 1085 patients with sequential core needle biopsy and aspiration biopsy of the same lung lesion under CT guidance in the First Affiliated Hospital of Wenzhou Medical University from January 2016 to January 2021. GenXpert MTB/RIF detection and BD BACTEC™ Mycobacterium/fungus culture were applied to identifying the microbiological results of these patients. We then compared the positive diagnostic rate, false negative rate, and diagnostic sensitivity rate of three methods including core needle biopsy alone, aspiration biopsy alone, and both core needle biopsy and aspiration biopsy.

Results: The pathological results of cutting histopathology and aspiration of cell wax were examined for 1085 patients. The diagnostic rates of cutting and aspiration pathology were 90.1% (978/1085) and 86.3% (937/1085), respectively, with no significant difference ($P > 0.05$). Considering both cutting and aspiration pathologies, the diagnostic rate was significantly improved, up to 98% (1063/1085) ($P < 0.001$). A total of 803 malignant lesions were finally diagnosed (803/1085, 74.0%). The false negative rate by cutting pathology was 11.8% (95/803), which was significantly lower than that by aspiration biopsy [31.1% (250/803), $P < 0.001$]. Compared with core needle biopsy alone, the false negative rate of malignant lesions decreased to 5.6% (45/803) ($P < 0.05$). Next, the aspirates of the malignant lesions highly suspected of corresponding infection were cultured. The results showed that 16 cases (3.1%, 16/511) were infected with Mycobacterium tuberculosis

complex, *Aspergillus niger*, and *Acinetobacter baumannii*, which required clinical treatment. 803 malignant tumors were excluded and 282 cases of benign lesions were diagnosed, including 232 cases of infectious lesions (82.3%, 232/282). The diagnostic rate of Mycobacterium/fungus culture for infectious lesions by aspiration biopsy (47.4%) was significantly higher than that by lung core needle biopsy (22.8%; $P < 0.001$). The diagnostic rate of aspiration biopsy combined with core needle biopsy was 56% (130/232). The parallel diagnostic rate of aspirated biopsy for GenXpert detection and Mycobacterium/fungal culture combined with core needle biopsy was 64.7% (150/232), which was significantly higher than that of lung core needle biopsy alone ($P < 0.001$). Finally, pulmonary tuberculosis was diagnosed in 90 cases (38.8%) of infectious lesions. Compared with the sensitivity of core needle biopsy to detect tuberculosis (27.8%, 25/90), the sensitivity of aspirating biopsy for GenXpert detection and Mycobacterium/fungal culture was significantly higher, at 70% (63/90) and 56.7% (51/90), respectively. Although there was no significant difference in the sensitivity of aspirated biopsy for GenXpert and Mycobacterium/fungal culture to detect pulmonary tuberculosis, the sensitivity was significantly increased to 83.3% ($P < 0.05$) when the two tests were combined. Moreover, when aspirated biopsies were combined with GenXpert detection, Mycobacterium/fungus culture, and core needle biopsy, the sensitivity was as high as 90% (81/90).

Conclusion: CT-guided lung aspiration biopsy has a significant supplementary effect on core needle biopsies, which is indispensable in clinical application. Additionally, the combination of aspiration biopsy and core needle biopsy can significantly improve the diagnostic rate of benign and malignant lesions. Aspiration biopsy showed that pulmonary malignant lesions are complicated with pulmonary tuberculosis, aspergillus, and other infections. Finally, the diagnostic ability of lung puncture core needle biopsy and aspiration biopsy combined with routine microbial detection under CT positioning in the diagnosis of pulmonary infectious diseases was significantly improved.

KEYWORDS

CT-guided lung biopsy, core needle biopsy, aspiration biopsy, pathology, microbial diagnosis

Introduction

Lung cancer is the leading cause of cancer-related death worldwide. The 5-year survival rate of lung cancer across all stages is only 4–17% (Nasim et al., 2019), which is mainly due to the high rates of recurrence and metastasis (He et al., 2020; Liu et al., 2021). Therefore, early diagnosis and intervention are crucial to successful treatment of lung cancer. The continuous development of computed tomography (CT) imaging technology has increased the ability to detect suspicious lung lesions, which may have otherwise been missed (Zurstrassen et al., 2020). A large number of the lung lesions found by CT are caused by infection rather than cancer, and their rapid progress can lead to systemic multiple organ failure. Due to the inability to identify the pathogen, treatment is often delayed, which is equally as life-threatening as the cancer itself. However, CT cannot accurately determine the benign and malignant lesions, which instead require the use of small biopsy or surgical pathology.

The most common methods of small lung biopsy include endobronchial ultrasound-guided biopsy, image-guided transthoracic lung biopsy, and video-assisted thoracoscopic biopsy. The combination of radiography and biopsy has developed to such an extent that image-guided transthoracic lung puncture is now considered as a safe and effective diagnostic method (Lee et al., 2019; Mallow et al., 2019; Ma et al., 2020; Chen et al., 2021), which has the advantages of high sensitivity and specificity, and low cost. CT-assisted

lung puncture mainly includes cutting needle biopsy and aspiration needle biopsy. The cut specimens are subjected to histopathological analysis, while the needle aspiration specimens are commonly used for cytological evaluation. Studies have shown that the simultaneous use of both methods under CT guidance has stronger diagnostic ability than the use of one method alone. Indeed, the sensitivity and specificity are as high as 92.52% ± 3.14% and 97.98% ± 3.28%, respectively, while the puncture risk is not significantly increased (Yamagami et al., 2003; Choi et al., 2013).

Numerous studies have confirmed that core needle biopsy, also known as core biopsy, is the preferred choice for the diagnosis of malignant lung lesions. The diagnostic accuracy of pathological tissue analysis by cutting needle biopsy is higher than that by aspiration biopsy (McLean et al., 2018; Zhang et al., 2018; Sattar et al., 2019; Li et al., 2020; Tsai et al., 2020; Ye et al., 2022), but there is insufficient basis for identifying the pathogen responsible for lung infections. Lung aspiration biopsy can directly connect the aspirated tissue with the sterile culture bottle through the aspirating needle and attract the tissue through negative pressure. Compared with cutting into strip tissue, this technique can avoid the crushing and pollution of lung tissue and improve the detection rate of pathogens. However, in aspiration biopsy, the aspirate contains more bloody fluid, which affects the pathological diagnosis of aspirated cell wax and increases the false negative rate. Here, we focus on the diagnostic efficacy of lung cutting combined with aspiration biopsy, specifically, the clinical utility and potential value of combined routine microbial detection in clinical application, with the aim to provide a basis for diagnosis and decision-making.

Materials and methods

We retrospectively collected the electronic medical records, CT images, pathology, and other data of 1085 patients who underwent continuous CT-guided core needle biopsy and aspiration biopsy of the same lung lesion at our institution (provincial first-class hospital) from January 2016 to January 2021. GenXpert MTB/RIF detection and BD BACTECTM Mycobacterium/fungus culture were applied to identifying the microbiological results of these patients. All included patients provided informed consent for the study. Before the biopsy, a thoracic interventional radiologist with 30 years of experience evaluated the radiological characteristics of the patients' CT pulmonary lesions and determined the appropriate puncture point under CT positioning. The biopsy was completed by a senior pulmonary physician with extensive experience in interventional technology using a coaxial biopsy needle. Cytopathologists and Cytotechnologists were present at all of the biopsies to assess the adequacy of the samples. The inclusion criteria and exclusion criteria of this study are shown in **Figure 1**.

Biopsy procedure

Before biopsy

For patients with lung lesions screened by CT, the necessity of lesion biopsy was first preliminarily evaluated. Next, the hospitalization was arranged, while considering the patient's medical history in detail before biopsy. The results of blood routine examination and blood coagulation, and lung function were improved, and a puncture needle with appropriate specifications was selected. The puncture method and path were designed in advance, avoiding blood vessels, the heart, and lung bullae.

Biopsy procedure

The patient was placed in the supine, lateral, or prone position according to the location of the lesion. Usually, 2-mm thick spiral CT scanning was performed to determine the puncture focus, and a self-made fence-like metal surface locator was used to assist in determining the needle entry point of the chest wall skin. The skin at the puncture position was disinfected at least twice with an Iodophor cotton swab, and the diameter of the disinfection range was ≥15 cm. The operator wore sterile gloves, laid a sterile hole towel, and used a 5-ml syringe to extract 2% lidocaine for local infiltration anesthesia, being careful to avoid puncture to the blood vessel. The puncture needle was a semi-automatic combined biopsy needle (fine core biopsy needle; Nagano, Gyoda City, Saitama, Japan), with two specifications of 10 cm and 15 cm in length. The semi-automatic spring core needle biopsy gun is equipped with a cutting needle core with 18 gauge or 20 gauge and a 2-cm groove (the length of the cutting groove can be adjusted to 1 cm according to the size of the lesion). The corresponding supporting sheath tube of 17 gauge or 19 gauge was used for aspiration biopsy, and the supporting sheath tube also has a needle core to assist in pre needle insertion. First, the matching sheath and sheath needle core were inserted into the lower edge of the chest wall, before conducting CT scanning to confirm the angle and needle distance of the puncture needle before inserting the needle into the edge of the target lesion. After plain CT scanning to confirm the correct position of the needle tip, the needle core was pulled out, the length of the cutting needle core groove was preset to 1 cm or 2 cm, the needle core of the core needle biopsy gun was inserted into the sheath, and the spring plug was pressed to complete the core needle biopsy. If the operator judged the tissue to be insufficient, the core needle biopsy needle core was re-inserted for repeated operation without pulling out the sheath. The biopsy needle was generally used to puncture 2–3 times to obtain 2–3 tissues. The biopsy prints tended to be made first, before the cut specimens were placed

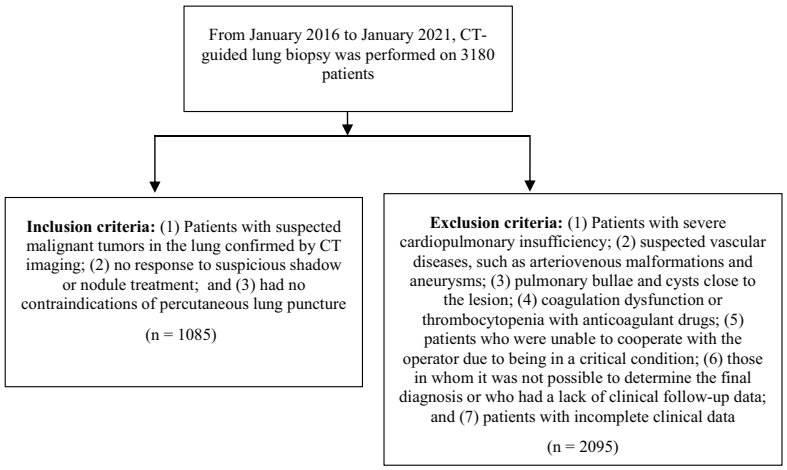


FIGURE 1
Inclusion and exclusion flowchart of the study.

in 10% formalin solution for histopathological evaluation. For specimens suspected to have specific infection, acid fast staining, silver hexamine staining (MSN), and Schiff periodate (PAS) staining can be used to perform further investigation. After the core needle biopsy gun was removed, the end of the tube sheath was connected with a 10-ml syringe barrel to form negative pressure suction. The aspirated tissue was placed in 10% formalin solution to prepare cell blocks for cytological evaluation. For those with clinical indications of pathogen infection and high suspicion of corresponding infection, the pulmonary physician decided whether to use part of the aspirates for microbiological examination, such as GenXpert MTB/RIF detection, BD BACTEC^{TM} Mycobacterium/fungal culture (Becton, Dickinson and Company, USA, and BacT/Alert aerobic and anaerobic microbial culture (bioMerieux, Inc., USA). Cut tissue samples were stained with hematoxylin eosin (HE) to observe the histomorphology under the microscope for further immunohistochemical analysis, or were used for gene testing and formulating individualized treatment plans. If the cut samples were too small, immunohistochemical analysis was conducted on paraffin sections of aspirated cells to determine the subtype or source of cancer. For the lesions with unsatisfactory materials, repeat CT scanning was conducted to confirm the position of the needle tip, followed by puncture and resampling. During the operation, the patients were closely observed for signs including chest tightness, shortness of breath, palpitation, hemoptysis, severe chest pain, and other abnormalities.

After biopsy

After the operation, routine CT scanning was performed to observe whether there were immediate complications related to biopsy, such as pneumothorax, intrapulmonary hemorrhage, and hemoptysis. Patients rested in the examination room for $\geq$3 h, during which time, their vital signs were closely monitored and they underwent chest plain film to detect whether there was delayed pneumothorax within 3 h after surgery. Some cases of asymptomatic pneumothorax (more stable pneumothorax and slight blood in the sputum) can be treated conservatively, and the clinical condition can be closely observed for improvement. However, when patients present with respiratory distress and progressive pneumothorax, a thoracic drainage tube should be placed for treatment. Moreover, in cases with high levels of hemoptysis, symptomatic treatment should be given with hemostatic drugs to prevent asphyxia.

Final diagnostic criteria

The final diagnosis was determined by a comprehensive analysis of the hospitalized patients' electronic medical record data and clinical follow-up data. The final determination of malignant lesions was based on the following: (1) in patients who underwent surgery, the final diagnosis is surgical pathology; (2) other non-surgical biopsy pathology considers malignancy, including CT-guided lung puncture or secondary lung puncture pathology, which clearly considers malignancy, and the malignant tumor is confirmed by endobronchial ultrasound-guided transbronchial needle aspiration (EBUS-TBNA), transbronchial lung biopsy (TBLB), pleural effusion cell block, or cervical lymph node metastasis biopsy; and (3) a typical malignant growth process is observed in the clinic. Positron emission computed tomography (PET-CT) tumor imaging considers malignancy, CT image follow-up, progressive enlargement of primary lesions, and occurrence of metastases, and can be used to initiate the treatment of malignant tumors.

The final diagnosis of benign lesions was based on the following criteria, provided that the lesions had no malignant

basis: (1) the biopsy lesions were confirmed to be benign by surgery and pathology; (2) the biopsy lesions were confirmed to have other benign changes determined by non-surgical biopsy pathology, such as pulmonary tuberculosis, pulmonary cryptococcosis, pulmonary aspergillosis, and hamartoma; (3) clinical imaging follow-up after discharge showed that the diameter of the lesion decreased by ≥20%, the lesion subsided, the lesion was stable for ≥24 months without special treatment (Min et al., 2009; Fontaine-Delaruelle et al., 2015; Li et al., 2020), there were no new solid components or invasive changes (e.g., short hair prick sign, lesion enlargement, pleural adhesion); (4) clear findings of microbial pathogens, such as acid-fast bacteria detected by tissue acid fast staining, Mycobacterium tuberculosis complex, non-Mycobacterium tuberculosis, pulmonary Aspergillus, or Cryptococcus cultured in lung tissue puncture or bronchoscopic alveolar lavage fluid, and genes of Mycobacterium tuberculosis were detected, which facilitated the initiation of relevant treatment; and (5) clear discharge clinical diagnosis should be considered benign. Considering benign lesions along with outpatient follow-up records, the follow-up time was generally 24 months.

In conclusion, we excluded from this analysis patients in whom it was not possible to determine the final diagnosis, or those without clinical follow-up data.

Data collection and definition of diagnostic results

The following information was collected from the medical electronic medical record system and CT images: (1) basic data, including the patient's age, sex, smoking history, history of extrapulmonary malignant tumor, presence of lesion cavity and type of lesion (the density under the lung window of CT image is divided into pure ground glass, partial solid, and solid), and smoking history, including never smoking (no smoking history), previous smoking (i.e., no smoking for the 3 months prior to the biopsy), and current smoking (i.e., smoking within the 3 months prior to the biopsy); (2) biopsy process data, including the patient's body position (supine, lateral, or prone), the size of the biopsy target focus (the longest axial diameter of the cross section of the focus measured under the lung window), the lung lobe (left upper lobe, left lower lobe, right upper lobe, right lower lobe, and middle lobe or interlobular fissure) of the puncture biopsy, and the puncture depth (distance of the focus passing through the lung parenchyma along the puncture path); and (3) biopsy results, including the histopathology of the core needle biopsy, the pathology of the cell block of the aspiration biopsy, the corresponding immunohistochemical analysis, the microbial results, and complications related to biopsy [e.g., hemoptysis (excluding hemoptysis caused by primary diseases), pneumothorax, further placement of thoracic drainage tube, and other rare and serious complications].

According to the description of the pathological report, the pathological results of core needle biopsy and aspiration biopsy under CT positioning were divided into three main categories: (1) malignant, with malignant tumor cells, including heterocyst cells showing a tendency toward malignancy; (2) benign, no obvious malignant findings; and (3) insufficient specimens, such as only bloody fluid, normal lung tissue, or too few puncture objects directly indicated in the operation record.

Histopathologically positive infectious diagnosis included the following: (1) pulmonary tuberculosis, as evidenced by granulomatous inflammation with caseous necrosis and surrounding Langhans giant cells, with or without positive acid fast staining; (2) pulmonary cryptococcosis (Setianingrum et al., 2019), in which cryptococcal spores or bacteria are observed, which may manifest as granulomatous inflammation or pneumonia of multinucleated giant and epithelioid cells, with positive PAS and hexamine silver staining; (3) pulmonary aspergillosis, as shown by the Aspergillus filaments or spheroids; and (4) other pulmonary fungal (Roden and Schuetz, 2017) or bacterial infections, in which fungal filaments or bacteria can be observed under the microscope.

Statistical analysis

When analyzing the diagnostic rate of lung puncture for infectious diseases, if the histopathology only indicates granulomatous inflammation, organized pneumonia, interstitial pneumonia, and chronic inflammation, no infectious diagnosis can be made; thus, such cases were not be included in the calculation of the positive diagnostic rate of infectious diseases in this study.

The measurement data are expressed as the mean ± standard deviation or median (range) according to whether the data were normally distributed. The count data are expressed as the rate. The sensitivity or diagnostic rate of the two methods were compared using chi-square test and McNemar's test, and the p-value was calculated. A two-tailed $p < 0.05$ was considered to indicate statistical significance. SPSS software version 22.0 was used to conduct all statistical analyses.

Results

Pathological diagnostic value of core needle biopsy and aspiration biopsies

This study included 1085 cases who underwent pathological analysis of cutting histopathology and aspiration of cell block in parallel (**Table 1**). The median age of the patients was 63 (19–93) years, and 62.5% were male. The cases comprised 981 solid lesions (90.4%), 94 sub-solid lesions (8.7%), and 10 pure ground glass density lesions (0.9%). Moreover, there were 331

biopsy lesions with a diameter ≤20 mm, accounting for 30.5%; 369 cases with a diameter >20 mm and ≤40 mm, accounting for 34.0%; and 385 cases with a diameter >40 mm, accounting for 35.5%. We observed cavities in 5.2% (n = 56) of all cases. The corresponding lung lobes punctured were 22.7% (n = 246) in the left upper lobe, 19.9% (n = 216) in the left lower lobe, 23.0% (n = 250) in the right upper lobe, 25.1% (n = 272) in the right lower lobe, and 9.3% (n = 101) in the middle lobe or interlobular fissure. The body position distribution of patients during the operation the supine position in 397 (36.6%), the prone position in 660 (60.8%), and the lateral position in 28 (2.6%). The median puncture depth was 15 (0–79) mm. Pneumothorax occurred in 388 cases (35.8%) after puncture, of which 28 cases (2.6%) required thoracic tube drainage, while 88 cases (8.1%) had hemoptysis and recovered after conservative treatment. No rare or serious complications were found.

Pathological diagnostic rate of core needle biopsy and aspiration biopsy

Table 2 lists the pathological classification of cut tissues and aspirated cells, which were mainly divided into malignant, benign, and insufficient. Of the analyzed lesions, 708 (65.3%) were malignant, 270 (25.0%) were benign, and 107 (9.9%) were insufficient specimens. The diagnostic rate was 90.1% (978/1085). The pathological diagnosis by aspiration biopsy identified 553 cases of malignant lesions (51.0%), 384 cases of benign lesions (35.5%), and 148 cases of insufficient specimens (13.7%), with a diagnostic rate of 86.3% (937/1085). There was no significant difference between the pathological diagnosis rate of cutting and aspiration (P > 0.05). Compared with the diagnostic rate of cutting or aspiration alone, when considering the pathology of both, the number of insufficient specimens decreased to 22 cases, and the diagnostic rate was significantly improved, up to 98.0% (1063/1085, P < 0.001).

False negative rate of core needle biopsy and aspiration biopsy pathology for malignant lesions

We found no misdiagnosis of malignant tumors in the pathological dataset used in this study. The histopathology of core needle biopsy, while 41 cases (15.2%, 41/270) and 54 cases (50.5%, 54/107) were finally diagnosed as malignant. The false negative rate of malignant lesions was 11.8% (95/803). Moreover, 107 (27.9%, 107/384) and 143 (96.6%, 143/148) cases were finally diagnosed as malignant, with a false negative rate of malignant lesions of 31.1% (250/803). The false negative rate of cut tissue pathology was significantly lower than that of aspiration (P < 0.001); however, compared with cutting alone, when considering cutting histopathology and aspiration cell pathology, the false negative rate of malignant lesions decreased significantly (5.6%, 45/803; P < 0.05).

Clinical significance of aspiration tissue culture in malignant lesions

Among the above 803 cases of pulmonaryung malignant lesions, 511 cases (63.6%) were highly suspected of pathogen infection due to the relevant pathogen signs and clinical

TABLE 1 General patient information related to biopsy (n = 1085).

Basic information	Number of cases (%)
Age (years)	
Median (range)	63 (19–93)
Sex	
Male	678 (62.5)
Female	407 (37.5)
Smoking history	
Never smoke	663 (61.0)
Previous smoking	227 (20.9)
Current smoking	195 (18.0)
Puncture lung lobes	
Left upper lobe	246 (22.7)
Right upper lobe	250 (23.0)
Middle lobe or cleft lungs	101 (9.3)
Left lower lobe	216 (19.9)
Right lower lobe	272 (25.1)
Lesion type	
Pure ground glass	10 (0.9)
Partial reality	94 (8.7)
Reality	981 (90.4)
Lesion size	
≤20 mm	331 (30.5)
20–40 mm	369 (34.0)
>40 mm	385 (35.5)
Puncture depth (mm)	
Median (range)	15 (0–79)
Puncture depth	
≤10 mm	453 (41.8)
10–30 mm	326 (30.0)
>30 mm	306 (28.2)
Cavity focus	56 (5.2)
Puncture position	
Supine	397 (36.6)
Prone	660 (60.8)
Lateral	28 (2.6)
Final diagnosis	
Malignant	803 (74.0)
Benign	282 (26.0)
Complication	
Pneumothorax	388 (35.8)
Thoracic tube drainage	28 (2.6)
Hemoptysis	88 (8.1)

TABLE 2 Pathological classification of cutting and aspiration.

Core needle biopsy tissue pathology, n (%)	Aspiration biopsy tissue pathology, n (%)			Total
	Malignant	Benign	Insufficient specimens	
Malignant	503 (46.4)	107 (9.9)	98 (9.0)	708 (65.3)
Benign	18 (1.7)	224 (20.6)	28 (2.6)	270^{b41} (25.0)
Insufficient specimens	32 (2.9)	53 (4.9)	22 (2.0)	107^{b54} (9.9)
Total	553 (51.0)	384^{b107} (35.5)	148^{b143} (13.7)	1085 (100)

^{bn} n cases were finally diagnosed as malignant, that is, n cases of malignant lesions were missed.

test results, and their aspirates were subjected to microbial culture. The results of aspirated tissue culture showed that of the 16 cases (3.1%, 16/511) with primary malignant lesions of the lung, nine cases were simultaneously infected with Mycobacterium tuberculosis complex, four cases were simultaneously infected with *Aspergillus niger*, and three cases were simultaneously infected with *Acinetobacter baumannii*; clinical intervention and targeted treatment measures were required in all cases with simultaneous bacterial infection.

Significance of routine culture of aspirated tissue in the diagnosis of benign lesions

The above 282 cases of pathological exclusion of malignant tumors were analyzed retrospectively. The aspirates were routinely subjected to GenXpert MTB/RIF detection (Xpert) and BD BACTEC^{TM} Myco/F lytic culture (MFC), with or without BacT/Alert aerobic and anaerobic microbial culture.

Classification of infectious and non-infectious benign lesions

The final diagnosis of 282 cases of benign diseases was divided into two categories. First, there were 232 cases (82.3%, 232/282) of infectious diseases, including 90 cases of pulmonary tuberculosis, three cases of atypical mycobacterial lung disease, 52 cases of pulmonary fungal infection [35 cases of pulmonary cryptococcosis, 14 cases of pulmonary aspergillosis (one of which was complicated with *Escherichia coli*), one case of pulmonary marneffei basket fungus infection, one case of cerdospora infection at the tip of the lung, and one case of pulmonary filamentous fungus infection], and 29 cases of bacterial pneumonia. Additionally, there were 58 cases of pulmonary infection with unknown pathogens; in these cases, after empirical anti-infective treatment, the CT follow-up lesions subsided significantly or the clinical symptoms were relieved, but no pathogen was found. Second, there were 50 cases of non-infectious lesions (17.7%, 50/282), including 12 cases of benign tumors (three cases of sclerosing pneumocytoma, three cases of pulmonary hamartoma, two cases of schwannoma, two cases of thymoma, one case of pleural solitary fibrous tumor, one case of inflammatory myofibroblastic tumor), one case of interstitial pneumonia, two cases of cryptogenic organic pneumonia, seven cases of pneumoconiosis, and 28 cases of other non-infectious benign lesions (lesions were stable or reduced at follow-up of $\geq$1 year). Among the non-infectious lesions, histopathological diagnosis included three cases of sclerosing alveolar cell tumor (100%, 3/3), two cases of pulmonary hamartoma (66.7%, 2/3), two cases of thymoma (100%, 2/2), one case of interstitial pneumonia (100%, 1/1), and seven cases of pneumoconiosis pathological changes (100%, 7/7), with one case of suspected organic pneumonia. No specific cause was found in other clinical examinations, and the final diagnosis was cryptogenic organic pneumonia (50%, 1/2). The histopathology of the remaining 34 cases (68%, 34/50) only suggested inflammatory changes and it was not possible to make a specific benign-type diagnosis. In non-infectious benign lesions, aspiration culture was negative, and no signs of infection were found in the clinical follow up.

Diagnostic rate of cutting pathology, aspiration biopsy for mycobacterial/fungal culture, and GenXpert for infectious lesions

The diagnostic rate of core needle biopsy for pulmonary infectious diseases was 22.8% (53/232), including 25 cases of pulmonary tuberculosis, 20 cases of pulmonary cryptococcosis, four cases of pulmonary aspergillosis, three cases of fungi, and one case of bacteria. The remaining 179 cases (77.2%, 179/232) had no specific infection diagnosis. The diagnostic rate of Mycobacterium/fungus culture in infectious lesions was 47.4% (110/232). Fifty-one cases of Mycobacterium tuberculosis complex, three cases of non-Mycobacterium tuberculosis (one case of abscess Mycobacterium and two cases of intracellular Mycobacterium), 36 cases of fungi (23 cases of Cryptococcus, 10 cases of Aspergillus, one case of filamentous fungi, one case of marneffei cyanobacteria, and one case of *Cercospora apicalis*), 20 cases of bacteria, and four cases of excluding contaminated bacteria were cultured. The diagnostic rates of mycobacterial/fungal culture of cut biopsy and aspiration biopsy were compared (**Table 3**), and 33 cases of infectious

lesions were consistent between the two. The diagnostic rate of mycobacterial/fungal culture of aspiration biopsy was significantly better than that of cut biopsy (22.8%, McNemar's test, $P < 0.001$). When combined with parallel diagnosis, the diagnostic rate reached 56% (130/232). The lung aspirated tissue was routinely cultured with Xpert and Mycobacterium/fungus. When either of the two results was positive, the aspirated biopsy result was considered to be positive for microbial culture. The positive number of aspirated biopsies for Xpert combined with Mycobacterium/fungus culture was 134 (57.8%, 134/232), and the positive number of both and core needle biopsies was 37 (Table 3), which was significantly higher than the diagnostic rate of core needle biopsies alone (22.8%, McNemar's test, $P < 0.001$). The parallel diagnostic rate of the three methods was 64.7% (150/232), which was 8.7% higher than that of Mycobacterium/fungus culture combined with core needle biopsy, although without statistical significance ($P = 0.058$).

TABLE 3 Infectious diagnosis of core needle biopsy and aspiration biopsy ($n = 232$).

Aspiration biopsy	Result	Core needle biopsy		Total	*P*-value
		Positive	Negative		
MFC	Positive	33	77	110	$P < 0.001$*
	Negative	20	102	122	
	Total	53	179	232	
Xpert + MFC	Positive	37	97	134	$P < 0.001$*
	Negative	16	82	98	
	Total	53	179	232	

Xpert: GeneXpert MTB/RIF, MFC: BD BACTEC™ Mycobacterium/fungus culture. *McNemar's test.

TABLE 4 Results of BacT/ALERT culture and BD BACTEC™ culture ($n = 57$).

Final diagnosis	No. of cases	MFC (%)	BacT ALERT (%)
Pulmonary tuberculosis	15	7 (46.7)	0
Non-tuberculosis mycobacteria	1	1 (100)	0
Pulmonary fungal infection	10	6 (60)	1 (10)
Bacterial pneumonia	13	5 (38.5)	9 (69.2)
Pathogens not detected	18	–	–
Total	57	19 (33.3)	10 (17.5)

BacT/ALERT: BacT/ALERT microbial cultivation, MFC: BD BACTEC™ Mycobacterium/fungus culture.

Detection rate of BacT/Alert microbial culture

BacT/Alert microbial culture includes BacT/Alert fa (aerobic microorganism) and Sn (anaerobic bacteria). Fifty-seven cases (24.6%, 57/232) of pulmonary infectious diseases were sent for examination. The final diagnoses were 15 cases of pulmonary tuberculosis, one case of non-tuberculous Mycobacterium, 10 cases of fungal infection, 13 cases of bacterial infection, and 18 cases of unspecified pathogens. BacT/Alert microbial culture was positive in 10 cases (eight cases of anaerobic bacteria, one case of Legionella, and one case of Cryptococcus), and the diagnostic rate was only 17.5% (10/57). The positive diagnostic rate for core needle biopsies was 12.3% (7/57). The diagnostic rate of core needle biopsy combined with BacT/Alert microbial culture was 26.3% (15/57), which was not significantly different to that of core needle biopsy ($P = 0.058$). Comparing the results of the BacT/Alert microbial culture bottle with BD BACTEC™ Mycobacterium/fungus culture bottle (Table 4), only the detection rate of bacterial culture in the BacT/Alert microbial culture (69.2%) was higher than that in the Mycobacterium/fungus culture bottle (38.5%), but the difference was not statistically significant ($P > 0.05$).

Sensitivity of core needle biopsy, aspiration for mycobacterium culture, and Xpert in the diagnosis of pulmonary tuberculosis

According to the clinical diagnosis and follow-up results, 90 cases (38.8%, 90/232) were finally diagnosed as pulmonary tuberculosis. According to the histopathology of lung cutting, 25 cases were considered as tuberculosis, and the sensitivity was 27.8% (25/90) (Table 5). Compared with the sensitivity of cutting histopathology, the sensitivity of GenXpert MTB/RIF for aspiration biopsy was 70% (63/90), rifampicin resistance genes were detected in two cases (2.2%, 2/90), and the sensitivity of Mycobacterium/fungus culture for aspiration biopsy in the diagnosis of pulmonary tuberculosis was 56.7% (51/90), which was significantly increased ($P < 0.05$). GenXpert test and Mycobacterium/fungus culture for aspiration biopsy was both positive in 39 cases, and there was no significant difference in sensitivity ($P > 0.05$). However, compared with the single GenXpert MTB/RIF detection or Mycobacterium/fungal culture for aspiration biopsy, the combined detection identified 75 cases of Mycobacterium tuberculosis, and the sensitivity was significantly improved by 83.3% ($P < 0.05$). The sensitivity was as high as 90% (81/90).

Sensitivity of core needle biopsy and aspiration for fungal culture in the diagnosis of pulmonary fungal infection

According to the clinical diagnosis and follow-up results, 52 cases (22.4%, 52/232) were diagnosed as pulmonary fungal infections. The results of core needle biopsy and bacterial/fungal culture for aspiration biopsy are shown in Table 6. The sensitivity of combined detection (86.5%, 45/52) was significantly higher than that of single core needle biopsy (51.9%, 27/52; $P < 0.001$).

TABLE 5 Diagnosis of pulmonary tuberculosis by core needle biopsy and aspiration biopsy (n = 90).

Core needle biopsy	Aspiration biopsy					
	Xpert		MFC		Xpert + MFC	
	Positive	Negative	Positive	Negative	Positive	Negative
Positive	16	9	15	20	19	6
Negative	47	18	36	19	56	9
Total	63	27	51	39	75	15

Xpert: GeneXpert MTB/RIF, MFC: BD BACTEC™ mycobacterium/fungus culture.

Discussion

Transthoracic lung puncture under CT positioning is widely used in the clinic. Research has shown that the accuracy of cutting needle and aspiration needle biopsy is high, but the choice of biopsy method depends on the operator's operation experience (VanderLaan, 2016). A recent survey of American Thoracic Radiology members showed that 85% of radiologists used cutting needles with or without suction needles for lesion biopsy (Lee et al., 2017), while there are still differences in opinion regarding whether to diagnose lung lesions by combined cutting and aspiration biopsy (Aviram et al., 2007; Schoellnast et al., 2010; Choi et al., 2013; Coley et al., 2015; Marchiano et al., 2017). This retrospective analysis showed that core needle biopsy pathology combined with aspiration biopsy cell wax pathology can significantly improve the diagnostic rate of lung puncture lesions under CT localization and reduce the false negative rate of malignant lesions. The diagnostic rate of the combination is significantly higher than that of the single core needle biopsy, indicating that the core needle biopsy and aspiration biopsy under CT positioning are complementary and reduce the insufficient rate of samples, which is critical for accurate diagnosis and effective treatment decisions relating to lung lesions. Biopsy pathology is the gold standard for diagnosing lung lesions, and, crucially, may be repeated in cases with insufficient specimens in which malignant lesions are still suspected in combination with PET-CT images. However, repeat biopsy increases CT radiation exposure as well as the inherent risks related to biopsy, such as pneumothorax, focal bleeding, fever, and chest pain, which increases the medical burden. Therefore, the combination of core needle biopsy and aspiration biopsy improves the diagnostic ability of lesions and has greater clinical benefits.

TABLE 6 Diagnosis of pulmonary fungal infection by core needle biopsy and aspiration biopsy (n = 52).

Core needle biopsy	MFC for aspiration biopsy		Total
	Positive	Negative	
Positive	18	9	27
Negative	18	7	25
Total	36	16	52

MFC: BD BACTEC™ mycobacterium/fungus culture.

Studies have shown that core needle biopsy is the first choice for the diagnosis of malignant lung lesions. The false negative rate of cutting histopathology for malignant lesions is significantly lower than that of aspiration pathology. The analysis of pathological tissue has high accuracy in the diagnosis of malignant diseases (McLean et al., 2018; Zhang et al., 2018; Sattar et al., 2019; Li et al., 2020; Tsai et al., 2020; Yang et al., 2022; Ye et al., 2022), but it is insufficient to clarify the basis of lung infection. The diagnosis of benign diseases by cutting tissue is mainly analyzed from histopathology and corresponding special staining. When the pathological diagnosis of cutting tissue is non-specific inflammatory changes, it cannot be clearly diagnosed due to the lack of a pathogen basis. Lung aspiration biopsy can directly connect the aspirated tissue with the sterile culture bottle through the suction needle and suck it out through negative pressure. Compared with cutting tissues into strips, this method can save energy in the follow-up treatment ring, avoid crushing and pollution of lung tissue, and improve the detection rate of pathogens. However, in aspiration biopsy, due to more bloody fluid in the aspirate, a large number of red blood cells observed under the cell wax slice microscope may obscure heterotypic cells, thus increasing the false negative rate. Experienced doctors will choose to aspirate at different sites of the focus, but the judgment of operation will still be affected because the local bleeding usually shows that the focus is enlarged and the boundary is blurred on CT images. Additionally, it is easier to obtain necrotic and liquefied tissue by suction, which is not conducive to pathological analysis, affects the suction, and causes false negative. We also found that the false negative rate of malignant lesions was lower than that of cutting or aspiration alone. Chen et al. (2020) also believe that the application of two types of biopsies in the same lesion can effectively reduce false negative diagnosis, improve the diagnostic efficiency, and maximize the value of lung puncture. Clinically, most patients underwent lung biopsy

to determine the nature of the lesion because they either suspected or could not rule out malignant tumor. Based on various clinical auxiliary examinations and medical history analysis, even if the small biopsy pathology does not clearly indicate malignancy, if the possibility of malignancy is high, it is still be recommended to strengthen the clinical image follow-up or consider further diagnosis and treatment. Cutting combined with aspiration biopsy pathology can maximize the diagnostic ability of malignant lesions in a single operation, which optimizes the use of medical resources, reduces false negative, achieves early diagnosis, early decision-making and treatment, and improves the survival time of patients.

Immunohistochemical staining analysis is helpful to clarify the subtypes of lung cancer (particularly poorly differentiated cancer) and understand the source of cancer. In clinical application, because cutting can obtain a relatively complete small tissue, operators prefer to cut lung tissue for immunohistochemical analysis and gene detection. However, when the focus is located in the lower lobe and close to the diaphragm, the needle tip may deviate greatly during puncture due to the large respiratory amplitude, which is not conducive to accurate positioning. For small lesions ($\leq$20 mm) and subpleural lesions (Yu et al., 2020), the lesion tissue itself is fragile and the mucus changes, so it is difficult to cut the tissue. In such cases, the puncture passes through the lung parenchyma for a long distance, which is more likely to lead to pneumothorax and bleeding (Chen et al., 2021), affect the effectiveness of the operation, and lead to cutting failure. It is still necessary to attempt aspiration biopsy before the cutting becomes too difficult and the needle sheath is pulled out. Aspiration is the final chance to obtain suitable. Studies have shown that immunohistochemical analysis can be performed in cell wax samples (Lozano et al., 2015; Bayrak et al., 2021). In this retrospective study, we found that the immunohistochemical analysis of aspirated cell wax had a significant supplementary effect on malignant tumors.

Our results also showed that the risk of complications such as pneumothorax and hemoptysis was less when the routine suction was increased (Gupta et al., 2010b). The incidence of pneumothorax through sequential chest wall cutting and aspiration lung biopsy under CT guidance was 35.8%, which was equivalent to the 15.4–42.0% (Bae et al., 2020; Ruud et al., 2021) previously reported by interventional radiologists. Most pneumothorax can be observed conservatively, with only 2.6% requiring thoracic catheterization and drainage, which was lower than the pneumothorax catheterization rate of 4.3–7.3% reported by Heerink et al. (2017). Depending on the study population and the type of needle used, the incidence of CT-mediated transthoracic lung puncture hemoptysis ranges from 0.5 to 14.4% (Tai et al., 2016; Heerink et al., 2017). In this study, 8.1% of patients had hemoptysis, including some blood in the sputum. No serious complications, such as tumor needle metastasis, air embolism, and death, occurred. As lung suction is performed to generate negative pressure suction through the guide needle sheath after the cutting is completed, the sheath tube will be pulled out immediately after the completion, and the relevant specimens will be sent for examination without additional puncture needle placement, which will serve to reduce the operative duration. Moreover, lung aspirated tissue can be used for pathological HE staining, morphological analysis, and immunohistochemical analysis of cell blocks. Furthermore, especially in the case of insufficient cutting tissue or cutting failure, more tissue can be saved, which can also be used in subsequent molecular research (Zhao et al., 2014).

Some studies have shown that the combination of lung cutting and aspiration has no significant advantage in the diagnosis of certain inflammatory pathologies (e.g., pulmonary tuberculosis, pulmonary cryptococcosis) compared with cutting alone (Aviram et al., 2007; Schoellnast et al., 2010; Choi et al., 2013; Chen et al., 2020), and the specificity of cutting histopathology is significantly higher than that of aspiration pathology. This may be due to the histopathological evaluation of cutting needle biopsy and the cytological diagnosis of aspirates in most institutions. Histopathology can more effectively determine benign types, including tumors, under the microscope (e.g., pulmonary hamartoma, inflammatory pseudotumor, schwannoma, organized pneumonia, and granulomatous inflammation). However, in clinical practice, there remain deficiencies in the etiological diagnosis of lung cutting histopathology, with some studies showing that the sensitivity of lung cutting histopathology in the diagnosis of lung infection is 36%. Furthermore, most pathological results can only suggest chronic inflammation and cannot identify a specific infection. In lesions suspected of pulmonary infection, Kim et al. (2020) found that aspiration could detect more pathogenic microorganisms than cutting. Microbial culture is the gold standard for diagnosis, and its role in benign lesions, especially in infectious lesions, cannot be ignored. The aspirated tissue can be quickly and easily inhaled into the culture bottle through the sheath, which is simple and convenient to operate and has a low risk of pollution.

Recently, the application of aspiration in the diagnosis of infectious diseases has been increasing (Haas et al., 2017). Previous studies have found that the diagnostic rate of lung aspirate culture under CT localization for pulmonary opportunistic infection was 36.5%–80% in patients with immune impairment (Hwang et al., 2000; Carrafiello et al., 2006; Gupta et al., 2010a; Ideh et al., 2011; Hsu et al., 2012; Clement et al., 2014; Liu et al., 2022). The diagnostic rate of Mycobacterium/fungus cultures of lung aspirates was 47.4%, which was significantly higher than that of core needle biopsy. Although GenXpert MTB/RIF is only sensitive to Mycobacterium tuberculosis, the parallel diagnostic rate of GenXpert MTB/RIF detection and Mycobacterium/fungus culture for aspiration biopsy combined with lung cutting pathology is still 8.7% higher than that of Mycobacterium/fungus culture for aspiration biopsy combined

with core needle biopsy. Abundant materials can be obtained by one suction for auxiliary examination. Compared with core needle biopsy alone, GenXpert MTB/RIF detection and Mycobacterium/fungus culture by increasing the suction can substantially improve the overall diagnostic rate of pulmonary infectious diseases, and reduce subsequent invasive operations and targeted antiviral therapy.

China has a high incidence of tuberculosis, with the infection rate ranking second worldwide (MacNeil et al., 2020). The hidden onset of tuberculosis has become a major burden to public health, and the situation of prevention and control is grim. Early diagnosis and treatment are crucial to controlling the progression of pulmonary tuberculosis and reducing infection. Pulmonary lesions cannot be routinely screened for Mycobacterium tuberculosis infection by CT imaging, so effective clinical detection methods are required. According to the recommendations of the World Health Organization, among all adults with suspected tuberculosis, the Xpert MTB/RIF test should be preferentially used as the initial diagnosis method. In our study, the sensitivity of the lung cutting histopathology in the diagnosis of pulmonary tuberculosis was only 27.8%, which is similar to that reported in previous studies (Montenegro et al., 2014; Jiang et al., 2016). For histopathology describing inflammation, such as granulomatous inflammation, with or without coagulative necrosis, mycobacterium culture or molecular detection must be further clarified clinically. The Xpert MTB/RIF reported in the literature shows considerable variation in the detection sensitivity of tissue samples, with a reported range of 42%–100%. This variation may be related to the heterogeneity of sampling lesions, resulting in the actual detected tissues not being the most representative samples. In this study, there was no significant difference in the sensitivity of GenXpert MTB/RIF to pulmonary tuberculosis compared to Mycobacterium/fungal culture, but the sensitivity was significantly improved when they were diagnosed in parallel (up to 83.3%). Mycobacterium tuberculosis grows slowly and requires high nutrition culture medium, with the culture taking 42 days on average. GenXpert MTB/RIF detection has high sensitivity and specificity, requires less tissue, is simple and safe to operate, and has fast detection time. Additionally, it can detect Mycobacterium tuberculosis complex DNA and rifampicin resistance within 2 h (Yu et al., 2019), which can facilitate individualized anti-tuberculosis treatment as soon as possible. Therefore, as long as conditions permit, the Xpert gene detection of lung aspirates and mycobacterium culture should be conducted in parallel to diagnose the infection early, reduce transmission (Han et al., 2021), and adopt individualized anti-tuberculosis treatment more effectively. For patients with rapid progress of tuberculosis, this combination can be employed to better control tuberculosis activity and improve prognosis.

Additionally, 57 cases underwent BacT/Alert microbial culture simultaneously, and most of the results detected anaerobic bacteria. The diagnostic rate of combined detection with histopathology did not increase significantly, which may be related to the inability of BacT/Alert FA and Sn culture bottles to culture mycobacteria and the low detection rate of fungi. The average culture time of the BacT/Alert bottle is 5 days, which cannot meet the growth time of most fungi, whereas the culture time of the BD BACTEC^{TM} Mycobacterium/fungus bottle is 42 days. The detection rate of BacT/Alert microbial culture bacteria is higher than that of the BD BACTEC^{TM} Mycobacterium/fungus culture bottle, which is related to the detection of obligate anaerobic bacteria, while the BD BACTEC^{TM} Mycobacterium/fungus culture bottle cannot detect obligate anaerobic bacteria. Despite no statistical difference between the two, this may have relevance for selecting clinical antibiotics. The results of this study suggest that GenXpert MTB/RIF detection, BD BACTEC^{TM} Mycobacterium/fungus culture, and BacT/Alert specific anaerobic bacteria culture should be conducted routinely to maximize the detection rate of pathogens.

In this study, 20.6% of pulmonary infections did not identify the pathogen in other auxiliary examinations, such as blood culture, related serum antibody detection, and bronchoscopic interventional diagnosis and treatment, and the CT follow-up lesions subsided significantly or the clinical symptoms were relieved after empirical anti-infection treatment. The early use of antibiotics may make it difficult to detect sensitive bacteria in the follow-up examination. Additionally, many microorganisms are difficult to cultivate and require specific culture medium and strict culture conditions, such as Brucella, Chlamydia trachomatis, flagellin spirochete, and Neisseria gonorrhoeae (Glaser and Montone, 2020). Aspiration culture should be performed as soon as clinically feasible. Additionally, molecular technology, such as high-throughput sequencing, is not restricted by culture conditions and can directly detect nucleic acids, which is worthy of popularization.

In 50 cases of benign non-infectious lesions, the histopathological diagnosis was consistent with the imaging and clinical analysis. Indeed, the clinical CT follow-up was sufficient to exclude malignancy and consider the lesion to be benign. The combination of cutting histopathology and aspiration microbial detection is more conducive to the diagnosis of benign lesions, with the exception of infection (Hwang et al., 2000). In this study, in 3.1% of malignant lesions, infectious bacteria was detected simultaneously, including *Mycobacterium tuberculosis*, *Aspergillus*, and *Acinetobacter baumannii*. For patients with cancer, timely effective anti-infection programs will be crucial to control the spread and progress of infection.

Aspiration has a unique complementary value in the diagnosis of benign and malignant diseases. Aspiration is an easy means to process microbial samples and plays a decisive role in the diagnosis of infectious pathogens. However, this study has the following limitations: (1) it is a retrospective study, only evaluated the lung puncture data of one hospital, was limited to the operation of a single biopsy needle, and did not analyze the situation of other institutions; (2) we could not directly compare

the differences in immunohistochemistry in histopathology and cell pathology because immunohistochemical analysis is the first choice of histopathological specimens, and cell pathology is often used as auxiliary research, especially when histopathology is insufficient; and (3) due to the lack of corresponding standards, the local bleeding of lesions observed on CT images was not graded and evaluated.

Data availability statement

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

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

X-RJ conceived and designed this study. J-HH and J-XR performed the analyses and wrote the manuscript. YL, Z-DH, XC, DH, and C-SC performed the literature review and assisted with data collection. All authors have read and approved the final manuscript.

Funding

This study was supported by grants from the Key Laboratory of Interventional Pulmonology of Zhejiang Province (2019E10014) and the Zhejiang Provincial Key Research and Development Program (2020C03067), and the Wu Jieping Medical Foundation (No. 320.6750).

Acknowledgments

We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

Conflict of interest

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

Publisher's note

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

References

Aviram, G., Greif, J., Man, A., Schwarz, Y., Marmor, S., Graif, M., et al. (2007). Diagnosis of intrathoracic lesions: Are sequential fine-needle aspiration (FNA) and core needle biopsy (CNB) combined better than either investigation alone? *Clin. Radiol.* 62, 221–226. doi: 10.1016/j.crad.2006.11.003

Bae, K., Ha, J. Y., and Jeon, K. N. (2020). Pneumothorax after CT-guided transthoracic lung biopsy: A comparison between immediate and delayed occurrence. *PLoS One* 15:e0238107. doi: 10.1371/journal.pone.0238107

Bayrak, B. Y., Paksoy, N., and Vural, C. (2021). Diagnostic utility of fine needle aspiration cytology and core biopsy histopathology with or without immunohistochemical staining in the subtyping of the non-small cell lung carcinomas: Experience from an academic centre in Turkey. *Cytopathology* 32, 331–337.

Carrafiello, G., Lagana, D., Nosari, A. M., Guffanti, C., Morra, E., Recaldini, C., et al. (2006). Utility of computed tomography (CT) and of fine needle aspiration biopsy (FNAB) in early diagnosis of fungal pulmonary infections. Study of infections from filamentous fungi in haematologically immunodeficient patients. *Radiol. Med.* 111, 33–41. doi: 10.1007/s11547-006-0004-9

Chen, C., Xu, L. C., Sun, X. F., Liu, X. X., Han, Z., and Li, W. T. (2021). Safety and diagnostic accuracy of percutaneous CT-guided transthoracic biopsy of small lung nodules (<= 20 mm) adjacent to the pericardium or great vessels. *Diagn. Interv. Radiol.* 27, 94–101.

Chen, L., Jing, H., Gong, Y., Tam, A. L., Stewart, J., Staerkel, G., et al. (2020). Diagnostic efficacy and molecular testing by combined fine-needle aspiration and core needle biopsy in patients with a lung nodule. *Cancer Cytopathol.* 128, 201–206. doi: 10.1002/cncy.22234

Choi, S. H., Chae, E. J., Kim, J.-E., Kim, E. Y., Oh, S. Y., Hwang, H. J., et al. (2013). Percutaneous CT-guided aspiration and core biopsy of pulmonary nodules smaller than 1 cm: Analysis of outcomes of 305 procedures from a tertiary referral center. *Am. J. Roentgenol.* 201, 964–970. doi: 10.2214/AJR.12.10156

Clement, C., Nawgiri, R., Williams-Bouyer, N., and Schnadig, V. (2014). Correlation of fine-needle aspiration cytology with microbiologic culture: A 14-year experience at a single institution. *Mod. Pathol.* 27:391A. doi: 10.1002/cncy.21590

Coley, S. M., Crapanzano, J. P., and Saqi, A. (2015). FNA, core biopsy, or both for the diagnosis of lung carcinoma: Obtaining sufficient tissue for a specific diagnosis and molecular testing. *Cancer Cytopathol.* 123, 318–326.

Fontaine-Delaruelle, C., Souquet, P. J., Gamondes, D., Pradat, E., De Leusse, A., Ferretti, G. R., et al. (2015). Negative predictive value of transthoracic core-needle biopsy a multicenter study. *Chest* 148, 472–480. doi: 10.1378/chest.14-1907

Glaser, L. J., and Montone, K. T. (2020). A practical guide to the role of ancillary techniques in the diagnosis of infectious agents in fine needle aspiration samples. *Acta Cytol.* 64, 81–91. doi: 10.1159/000497076

Gupta, S., Sultenfuss, M., Romaguera, J. E., Ensor, J., Krishnamurthy, S., Wallace, M. J., et al. (2010a). CT-guided percutaneous lung biopsies in patients

with haematologic malignancies and undiagnosed pulmonary lesions. *Hematol. Oncol.* 28, 75–81. doi: 10.1002/hon.923

Gupta, S., Wallace, M. J., Cardella, J. F., Kundu, S., Miller, D. L., and Rose, S. C. (2010b). Quality improvement guidelines for percutaneous needle biopsy. *J. Vasc. Interv. Radiol.* 21, 969–975. doi: 10.1016/j.jvir.2010.01.011

Haas, B. M., Clayton, J. D., Elicker, B. M., Ordovas, K. G., and Naeger, D. M. (2017). CT-Guided percutaneous lung biopsies in patients with suspicion for infection may yield clinically useful information. *Am. J. Roentgenol.* 208, 459–463. doi: 10.2214/AJR.16.16255

Han, S. B., Park, J., Ji, S. K., Jang, S. H., Shin, S., Kim, M. S., et al. (2021). The impact of the Xpert MTB/RIF screening among hospitalized patients with pneumonia on timely isolation of patients with pulmonary tuberculosis. *Sci. Rep.* 11:1694. doi: 10.1038/s41598-020-79639-7

He, B., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., et al. (2020). TOOme: A novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression. *Front. Bioeng. Biotechnol.* 8:394. doi: 10.3389/fbioe.2020.00394

Heerink, W. J., De Bock, G. H., De Jonge, G. J., Groen, H. J. M., Vliegenthart, R., and Oudkerk, M. (2017). Complication rates of CT-guided transthoracic lung biopsy: Meta-analysis. *Eur. Radiol.* 27, 138–148. doi: 10.1007/s00330-016-4357-8

Hsu, J. L., Kuschner, W. G., Paik, J., Bower, N., Guillamet, M. C. V., and Kothary, N. (2012). The diagnostic yield of CT-guided percutaneous lung biopsy in solid organ transplant recipients. *Clin. Transplant.* 26, 615–621. doi: 10.1111/j.1399-0012.2011.01582.x

Hwang, S. S., Kim, H. H., Park, S. H., Jung, J. I., and Jang, H. S. (2000). The value of CT-guided percutaneous needle aspiration in immunocompromised patients with suspected pulmonary infection. *Am. J. Roentgenol.* 175, 235–238. doi: 10.2214/ajr.175.1.1750235

Ideh, R. C., Howie, S. R. C., Ebruke, B., Secka, O., Greenwood, B. M., Adegbola, R. A., et al. (2011). Transthoracic lung aspiration for the aetiological diagnosis of pneumonia: 25 years of experience from The Gambia. *Int. J. Tuberc. Lung Dis.* 15, 729–735. doi: 10.5588/ijtld.10.0468

Jiang, F. M., Huang, W. W., Wang, Y., Tian, P. W., Chen, X. R., and Liang, Z. A. (2016). Nucleic acid amplification testing and sequencing combined with acid-fast staining in needle biopsy lung tissues for the diagnosis of smear-negative pulmonary tuberculosis. *PLoS One* 11:e0167342. doi: 10.1371/journal.pone.0167342

Kim, J., Lee, K. H., Cho, J. Y., Kim, J., Shin, Y. J., and Lee, K. W. (2020). Usefulness of CT-guided percutaneous transthoracic needle lung biopsies in patients with suspected pulmonary infection. *Korean J. Radiol.* 21, 526–536.doi: 10.3348/kjr.2019.0492

Lee, C., Guichet, P. L., and Abtin, F. (2017). Percutaneous lung biopsy in the molecular profiling era a survey of current practices. *J. Thorac. Imaging* 32, 63–67. doi: 10.1097/RTI.0000000000000237

Lee, K. H., Lim, K. Y., Suh, Y. J., Hur, J., Han, D. H., Kang, M. J., et al. (2019). Diagnostic accuracy of percutaneous transthoracic needle lung biopsies: A multicenter study. *Korean J. Radiol.* 20, 1300–1310. doi: 10.3348/kjr.2019.0189

Li, H. C., Chen, R., and Zhao, J. (2020). Correlation between percutaneous transthoracic needle biopsy and recurrence in stage I lung cancer: A systematic review and meta-analysis. *BMC Pulm. Med.* 20:198. doi: 10.1186/s12890-020-01235-2

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. *Front. Cell. Dev. Biol.* 9:619330.

Liu, J., Lan, Y., Tian, G., and Yang, J. (2022). A systematic framework for identifying prognostic genes in the tumor microenvironment of colon cancer. *Front. Oncol.* 2:899156. doi: 10.3389/fonc.2022.899156

Lozano, M. D., Labiano, T., Echeveste, J., Gurpide, A., Martin-Algarra, S., Zhang, G. L., et al. (2015). Assessment of EGFR and KRAS mutation status from FNAs and core-needle biopsies of non-small cell lung cancer. *Cancer Cytopathol.* 123, 230–236. doi: 10.1002/cncy.21513

Ma, X., Xi, B., Zhang, Y., Zhu, L., Sui, X., Tian, G., et al. (2020). A machine learning-based diagnosis of thyroid cancer using thyroid nodules ultrasound images. *Curr. Bioinform.* 15, 349–358. doi: 10.2174/1574893614666191017091959

MacNeil, A., Glaziou, P., Sismanidis, C., Date, A., Maloney, S., and Floyd, K. (2020). Global epidemiology of tuberculosis and progress toward meeting global targets - worldwide, 2018. *MMWR Morb. Mortal. Wkly. Rep.* 69, 281–285.

Mallow, C., Lee, H., Oberg, C., Thiboutot, J., Akulian, J., Burks, A. C., et al. (2019). Safety and diagnostic performance of pulmonologists performing electromagnetic guided percutaneous lung biopsy (SPiNperc). *Respirology* 24, 453–458. doi: 10.1111/resp.13471

Marchiano, A. V., Cosentino, M., Di Tolla, G., Greco, F. G., Silva, M., Sverzellati, N., et al. (2017). FNA and CNB in the diagnosis of pulmonary lesions: A single-center experience on 665 patients, comparison between two periods. *Tumori* 103, 360–366. doi: 10.5301/tj.5000633

McLean, A. E. B., Barnes, D. J., and Troy, L. K. (2018). Diagnosing lung cancer: The complexities of obtaining a tissue diagnosis in the era of minimally invasive and personalised medicine. *J. Clin. Med.* 7:163. doi: 10.3390/jcm7070163

Min, J. W., Lee, S. M., Chung, D. H., Yim, J. J., Yang, S. C., Yoo, C. G., et al. (2009). Clinical significance of non-diagnostic pathology results from percutaneous transthoracic needle lung biopsy: Experience of a tertiary hospital without an on-site cytopathologist. *Respirology* 14, 1042–1050. doi: 10.1111/j.1440-1843.2009.01610.x

Montenegro, S., Delgado, C., Pineda, S., Reyes, C., De La Barra, T., Cabezas, C., et al. (2014). Efficacy of PCR for the differential diagnosis of tuberculosis in granulomatous lesions of paraffin-embedded formalin fixed tissues. *Rev. Chilena Infectol.* 31, 676–681. doi: 10.4067/S0716-10182014000600006

Nasim, F., Sabath, B. F., and Eapen, G. A. (2019). Lung cancer. *Med. Clin. North Am.* 103, 463–473. doi: 10.1016/j.mcna.2018.12.006

Roden, A. C., and Schuetz, A. N. (2017). Histopathology of fungal diseases of the lung. *Semin. Diagn. Pathol.* 34, 530–549. doi: 10.1053/j.semdp.2017.06.002

Ruud, E. A., Stavem, K., Geitung, J. T., Borthne, A., Soyseth, V., and Ashraf, H. (2021). Predictors of pneumothorax and chest drainage after percutaneous CT-guided lung biopsy: A prospective study. *Eur. Radiol.* 31, 4243–4252. doi: 10.1007/s00330-020-07449-6

Sattar, A., Khan, S. A., Al-Qamari, N., Adel, H., and Adil, S. O. (2019). Diagnostic accuracy and associated complications of percutaneous computed tomography guided core needle biopsy of pulmonary lesions using coaxial technique. *J. Pakistan Med. Assoc.* 69, 1711–1713. doi: 10.5455/JPMA.272057

Schoellnast, H., Komatz, G., Bisail, H., Talakic, E., Fauster, M., Ehammer, T., et al. (2010). CT-guided biopsy of lesions of the lung, liver, pancreas or of enlarged lymph nodes: Value of additional Fine Needle Aspiration (FNA) to Core Needle Biopsy (CNB) in an offsite pathologist setting. *Acad. Radiol.* 17, 1275–1281. doi: 10.1016/j.acra.2010.05.015

Setianingrum, F., Rautemaa-Richardson, R., and Denning, D. W. (2019). Pulmonary cryptococcosis: A review of pathobiology and clinical aspects. *Med. Mycol.* 57, 133–150. doi: 10.1093/mmy/myy086

Tai, R., Dunne, R. M., Trotman-Dickenson, B., Jacobson, F. L., Madan, R., Kumamaru, K. K., et al. (2016). Frequency and severity of pulmonary hemorrhage in patients undergoing percutaneous CT-guided transthoracic lung biopsy: Single-institution experience of 1175 cases. *Radiology* 279, 287–296. doi: 10.1148/radiol.2015150381

Tsai, P. C., Yeh, Y. C., Hsu, P. K., Chen, C. K., Chou, T. Y., and Wu, Y. C. (2020). CT-Guided core biopsy for peripheral sub-solid pulmonary nodules to predict predominant histological and aggressive subtypes of lung adenocarcinoma. *Ann. Surg. Oncol.* 27, 4405–4412. doi: 10.1245/s10434-020-08511-9

VanderLaan, P. A. (2016). Fine-needle aspiration and core needle biopsy: An update on 2 common minimally invasive tissue sampling modalities. *Cancer Cytopathol.* 124, 862–870. doi: 10.1002/cncy.21742

Yamagami, T., Iida, S., Kato, T., Tanaka, O., and Nishimura, T. (2003). Combining fine-needle aspiration and core biopsy under CT fluoroscopy guidance: A better way to treat patients with lung nodules? *Am. J. Roentgenol.* 180, 811–815. doi: 10.2214/ajr.180.3.1800811

Yang, M., Yang, H., Ji, L., Hu, X., Tian, G., Wang, B., et al. (2022). A multi-omics machine learning framework in predicting the survival of colorectal cancer patients. *Comput. Biol. Med.* 146:105516. doi: 10.1016/j.compbiomed.2022.105516

Ye, Z., Zhang, Y., Liang, Y., Lang, J., Zhang, X., Zang, G., et al. (2022). Cervical cancer metastasis and recurrence risk prediction based on deep convolutional neural network. *Curr. Bioinform.* 17, 164–173. doi: 10.2174/1574893616666210708143556

Yu, G. C., Zhong, F. M., Ye, B., Xu, X. D., Chen, D., and Shen, Y. Q. (2019). Diagnostic accuracy of the xpert MTB/RIF assay for lymph node tuberculosis: A systematic review and meta-analysis. *Biomed Res. Int.* 2019:4878240. doi: 10.1155/2019/4878240

Yu, J. H., Li, B., Yu, X. X., Du, Y., Yang, H. F., Xu, X. X., et al. (2020). CT-guided core needle biopsy of small (=20 mm) subpleural pulmonary lesions: Value of the long transpulmonary needle path. *J. Clin. Radiol.* 74, 570.e513–570.e18. doi: 10.1016/j.crad.2019.03.019

Zhang, L., Shi, L., Xiao, Z. P., Qiu, H., Peng, P., and Zhang, M. S. (2018). Coaxial technique-promoted diagnostic accuracy of CT-guided percutaneous cutting needle biopsy for small and deep lung lesions. *PLoS One* 13:e0192920. doi: 10.1371/journal.pone.0192920

Zhao, C., Li, X. F., Li, J. Y., Zhang, Y. S., Ren, S. X., Chen, X. X., et al. (2014). Detecting ALK, ROS1 and RET fusion genes in cell block samples. *Transl. Oncol.* 7, 363–367. doi: 10.1016/j.tranon.2014.04.013

Zurstrassen, C. E., Tyng, C. J., Guimaraes, M. D., Barbosa, P., Pinto, C. A. L., Bitencourt, A. G. V., et al. (2020). Functional and metabolic imaging in transthoracic biopsies guided by computed tomography. *Eur. Radiol.* 30, 2041–2048.doi: 10.1007/s00330-019-06591-0

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 15 September 2022
DOI 10.3389/fmicb.2022.993933

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Jiyuan Hu,
New York University, United States
Xuehai Hu,
Huazhong Agricultural University, China
Xie Xianhua,
Gannan Normal University, China

*CORRESPONDENCE
Hongping Guo
guohongping@hbnu.edu.cn

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

RECEIVED 14 July 2022
ACCEPTED 24 August 2022
PUBLISHED 15 September 2022

CITATION
Guo H, Li T and Wen H (2022) Identifying shared genetic loci between coronavirus disease 2019 and cardiovascular diseases based on cross-trait meta-analysis. *Front. Microbiol.* 13:993933. doi: 10.3389/fmicb.2022.993933

Identifying shared genetic loci between coronavirus disease 2019 and cardiovascular diseases based on cross-trait meta-analysis

Hongping Guo[1]*, Tong Li[1] and Haiyang Wen[2]

[1]School of Mathematics and Statistics, Hubei Normal University, Huangshi, China, [2]School of Computational Science and Electronics, Hunan Institute of Engineering, Xiangtan, China

People with coronavirus disease 2019 (COVID-19) have different mortality or severity, and this clinical outcome is thought to be mainly attributed to comorbid cardiovascular diseases. However, genetic loci jointly influencing COVID-19 and cardiovascular disorders remain largely unknown. To identify shared genetic loci between COVID-19 and cardiac traits, we conducted a genome-wide cross-trait meta-analysis. Firstly, from eight cardiovascular disorders, we found positive genetic correlations between COVID-19 and coronary artery disease (CAD, R_g = 0.4075, P = 0.0031), type 2 diabetes (T2D, R_g = 0.2320, P = 0.0043), obesity (OBE, R_g = 0.3451, P = 0.0061), as well as hypertension (HTN, R_g = 0.233, P = 0.0026). Secondly, we detected 10 shared genetic loci between COVID-19 and CAD, 3 loci between COVID-19 and T2D, 5 loci between COVID-19 and OBE, and 21 loci between COVID-19 and HTN, respectively. These shared genetic loci were enriched in signaling pathways and secretion pathways. In addition, Mendelian randomization analysis revealed significant causal effect of COVID-19 on CAD, OBE and HTN. Our results have revealed the genetic architecture shared by COVID-19 and CVD, and will help to shed light on the molecular mechanisms underlying the associations between COVID-19 and cardiac traits.

KEYWORDS

COVID-19, cardiovascular diseases, shared genetics, meta-analysis, GWAS

Introduction

The coronavirus disease 2019 (COVID-19) arises from severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, and it rapidly outbreak since November 2019 and recently become a public health emergency of international concern (The Severe Covid-19 Gwas Group, 2020). Up to now, there have been more than 170 million confirmed cases and nearly 3.9 million deaths globally. However, its etiology is not fully understood.

People with COVID-19 have different mortality or severity, and the clinical outcome are worse in patients with cardiovascular related disorders, which suggests the comorbidity of COVID-19 and cardiovascular diseases (CVD) (Guan et al., 2020b). More evidences showed the concordant result (Guan et al., 2020a; Ruan et al., 2020; Sisnieguez et al., 2020; Wang et al., 2020; Yang et al., 2020). On the one hand, it was reported that hypertension (21.1%) and diabetes (9.7%) ranked as the top two most prevalent comorbidities for COVID-19 (Wang et al., 2020). The odd ratios of hypertension (2.36) and coronary heart disease (3.42) were larger than 1 when comparing severe COVID-19 patients to non-severe cases (Yang et al., 2020). On the one hand, genome-wide association studies (GWAS) have identified several associated-variants involved in COVID-19 and cardiovascular disease-related traits. For example, a gene known as *ERI3* has been associated with COVID-19 related mortality, coronary artery disease and type 2 diabetes (MacArthur et al., 2017). Moreover, COVID-19 cardiovascular epidemiology showed that nearly 12% of COVID-19 cases have been found to have sustained cardiac injuries, COVID-19 might have a direct and indirect effect on the cardiovascular system (Tajbakhsh et al., 2021). The etiologic agent of COVID-19 can infect the heart, vascular tissues, and circulating cells through the host cell receptor for the viral spike protein (Chung et al., 2021). All the above studies lead us to wonder whether the comorbidity between COVID-19 and CVD is due to the potential shared genetic factors. However, there is few genetic study to reveal the common genetic architecture between COVID-19 and CVD. To this end, the goal of this study was to identify genetic loci shared between COVID-19 and cardiac traits by conducting a large-scale genome-wide cross-trait meta-analysis, and provide more knowledge about common molecular mechanisms of them.

Our study mainly includes three parts. Firstly, we estimated both the overall and local genetic correlation between COVID-19 and eight cardiac traits, including coronary artery disease (CAD), type 2 diabetes (T2D), hypertension (HTN), obesity (OBE), high-density lipoproteins (HDL), low-density lipoproteins (LDL), triglycerides (TC), and total cholesterol (TG). Secondly, we carried out a large-scale cross-trait meta-analysis to identify shared genetic loci between trait pairs that showed significant genetic correlation in the first part of the study. Finally, we conducted transcriptome-wide association study (TWAS), pathway enrichment analysis and Mendelian randomization (MR) analysis to obtain more biological insight.

The overall study design is shown in **Figure 1**.

Materials and methods

Data sources

The GWAS summary statistic for COVID-19 was extracted from the Genetics of Mortality in Critical Care (GenOMICC) study, which performed GWAS on 2244 critically ill patients with COVID-19 in 208 UK intensive care units (Pairo-Castineira et al., 2021). We downloaded the summary statistic with European cases vs UK Biobank controls in this study. We also retrieved the summary statistics of eight cardiac traits in the following public available datasets. The summary statistic for CAD was from the Coronary ARtery DIsease Genome Wide Replication and Meta-analysis plus The Coronary Artery Disease Genetics (CARDIoGRAMplusC4D) Consortium (60,801 cases and 123,504 controls) (Nikpay et al., 2015). The summary statistic for T2D was from the Diabetes Genetics Replication and Meta-Analysis (DIAGRAM) Consortium (26,676 cases and 132,532 controls) (Scott et al., 2017). The summary statistic for OBE was from the Genetic Investigation of ANthropometric Traits (GIANT) Consortium (32,858 cases and 65,839 controls) (Berndt et al., 2013). The summary statistic for HTN was from the Genome wide association study ATLAS (GWASATLAS) database (99,665 cases and 189,642 controls) (Watanabe et al., 2019). The summary statistics for four lipid traits (LDL, HDL, TC, and TG) were from the Global Lipids Genetics Consortium (GLGC) Consortium (188,577 samples) (Willer et al., 2013). The details of each summary statistic dataset are provided in **Supplementary Table 1**.

Genome-wide genetic correlation analysis

We employed the high-definition likelihood methodology (Ning et al., 2020) to estimate the genetic correlation between COVID-19 and eight cardiac traits. This approach provides more accurate estimation by fully accounting for linkage disequilibrium (LD) information across the whole genome. The χ^2 statistic of single nucleotide polymorphisms (SNPs) in high LD regions is higher than that of those in low LD regions, and similar results are observed by replacing one study test statistic with the product of two z-scores in the study. We used the reference panel with imputed HapMap3 SNPs, which are based on genotypes in UK Biobank.

Local genetic correlation

We applied ρ-HESS (Shi et al., 2017) to investigate whether COVID-19 and cardiac traits show local genetic correlation. ρ-HESS quantifies the correlation between traits at each LD-independent region of the genome due to genetic variation. A total of approximately 1.5 Mb was used for estimating local genetic heritabilities and genetic covariances from independent LD blocks. We chose the cardiac traits that showed significant genetic correlation with COVID-19 in this analysis, thus, four pairs of traits were included (COVID-19 and CAD, COVID-19 and T2D, COVID-19 and OBE, COVID-19 and HTN). Notice

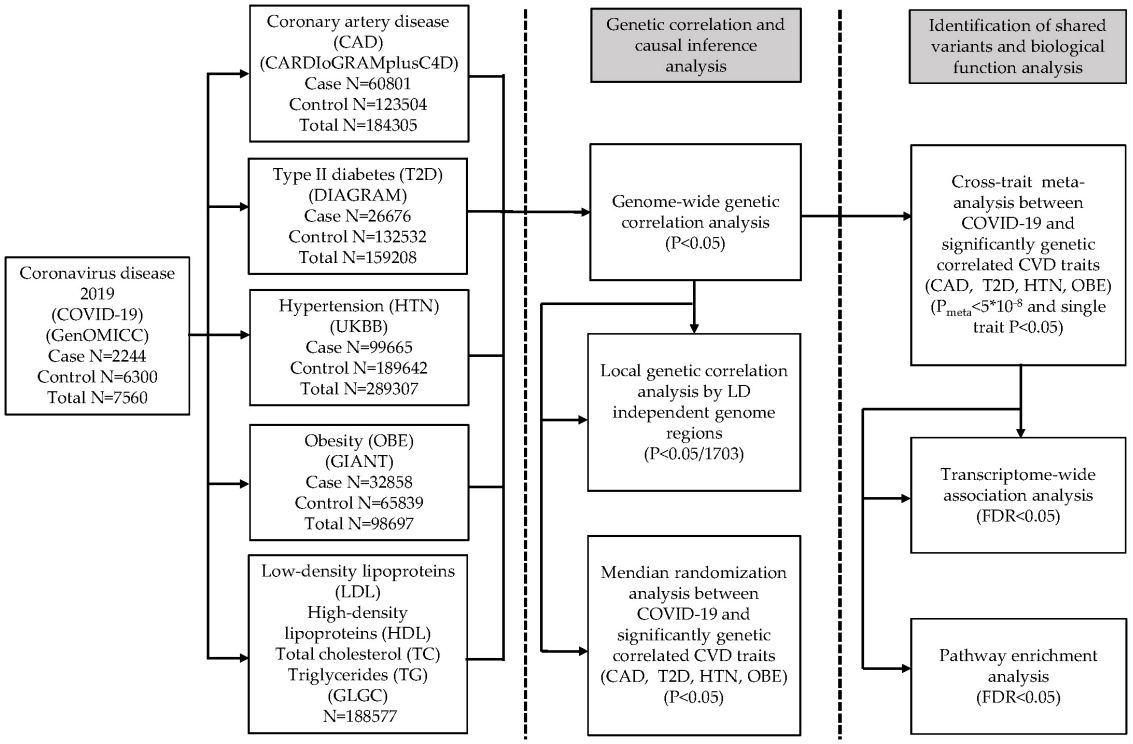


FIGURE 1
Flow chart of the present work.

that we removed the empty loci (with no SNP in it) in each local region in ρ-HESS.

Cross-trait meta-analysis

We conducted a large-scale cross-trait meta-analysis to identify genetic loci shared between severe COVID-19 and cardiac traits that showed significant genetic correlation, using PLEIO framework (Lee et al., 2021). PLEIO is a summary-statistics approach to mapping pleiotropic loci in a multiple trait analysis, either binary, quantitative, independent or correlated traits. Besides, this method can maximize power by adequately modeling the genetic architectures (genetic correlation and heritability) and control false positive rate by accounting for environmental correlation. SNPs with $P_{meta} < 5 \times 10^{-8}$ and trait-specific $P < 0.05$ were considered to be significant for both traits. We performed the operations on a computer of Intel Xeon E5-2695 CPU 2.10 GHz. For each disease pair, it will waste 8–10 mins for the standardization of raw summary statistics first, and then about 2 mins for the identification of pleiotropic loci with PLEIO.

The independent loci were identified using the clumping function of PLINK (version 1.9) tool (Purcell et al., 2007) with clumping parameters $p_1 = 5 \times 10^{-8}$, $p_2 = 1 \times 10^{-5}$, $r^2 = 0.1$, and kb = 500, that is, SNPs with p value less than 1×10^{-5}, r^2 greater than 0.1 and distance less than 500 kb from the peak will be assigned to the clump with that peak. Distance to the nearest gene was calculated using NCBI human genome build37 gene annotation.

Transcriptome-wide association study

We performed transcriptome-wide association study (TWAS) to detect gene expression associations in specific tissues for COVID-19 and cardiac traits, using FUSION software (Gusev et al., 2016) based on 43 Genotype-Tissue Expression Project (GTEx: version 6) tissue expression weights. FUSION is a powerful strategy that uses cis-regulated gene expression measurements to identify genes associated with complex traits through large-scale summary statistics. TWAS p values for each trait were corrected for multiple testing by using Benjamini-Hochberg's False Discovery Rate (FDR) procedure (FDR < 0.05).

Pathway enrichment analysis

To obtain biological insight for shared risk genes that were identified from cross-trait meta-analysis, we used Enrichr

tool (Kuleshov et al., 2016) to perform Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. The Benjamini-Hochberg procedure was used on *p* value to account for multiple testing.

Mendelian randomization analysis

In order to examine the causal relationships between COVID-19 and cardiac traits, we conducted MR analysis using MR-PRESSO test (Verbanck et al., 2018). The MR-PRESSO method estimates exposure effects in multi-instrument MR using SNPs significantly associated with exposure, as well as horizontal pleiotropy in multi-instrument MR utilizing summary statistics. Instruments were constructed using LD-independent SNPs with *p* values lower than 5×10^{-8}.

Results

Overall and local genetic correlations between coronavirus disease 2019 and cardiac traits

We estimated the genetic correlation between COVID-19 and eight cardiac traits using high-definition likelihood method. Four out of eight cardiac traits showed strong and significant genetic correlation with COVID-19. There was the strongest genetic correlation between COVID-19 and CAD (R_g = 0.4075, P = 0.0031), followed by T2D and HTN in a similar magnitude (R_g = 0.232, P = 0.0043 and R_g = 0.233, P = 0.0026, respectively). Moreover, a positive genetic correlation was also found with COVID-19 in OBE (R_g = 0.3451, P = 0.0061). However, no significant genetic correlation was found between COVID-19 and four lipid traits (LDL, HDL, TC, and TG). The detailed results of genetic correlation are displayed in **Table 1**.

TABLE 1 Genetic correlation between coronavirus disease 2019 and cardiac traits.

Phenotype 1	Phenotype 2	R_g	SE	*P*
COVID-19	CAD	0.4075	0.1379	0.0031
	T2D	0.232	0.1147	0.0043
	OBE	0.3451	0.1259	0.0061
	HTN	0.233	0.0774	0.0026
	LDL	0.0335	0.1058	0.7510
	HDL	−0.1923	0.1169	0.1000
	TC	0.0292	0.0852	0.7320
	TG	0.1928	0.1049	0.0661

R_g, genetic correlation estimate; SE, standard error of genetic correlation; COVID-19, coronavirus disease 2019; CAD, coronary artery disease; T2D, type 2 diabetes; OBE, obesity; HTN, hypertension; LDL, low-density lipoproteins; HDL, high-density lipoproteins; TC, total cholesterol; TG, triglycerides.

Due to the significant genetic correlation between COVID-19 and four cardiac traits (CAD, T2D, OBE, and HTN), we conducted ρ-HESS to explore whether there is a genetic correlation between COVID-19 and cardiac traits in certain regions of the genome. Result of the COVID-19/CAD trait pair showed that the 19p13.2 region (chromosome 19: 9238393-11284028) had strong local genetic correlation ($P = 3.76 \times 10^{-6}$). Besides, result of the COVID-19/T2D trait pair showed strong local genetic correlation ($P = 1.39 \times 10^{-7}$) in the 4q21.23 region (chromosome 4: 83372593-84799656). We did not find significant local genetic correlations for neither COVID-19/HTN nor COVID-19/OBE trait pair (**Figure 2**).

Cross-trait meta-analysis results between coronavirus disease 2019 and cardiac traits

We performed a large-scale genome-wide cross-trait meta-analysis to improve the statistical power to identify shared genetic loci between COVID-19 and four cardiac traits that show significant genetic correlations. We considered SNPs with $P_{meta} < 5 \times 10^{-8}$ and trait-specific $P < 0.05$ to be significant for both COVID-19 and cardiac traits. Based on these criteria, we identified 39 independent loci significantly associated with COVID-19 and cardiac traits, of which eight loci failed to be detected in trait-specific GWAS of COVID-19 and cardiac traits (**Tables 2**, **3**).

We observed two overlapped significant loci in the cross-trait meta-analysis of COVID-19/CAD and COVID-19/HTN. The first association signal was 9q34.2 (index SNP: rs495828, $P_{meta} = 1.19 \times 10^{-12}$ for COVID-19/CAD; $P_{meta} = 1.61 \times 10^{-12}$ for COVID-19/HTN). This locus was located at the ABO blood group, which contributed to the immunopathogenesis of SARS-CoV-infection (The Severe Covid-19 Gwas Group, 2020). Similarly, it was concluded that group A individuals had a higher risk of COVID-19 respiratory failure while group O individuals had a protective effect *via* blood type-specific analysis (Deleers et al., 2021). The other locus (index SNP: rs4691707, $P_{meta} = 6.15 \times 10^{-9}$ for COVID-19/CAD; $P_{meta} = 3.03 \times 10^{-10}$ for COVID-19/HTN) was in the intergenic region closet to the *MTND1P22* gene, which may have a role in transcription regulation.

In addition to rs495828 and rs4691707, a further eight loci were identified to be associated with COVID-19 and CAD (**Table 2**). The strongest association signal (index SNP: rs1122608, $P_{meta} = 2.23 \times 10^{-13}$) was found near gene *SMARCA4* on chromosome 19, which was previously reported to regulate atherosclerosis (Ma et al., 2019) and play a protective role to against the risk of HTN (Xiong et al., 2014).

Three loci were identified in a cross-trait meta-analysis of COVID-19 and T2D (**Table 2**). The first locus (index SNP:

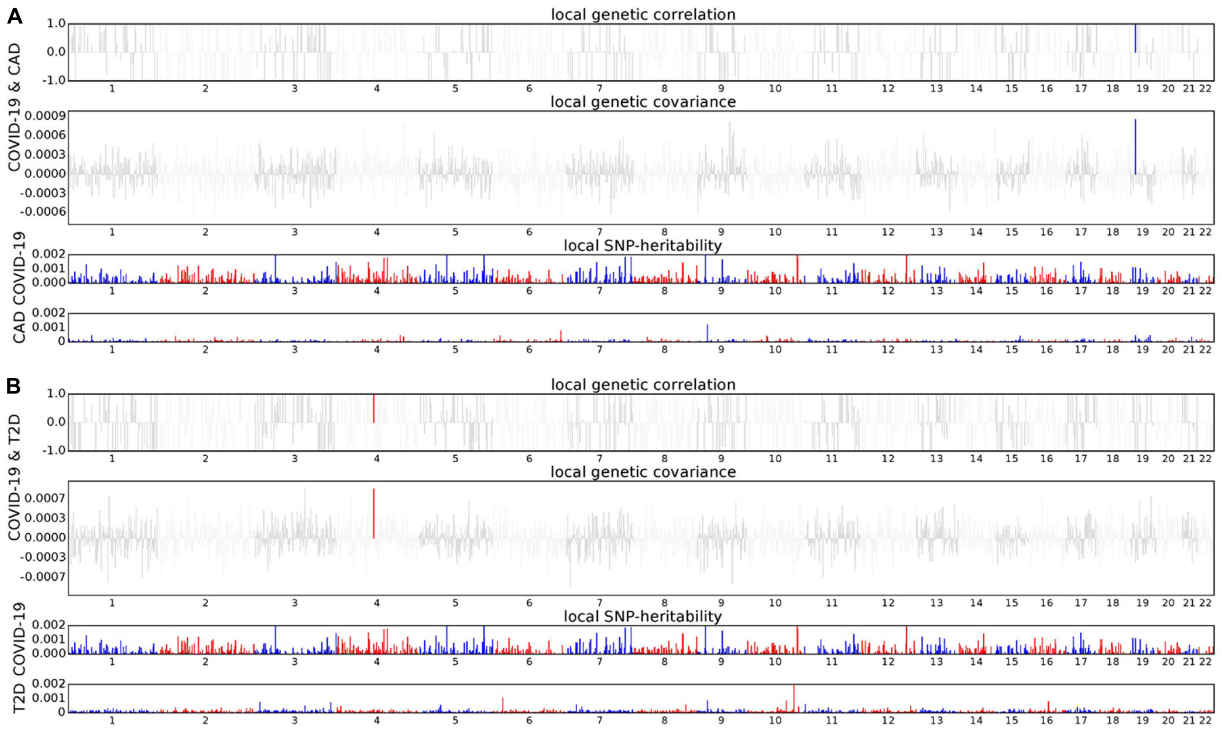


FIGURE 2
Local genetic correlation and local SNP-heritability between COVID-19 and CAD **(A)**, T2D **(B)**, respectively. For each subfigure, the top part represents local genetic correlation, the middle part represents local genetic covariance, and blue or red highlights indicate significant local genetic correlation and covariance after multiple testing correction, the bottom part represents local SNP heritability for each trait.

rs6446490, $P_{meta} = 2.30 \times 10^{-13}$) was mapped on *PPP2R2C*, a gene that increased insulin resistance (Daily et al., 2019). The second locus represented by rs6798189 ($P_{meta} = 1.08 \times 10^{-10}$) was mapped on *ADCY5*, a gene coupled glucose to insulin secretion in human islets (Hodson et al., 2014). The third locus (index SNP: rs1359790, $P_{meta} = 3.89 \times 10^{-9}$) located in intergenic region, which was previously reported to be associated with T2D (Flannick et al., 2019).

We also found five significant loci that were associated with both COVID-19 and OBE (**Table 2**). The top locus (index SNP: rs16917237, $P_{meta} = 8.07 \times 10^{-14}$) was mapped on *BDNF*, a gene was not only associated with body mass index but also CAD (Winkler et al., 2015; van der Harst and Verweij, 2018). The second locus (index SNP: rs3136673, $P_{meta} = 5.90 \times 10^{-10}$) was originally significant associated with COVID-19 ($P = 6.87 \times 10^{-9}$), the mapped gene *CCR1* involved in heart and blood communication in cardiac diseases.

In the cross-trait meta-analysis of COVID-19 and HTN, we identified 21 significant loci (**Table 3**). One of the most important loci is characterized by the *ATP2B1* gene (index SNP: rs1401982, $P_{meta} = 4.32 \times 10^{-32}$), which plays a key role in regulating blood pressure by altering calcium handling and vasoconstriction in vascular smooth muscle cells (Wain et al., 2011).

Results of transcriptome-wide association analysis, pathway enrichment analysis, and Mendelian randomization analysis

To identify association between COVID-19 and cardiac traits with gene expression in specific tissue, we performed TWAS in 43 GTEx tissues. A total of 20 gene-tissue pairs were significantly associated with COVID-19, in addition to 263 gene-tissue pairs with CAD, 142 gene-tissue pairs with T2D, 2030 gene-tissue pairs with HTN, and 256 gene-tissue pairs with OBE (**Supplementary Tables 2–6**). There is no gene-tissue pair overlapped between COVID-19 and the four cardiac traits in TWAS.

To investigate the biological pathways represented by shared genes, we assessed enrichment of shared genes between COVID-19 and cardiac traits. KEGG pathway enrichment analysis revealed cGMP-PKG signaling pathway as the most significant pathway, as well as other signaling pathways and secretion pathways (**Figure 3**).

We identified three significant causal relationships by using MR-PRESSO test, including the effect of COVID-19 on CAD (causal estimate = 0.0045, $P = 3.70 \times 10^{-6}$), OBE (causal estimate = 0.0494, $P = 1.86 \times 10^{-4}$), and HTN (causal

estimate = 0.0019, $P = 1.20 \times 10^{-6}$). However, we did not observed causal effect of COVID-19 on T2D (causal estimate = –0.0026, $P = 0.2281$; **Supplementary Table 7**).

Discussion

To the best of our knowledge, this is the study to identify shared genetic architecture between COVID-19 and cardiac traits. Specifically, we found substantial and significant genetic correlation between COVID-19 and CAD, T2D, OBE, and HTN. These findings are consistent with the study which estimated the genetic correlation by LD score regression method (Chang et al., 2021), and further confirmed the fact that patients with certain underlying medical conditions (such as CAD, T2D, OBE, and HTN) are at increased risk for poor outcome in COVID-19 (Richardson et al., 2020).

In the original GWAS summary statistics, there were hundreds to thousands of significant loci ($P < 5 \times 10^{-8}$) in each of these diseases. However, no shared genetic locus was found between COVID-19 and any of the four cardiac traits. After cross-trait meta-analysis, we identified 10 shared loci between COVID-19 and CAD, three shared loci between COVID-19 and T2D, five shared loci between COVID-19 and OBE, and 21 shared locus between COVID-19 and HTN. This series of comparative data highlights the superiority of cross-trait meta-analysis. These shared genetic loci could be used to predict the occurrence of COVID-19 as well as the abnormal cardiac traits. In addition, we identified eight loci that failed to reach significance in trait-specific GWAS,

TABLE 2 Cross-trait meta-analysis results between coronavirus disease 2019 and CAD, T2D, and OBE ($P_{meta} < 5 \times 10^{-8}$; single trait $P < 0.05$).

Traits	SNP	Genome position	Eff. alle.	Ref. alle.	MAF	COVID-19 *P*	Cardiac trait *P*	Meta OR	Meta *P*	Genes within clumping region
CAD	rs1122608	chr19:10891239–11177408	T	G	0.259	0.017	2.73×10^{-11}	1.08	2.23×10^{-13}	*C19orf38, C19orf52, CARM1, DNM2, SMARCA4, TMED1*, and *YIPF2*
	rs495828	chr9:136154867–136154867	T	G	0.217	0.019	1.29×10^{-10}	0.93	1.19×10^{-12}	*ABO**
	rs6705971	chr2:85755357–85809989	C	A	0.468	0.004	4.52×10^{-10}	0.94	3.23×10^{-12}	*GGCX, MAT2A*, and *VAMP8*
	rs6694817	chr1:154401972–154426264	T	C	0.425	0.037	2.96×10^{-9}	0.95	1.59×10^{-10}	*IL6R*
	rs17678683	chr2:145286559–145286559	G	T	0.091	0.035	3.00×10^{-9}	0.91	2.61×10^{-10}	*LINC01412**
	rs2437935	chr10:44752268–44793299	G	A	0.358	0.014	6.98×10^{-9}	1.06	3.46×10^{-10}	*C10orf142*
	rs4691707	chr4:156441314–156441314	G	A	0.348	0.003	5.95×10^{-7}	0.95	6.15×10^{-9}	*MTND1P22**
	rs17612742	chr4:148401190–148414651	C	T	0.138	0.039	1.61×10^{-7}	0.93	1.29×10^{-8}	*EDNRA*
	rs3002124	chr1:222748085–222748085	G	A	0.293	0.011	7.98×10^{-7}	1.06	2.84×10^{-8}	*TAF1A*
	rs17251589	chr19:41756085–41756906	C	T	0.119	0.025	3.29×10^{-7}	0.92	3.44×10^{-8}	*AXL*
T2D	rs6446490	chr4:6323465–6325086	G	A	0.451	1.00×10^{-4}	1.70×10^{-10}	1.08	2.30×10^{-13}	*PPP2R2C*
	rs6798189	chr3:123095312–123095312	G	A	0.266	0.040	1.30×10^{-10}	0.91	1.08×10^{-10}	*ADCY5*
	rs1359790	chr13:80717156–80717156	G	A	0.288	0.011	1.40×10^{-8}	0.92	3.89×10^{-9}	Intergenic region
OBE	rs16917237	chr11:27702383–27702383	T	G	0.204	0.048	3.60×10^{-11}	1.11	8.07×10^{-14}	*BDNF*
	rs3136673	chr3:46031957–46272440	T	C	0.086	6.87×10^{-9}	0.0093	1.06	5.90×10^{-10}	*CCR1, FYCO1*, and *XCR1*
	rs7189927	chr16:28913787–28922149	C	T	0.356	0.013	3.40×10^{-7}	1.07	6.07×10^{-10}	*ATP2A1* and *RABEP2*
	rs1541984	chr2:25079770–25100328	G	A	0.428	0.049	1.80×10^{-8}	1.07	7.42×10^{-10}	*ADCY3*
	rs1766530	chr6:97576742–97576742	A	G	0.314	2.40×10^{-3}	6.90×10^{-6}	1.06	6.42×10^{-9}	*KLHL32* and *MIR548H3*

*The nearest genes to these loci. COVID-19, coronavirus disease 2019; CAD, coronary artery disease; T2D, type 2 diabetes; OBE, obesity; SNP, single nucleotide polymorphisms; chr, chromosome; Eff. alle., effect allele; Ref. alle., reference allele; MAF, minor allele frequency; OR, odds ratios.

demonstrating cross-trait meta-analysis' excellent statistical power similarly.

We performed GWAS-Catalog analysis to understand whether the shared genes have been reported in the previous studies (**Supplementary Table 8**). Gene ABO, mapped by the locus rs495828 in 9q34.2 region, was reported to be associated with COVID-19, CAD, OBE and HTN (Covid-19 Host Genetics Initiative., 2021). Additionally three genes (CCR1, FYCO1, and XCR1) were not only associated with COVID-19, but also at least two cardiac traits (Shelton et al., 2021). Beyond them, other shared genes were newly found. In terms of gene function, ABO, which determines blood type, may affect COVID-19 disease severity, but there was no evidence to confirm ABO blood group influences risk of COVID-19 infection or outcome (Lehrer and Rheinstein, 2021). Genes CCR1, FYCO1, and XCR1 were involved in T-cell and dendritic-cell function (Kaser, 2020).

In the local genetic correlation analysis between COVID-19 and cardiac traits that showed significant genetic correlation, we found that *SMARCA4* region to have genetic correlation between COVID-19 and CAD, which was also identified by

TABLE 3 Cross-trait meta-analysis result between coronavirus disease 2019 and HTN ($P_{meta} < 5 \times 10^{-8}$; single trait $P < 0.05$).

Traits	SNP	Genome position	Eff. alle.	Ref. alle.	MAF	COVID-19 *P*	HTN *P*	Meta OR	Meta *P*	Genes within clumping region
HTN	rs1401982	chr12:89989599–90441215	G	A	0.413	0.0056	8.50×10^{-28}	0.94	4.32×10^{-32}	*ATP2B1*
	rs35441	chr12:115552499–115553115	T	C	0.383	0.0256	3.04×10^{-25}	0.94	1.19×10^{-28}	Intergenic region
	rs2137320	chr11:1884342–1884342	A	G	0.387	0.022	3.82×10^{-23}	1.06	1.34×10^{-25}	*LSP1*
	rs17080093	chr6:150989698–151027008	T	C	0.069	0.0232	3.86×10^{-20}	0.90	3.23×10^{-22}	*PLEKHG1*
	rs936228	chr15:75131661–75225415	T	C	0.277	0.0081	9.97×10^{-19}	1.05	6.33×10^{-21}	*COX5A, FAM219B, MPI, SCAMP2,* and *ULK3*
	rs3942852	chr11:48028343–48136990	C	T	0.209	0.0329	9.92×10^{-19}	0.94	7.33×10^{-20}	*PTPRJ*
	rs6055976	chr20:8629857–8630692	A	C	0.229	0.0442	1.56×10^{-17}	0.94	1.25×10^{-18}	*PLCB1*
	rs2279500	chr1:113230394–113248791	T	C	0.167	0.0017	1.31×10^{-12}	0.95	1.87×10^{-13}	*MOV10* and *RHOC*
	rs17419291	chr5:87780432–88178683	C	T	0.086	0.0079	1.29×10^{-12}	0.93	3.20×10^{-13}	*MEF2C*
	rs2242261	chr11:47038220–47282024	G	T	0.155	0.0424	4.34×10^{-13}	0.94	1.05×10^{-12}	*ACP2, ARFGAP2, C11orf49, DDB2, NR1H3,* and *PACSIN3*
	rs495828	chr9:136139265–136154867	T	G	0.217	0.0187	1.11×10^{-12}	0.95	1.61×10^{-12}	*ABO*
	rs7716011	chr5:157525853–157525853	G	T	0.252	0.0366	6.21×10^{-12}	1.04	2.48×10^{-11}	*LINC02056**
	rs3744251	chr17:7760983–7760983	A	G	0.076	0.0483	3.94×10^{-11}	1.07	1.84×10^{-10}	*NAA38*
	rs1918966	chr3:169098791–169181582	A	G	0.455	0.0342	5.56×10^{-11}	1.04	2.04×10^{-10}	*MECOM*
	rs4691707	chr4:156441314–156499985	G	A	0.348	0.0025	1.36×10^{-9}	1.04	3.03×10^{-10}	*MTND1P22**
	rs11858678	chr15:41353079–41542591	G	A	0.428	0.0362	1.00×10^{-10}	1.04	3.64×10^{-10}	*CHP1, EXD1,* and *INO80*
	rs7254154	chr19:17169936–17178119	C	A	0.410	0.0258	7.99×10^{-9}	1.03	1.54×10^{-8}	*HAUS8*
	rs2228615	chr19:10403368–10403368	A	G	0.377	4.45×10^{-6}	7.67×10^{-6}	0.97	3.41×10^{-8}	*ICAM5*
	rs11707155	chr3:53608306–53608306	G	A	0.038	0.0183	3.14×10^{-8}	1.09	4.23×10^{-8}	*CACNA1D*
	rs3809278	chr12:111725185–111725185	A	C	0.130	0.0222	2.78×10^{-8}	0.95	4.31×10^{-8}	*CUX2*
	rs2348427	chr4:111414399–111414399	T	C	0.447	0.0016	5.00×10^{-7}	1.03	4.98×10^{-8}	*ENPEP*

*The nearest genes to these loci. COVID-19, coronavirus disease 2019; HTN, hypertension; SNP, single nucleotide polymorphisms; chr, chromosome; Eff. alle., effect allele; Ref. alle., reference allele; MAF, minor allele frequency; OR, odds ratios.

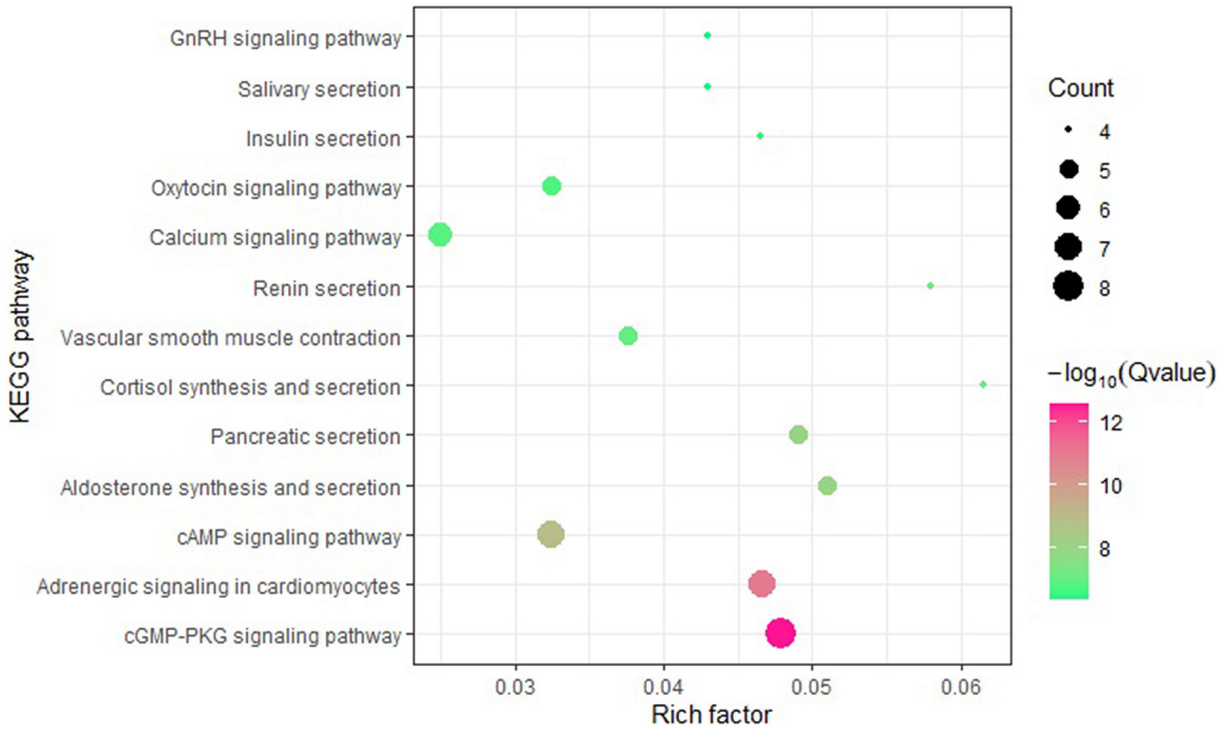


FIGURE 3
Bubble chart of enrichment analysis of shared genes.

cross-trait meta-analysis. *SMARCA4* is a well-known gene associated with CAD, and it mediated nucleosome remodeling which was considered another epigenetic mechanism that can affect the course of COVID-19 (Peng et al., 2020; Shirvaliloo, 2021). Moreover, we also identified *HELQ* region to be significantly associated with COVID-19 and T2D. *HELQ* is predominantly known for its ATP-dependent helicase activity and participation in DNA repair.

Post-GWAS function analyses provided biological insights into the shared genes between COVID-19 and four cardiac traits. In TWAS analysis, we detected 20 significant gene-tissue pair associated with COVID-19, 263 with CAD, 142 with T2D, 256 with OBE and 2030 with HTN. Of these, none of the gene-tissue pair significantly associated with COVID-19 and cardiac traits. In addition, we also performed GTEx tissue enrichment analysis, and did not identify any enrichment signal in tissues. These results suggest that the distribution of pleiotropic genes between COVID-19 and cardiac traits is scattered and not limited to a specific tissue. Moreover, KEGG pathway enrichment analysis showed that the shared genes enriched in some signaling pathways and secretion pathways, such as cGMP-PKG signaling pathway, pancreatic secretion and insulin secretion. The recent studies reported that signaling pathways significantly related to COVID-19 (Messina et al., 2021; Wang et al., 2021), and secretion pathways significantly related to cardiovascular diseases (Chae and Kwon, 2019).

Our MR analysis showed causal effect of COVID-19 on CAD, OBE, and HTN, these findings supported the idea that the genetic correlation of polygenic diseases may be due to both causality and pleiotropy (van Rheenen et al., 2019). Moreover, there is no causal relationship between COVID-19 and T2D, this result indicated the shared genetic effect between COVID-19 and T2D is more likely to be pleiotropic effect, rather than causal effect or mechanism.

As well as genetic factors, environmental factors and lifestyle also play an important role in the comorbidity of COVID-19 and cardiac traits. Although there are many studies on screening anti-SARS-CoV-2 drugs and discovering potential therapeutic drugs for COVID-19 (Zhou et al., 2020; Peng et al., 2021; Shen et al., 2022; Tian et al., 2022), home quarantine and staying away from infection prevention, vaccination, appropriate immunomodulatory diet and drugs that modulate cardiovascular system are currently the most effective approach to prevention (Lotfi et al., 2020).

We also acknowledge some limitations of our work. First, we restricted the analysis to the participants of European ancestors to avoid population stratification, so the findings may not be applicable to general populations. Second, we observed some positive genetic correlation between COVID-19 and TG as well as negative genetic correlation between COVID-19 and HDL, but they failed to reach the standard significant level. The genetic relationship between severe COVID-19 and lipid traits deserves

further study. Third, although large sample cohorts were used in this study, we did not perform replication with other COVID-19 cohorts, which would be meaningful to confirm our findings.

Conclusion

In conclusion, our genome-wide cross-trait meta-analysis confirmed the association between COVID-19 and cardiovascular disorders. Investigation of the shared genetic loci between COVID-19 and cardiac traits can be helpful to understand the common biological mechanisms underlying the comorbidity.

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 author.

Author contributions

HG conceived the project, conducted data analysis, and wrote the manuscript. HW performed the methodology and software. TL participated in the discussion and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by Natural Science Foundation of Hubei Province (Grant No. 2021CFB197), Young Talents Project of Scientific Research Plan of Hubei Provincial Department of Education (Grant No. Q20212506), and Talent Introduction Project of Hubei Normal University in 2021 (Grant No. HS2021RC013).

Acknowledgments

The authors would like to thank the reviewers and editors for the valuable comments and suggestions.

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.2022.993933/full#supplementary-material

References

Berndt, S. I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M. F., et al. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* 45, 501–512. doi: 10.1038/ng.2606

Chae, C. W., and Kwon, Y. W. (2019). Cell signaling and biological pathway in cardiovascular diseases. *Arch. Pharmacal. Res.* 42, 195–205. doi: 10.1007/s12272-019-01141-0256

Chang, X., Li, Y., Nguyen, K., Qu, H., Liu, Y., Glessner, J., et al. (2021). Genetic correlations between COVID-19 and a variety of traits and diseases. *Innovation* 2:100112. doi: 10.1016/j.xinn.2021.100112

Chung, M. K., Zidar, D. A., Bristow, M. R., Cameron, S. J., Chan, T., Harding, C. V., et al. (2021). COVID-19 and cardiovascular disease: From bench to beside. *Circul. Res.* 128, 1214–1236. doi: 10.1161/CIRCRESAHA.121.317997

Covid-19 Host Genetics Initiative. (2021). Mapping the human genetic architecture of COVID-19. *Nature* 600, 472–477. doi: 10.1038/s41586-021

Daily, J., Liu, M., and Parkb, S. (2019). High genetic risk scores of SLIT3, PLEKHA5 and PPP2R2C variants increased insulin resistance and interacted with coffee and caffeine consumption in middle-aged adults. *Nut. Metabolism Cardiovasc. Dis.* 29, 79–89. doi: 10.1016/j.numecd.2018.09

Deleers, M., Breiman, A., Daubie, V., Maggetto, C., Barreau, I., Besse, T., et al. (2021). Covid-19 and blood groups: ABO antibody levels may also matter. *Int. J. Infect. Dis.* 104, 242–249. doi: 10.1016/j.ijid.2020.12.025

Flannick, J., Mercader, J. M., Fuchsberger, C., Udler, M. S., Mahajan, A., Wessel, J., et al. (2019). Exome sequencing of 20791 cases of type 2 diabetes and 24440 controls. *Nature* 570, 71–76. doi: 10.1038/s41586-019-1231-2

Guan, W., Liang, W., He, J., and Zhong, N. (2020a). Cardiovascular comorbidity and its impact on patients with COVID-19. *Eur. Respiratory J.* 55:2001227. doi: 10.1183/13993003.01227-2020

Guan, W., Liang, W., Zhao, Y., Liang, H., Chen, Z., Li, Y., et al. (2020b). Comorbidity and its impact on 1590 patients with COVID-19 in china: A nationwide analysis. *Eur. Respiratory J.* 55:2000547. doi: 10.1183/13993003.00547-2020

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48, 245–252. doi: 10.1038/ng.3506

Hodson, D. J., Mitchell, R. K., Marselli, L., Pullen, T. J., Brias, S. G., Semplici, F., et al. (2014). ADCY5 couples glucose to insulin secretion in human islets. *Diabetes* 63, 3009–3021. doi: 10.2337/db13-1607

Kaser, A. (2020). Genetic risk of severe Covid-19. *New England J. Med.* 383, 1590–1591. doi: 10.1056/NEJMe2025501

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44:W90–W97. doi: 10.1093/nar/gkw377

Lee, C. H., Shi, H., Pasaniuc, B., Eskin, E., and Han, B. (2021). PLEIO: A method to map and interpret pleiotropic loci with GWAS summary statistics. *Am. J. Hum. Genet.* 108, 36–48. doi: 10.1016/j.ajhg.2020.11.017

Lehrer, S., and Rheinstein, P. H. (2021). ABO blood groups, COVID-19 infection and mortality. Blood Cells, *Mol. Dis.* 89:102571. doi: 10.1016/j.bcmd.2021.102571

Lotfi, M., Hamblin, M. R., and Reza, N. (2020). COVID-19: Transmission, prevention, and potential therapeutic opportunities. *Clin. Chimica Acta* 508, 254–266. doi: 10.1016/j.cca.2020.05.044

Ma, H., He, Y., Bai, M., Zhu, L., He, X., Wang, L., et al. (2019). The genetic polymorphisms of ZC3HC1 and SMARCA4 are associated with hypertension risk. *Mol. Genet. Genomic Med.* 7:e942. doi: 10.1002/mgg3.942

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45:D896–D901. doi: 10.1093/nar/gkw1133

Messina, F., Giombini, E., Montaldo, C., Sharma, A. A., Zoccoli, A., Sekaly, R.-P., et al. (2021). Looking for pathways related to COVID-19: Confirmation of pathogenic mechanisms by SARS-CoV-2-host interactome. *Cell Death Dis.* 12:788. doi: 10.1038/s41419-021-03881-8

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., et al. (2015). A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* 47, 1121–1130. doi: 10.1038/ng.3396

Ning, Z., Pawitan, Y., and Shen, X. (2020). High-definition likelihood inference of genetic correlations across human complex traits. *Nat. Genet.* 52, 859–864. doi: 10.1038/s41588-020-0653-y

Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A. D., Rawlik, K., Pasko, D., et al. (2021). Genetic mechanisms of critical illness in COVID-19. *Nature* 591, 92–98. doi: 10.1038/s41586-020-03065-y

Peng, L., Shen, L., Xu, J., Tian, X., Liu, F., Wang, J., et al. (2021). Prioritizing antiviral drugs against SARS-CoV-2 by integrating viral complete genome sequences and drug chemical structures. *Sci. Rep.* 11:6248. doi: 10.1038/s41598-021-83737-5

Peng, L., Tian, X., Shen, L., Kuang, M., Li, T., Tian, G., et al. (2020). Identifying effective antiviral drugs against SARS-CoV-2 by drug repositioning through virus-drug association prediction. *Front. Genet.* 11:577387. doi: 10.3389/fgene.2020.577387

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Richardson, S., Hirsch, J. S., Narasimhan, M., Crawford, J. M., McGinn, T., Davidson, K. W., et al. (2020). Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York city area. *J. Am. Med. Assoc.* 323, 2052–2059. doi: 10.1001/jama.2020.6775

Ruan, Q., Yang, K., Wang, W., Jiang, L., and Song, J. (2020). Clinical predictors of mortality due to COVID -19 based on an analysis of data of 150 patients from Wuhan. China. *Intensive Care Med.* 46, 846–848. doi: 10.1007/s00134-020-05991-x

Scott, R. A., Scott, L. J., Mägi, R., Marullo, L., Gaulton, K. J., Kaakinen, M., et al. (2017). An expanded genome-wide association study of type 2 diabetes in europeans. *Diabetes* 66, 2888–2902. doi: 10.2337/db16-1253

Shelton, J. F., Shastri, A. J., Ye, C., Weldon, C. H., Filshtein-Sonmez, T., Coker, D., et al. (2021). Transancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. *Nat. Genet.* 53, 801–808. doi: 10.1038/s41588-021-00854-7

Shen, L., Liu, F., Huang, L., Liu, G., Zhou, L., and Peng, L. (2022). VDA-RWLRLS: An anti-SARS-CoV-2 drug prioritizing framework combining an unbalanced bi-random walk and Laplacian regularized least squares. *Comput. Biol. Med.* 140:105119. doi: 10.1016/j.compbiomed.2021.105119

Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* 101, 737–751. doi: 10.1016/j.ajhg.2017.09.022

Shirvaliloo, M. (2021). Epigenomics in COVID-19, the link between DNA methylation, histone modifications and SARS-CoV-2 infection. *Epigenomics* 13:10. doi: 10.2217/epi-2021-0057

Sisnieguez, C. E. L., Espeche, W. G., and Salazar, M. R. (2020). Arterial hypertension and the risk of severity and mortality of COVID-19. *Eur. Respiratory J.* 55:2001148. doi: 10.1183/13993003.01148-2020

Tajbakhsh, A., Gheibi Hayat, S. M., Taghizadeh, H., Akbari, A., Inabadi, M., Savardashtaki, A., et al. (2021). COVID-19 and cardiac injury: Clinical manifestations, biomarkers, mechanisms, diagnosis, treatment, and follow up. *Exp. Rev. Anti Infect. Therapy* 19, 345–357. doi: 10.1080/14787210

The Severe Covid-19 Gwas Group. (2020). Genome-wide association study of severe Covid-19 with respiratory failure. *N. England J. Med.* 383, 1522–1534. doi: 10.1056/NEJMoa2020283

Tian, X., Shen, L., Gao, P., Huang, L., Liu, G., Zhou, L., et al. (2022). Discovery of potential therapeutic drugs for COVID-19 through logistic matrix factorization with kernel diffusion. *Front. Microbiol.* 13:740382. doi: 10.3389/fmicb.2022.740382

van der Harst, P., and Verweij, N. (2018). Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circul. Res.* 122, 433–443. doi: 10.1161/CIRCRESAHA.117.312086

van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H., and Wray, N. R. (2019). Genetic correlations of polygenic disease traits: From theory to practice. *Nat. Rev. Genet.* 20, 567–581. doi: 10.1038/s41576-019-0137-z

Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* 50, 693–698. doi: 10.1038/s41588-018-0099-7

Wain, L. V., Verwoert, G. C., O'Reilly, P. F., Shi, G., Johnson, T., Johnson, A. D., et al. (2011). Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nat. Genet.* 43, 1005–1011. doi: 10.1038/ng.922

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., et al. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *J. Am. Med. Assoc.* 323, 1061–1069. doi: 10.1001/jama.2020.1585

Wang, J., Wang, C., Shen, L., Zhou, L., and Peng, L. (2021). Screening potential drugs for COVID-19 based on bound nuclear norm regularization. *Front. Genet.* 12:749256. doi: 10.3389/fgene.2021.749256

Watanabe, K., Stringer, S., Frei, O., Mirkov, M. U., de Leeuw, C., Polderman, T. J., et al. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* 51, 1339–1348. doi: 10.1038/s41588-019-0481-0

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* 45, 1274–1283. doi: 10.1038/ng.2797

Winkler, T. W., Justice, A. E., Graff, M., Barata, L., Feitosa, M. F., Chu, S., et al. (2015). The influence of age and sex on genetic associations with adult body size and shape: A large-scale genome-wide interaction study. *PLoS Genet.* 570:e1005378. doi: 10.1371/journal.pgen.1005378

Xiong, X., Xu, C., Zhang, Y., Li, X., Wang, B., Wang, F., et al. (2014). BRG1 variant rs1122608 on chromosome 19p13.2 confers protection against stroke and regulates expression of pre-mRNA-splicing factor SFRS3. *Hum. Genet.* 133, 499–508. doi: 10.1007/s00439-013-1389-x

Yang, J., Zheng, Y., Gou, X., Pu, K., Chen, Z., Guo, Q., et al. (2020). Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: A systematic review and meta-analysis. *Int. J. Infect. Dis.* 94, 91–95. doi: 10.1016/j.ijid.2020.03.017

Zhou, L., Wang, J., Liu, G., Lu, Q., Dong, R., Tian, G., et al. (2020). Probing antiviral drugs against SARS-CoV-2 through virus-drug association prediction based on the KATZ method. *Genomics* 112, 4427–4434. doi: 10.1016/j.ygeno.2020.07.044

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 16 September 2022
DOI 10.3389/fmicb.2022.1007831

OPEN ACCESS

EDITED BY
Fei Ma,
Chinese Academy of Medical Sciences and Peking Union Medical College, China

REVIEWED BY
Min Chen,
Hunan Institute of Technology, China
Cangzhi Jia,
Dalian Maritime University, China

*CORRESPONDENCE
Changjun Li
licj@ouc.edu.cn
Lei Ji
jilei123@hust.edu.cn
Jialiang Yang
yangjl@geneis.cn

†These authors have contributed equally to this work

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

RECEIVED 31 July 2022
ACCEPTED 01 September 2022
PUBLISHED 16 September 2022

CITATION
Yuan X, Wang Z, Li C, Lv K, Tian G, Tang M, Ji L and Yang J (2022) Bacterial biomarkers capable of identifying recurrence or metastasis carry disease severity information for lung cancer. *Front. Microbiol.* 13:1007831. doi: 10.3389/fmicb.2022.1007831

Bacterial biomarkers capable of identifying recurrence or metastasis carry disease severity information for lung cancer

Xuelian Yuan[1†], Zhina Wang[2†], Changjun Li[1*], Kebo Lv[1], Geng Tian[3,4], Min Tang[5], Lei Ji[3,4*] and Jialiang Yang[3,4,6*]

[1]School of Mathematical Sciences, Ocean University of China, Qingdao, China, [2]Department of Respiratory and Critical Care, Emergency General Hospital, Beijing, China, [3]Geneis Beijing Co., Ltd., Beijing, China, [4]Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China, [5]School of Life Sciences, Jiangsu University, Zhenjiang, Jiangsu, China, [6]Chifeng Municipal Hospital, Chifeng, China

Background: Local recurrence and distant metastasis are the main causes of death in patients with lung cancer. Multiple studies have described the recurrence or metastasis of lung cancer at the genetic level. However, association between the microbiome of lung cancer tissue and recurrence or metastasis remains to be discovered. Here, we aimed to identify the bacterial biomarkers capable of distinguishing patients with lung cancer from recurrence or metastasis, and how it related to the severity of patients with lung cancer.

Methods: We applied microbiome pipeline to bacterial communities of 134 non-recurrence and non-metastasis (non-RM) and 174 recurrence or metastasis (RM) samples downloaded from The Cancer Genome Atlas (TCGA). Co-occurrence network was built to explore the bacterial interactions in lung cancer tissue of RM and non-RM. Finally, the Kaplan–Meier survival analysis was used to evaluate the association between bacterial biomarkers and patient survival.

Results: Compared with non-RM, the bacterial community of RM had lower richness and higher Bray–Curtis dissimilarity index. Interestingly, the co-occurrence network of non-RM was more complex than RM. The top 500 genera in relative abundance obtained an area under the curve (AUC) of 0.72 when discriminating between RM and non-RM. There were significant differences in the relative abundances of *Acidovorax*, *Clostridioides, Succinimonas,* and *Shewanella*, and so on between RM and non-RM. These biomarkers played a role in predicting the survival of lung cancer patients and were significantly associated with lung cancer stage.

Conclusion: This study provides the first evidence for the prediction of lung cancer recurrence or metastasis by bacteria in lung cancer tissue. Our results highlights that bacterial biomarkers that distinguish RM and non-RM are also associated with patient survival and disease severity.

KEYWORDS

lung cancer, recurrence or metastasis, bacterial community, random forest, survival

Introduction

Lung cancer is still the leading cause of cancer deaths worldwide. Local recurrence and distant metastasis are the primary causes of morbidity and mortality, and account for up to 95% of deaths related to lung cancer (Seyfried and Huysentruyt, 2013). Despite advances in therapeutic strategies, especially targeted therapy and immunotherapy, the prognosis remains poor because most patients have extensive metastases at diagnosis (Herbst et al., 2018; Liu et al., 2021). Clinically, a large number of patients with early-stage lung cancer relapse after surgery due to the neglected distant metastasis (Lu Y. et al., 2021). Thus, capturing the signal of metastasis in patients with early-stage lung cancer and continuously monitoring cancer progression after surgery is of great significance for reducing patient mortality.

Growing research has suggested that microbial communities influence the occurrence, progression, metastasis, and response to therapy of multiple cancers (Cullin et al., 2021; Yang M. et al., 2022). For example, studies have shown that *Fusobacterium nucleatum* may trigger cancer through multiple ways, and is related to cancer cell invasion and metastasis (Bullman et al., 2017). Recently, Bertocchi et al. found that intratumoral CRC-associated *Escherichia coli* could migrate to the liver following gut vascular barrier disruption and then prime the liver microenvironment to directly promote metastasis (Bertocchi et al., 2021). In addition, multiple studies have shown that enterotoxigenic *Bacteroides fragilis* could encode a toxin that ultimately induces chronic intestinal inflammation and tissue damage in colorectal cancer by targeting intestinal cells (Boleij et al., 2015; Cheng et al., 2020). However, the potential association between microbial communities of cancer tissue and lung cancer metastasis remains a knowledge gap.

A prominent reason for the high mortality rate of lung cancer is that it is initially asymptomatic and typically discovered at advanced stages (Nasim et al., 2019). Therefore, it is urgent to accurately identify the biomarkers in each stage of lung cancer and adjust the treatment measures for different stages (Yang et al., 2021). Zheng et al. identified 13 gut microbes as biomarkers with high accuracy in predicting early-stage lung cancer by 16 s rRNA sequencing analysis (Zheng et al., 2020). A survey of the gut and sputum microbiota of lung cancer patients at different stages by Lu et al. revealed that these two microbiomes were associated with distant metastasis and that microbial biomarkers across disease stages were largely shared (Lu H. et al., 2021). However, although the potential relationship between gut microbes and non-gut-related cancers is largely unraveled (Erdman and Poutahidis, 2015; Kwa et al., 2016; Zhao F. et al., 2021), the microbes at the original site of cancer development, the cancer tissue, deserve further exploration. The unclear mechanism of tissue microbiome in distant metastasis and lung cancer stage urgently needs to be investigated.

In this study, 174 samples of patients with recurrence or metastasis (RM) and 134 samples of patients without recurrence or metastasis (non-RM) were collected, and the tissue microbiome of all patients with lung cancer was characterized. The main objectives of this study were (1) to identify the bacterial biomarkers capable of discriminating between RM and non-RM, (2) to investigate the effect of smoking on RM and non-RM differential bacteria, and (3) to correlate bacterial biomarkers with survival and disease stage in lung cancer patients. Our study sheds light on the ability of tissue microbial markers of lung cancer to predict recurrence or metastasis and that these biomarkers are strongly associated with the survival and stage of patients with lung cancer.

Materials and methods

Patient cohorts and data preparation

Rob Knight's team rechecked the microbial readings from 18,116 cancer tissue samples included 10,481 patients and 33 cancers in The Cancer Genome Atlas[1] (TCGA; Poore et al., 2020). Of the $6.4'10^{12}$ sequencing readings in TCGA, 7.2% were classified as non-human, of which 35.2% were assigned to bacteria, archaea, or viruses; the sequencing readings were clustered into operational taxonomic units (OTUs) to the genus level by Kraken (Wood and Salzberg, 2014). Microbial sequencing technologies included whole-genome sequencing (WGS) and whole-transcriptome sequencing (RNA-seq). To obtain more tissue samples from lung cancer patients, we downloaded the microbial data obtained by RNA-seq in TCGA database, and obtained clinical indicators and patient information of all samples.

In total, we obtained 308 lung tissue samples from 298 patients with lung cancer. We divided the samples into two groups based on the presence of recurrence or metastasis within 3 years after the initial diagnosis of lung cancer. Specifically, we defined patient samples without recurrence or metastasis within 3 years as non-RM, and defined patient samples with recurrence or metastasis, or both recurrence and metastasis as RM. We obtained 174 RM samples and 134 non-RM samples. We also collected related important clinical indicators of the patients, such as age, gender, TNM stage, smoking history, etc. Specific information for all samples is provided in Table 1.

Network analyses and keystone taxa

We performed network analysis to assess microbiome complexity and identify potential keystone genera for RM and non-RM. We used Spearman's rank correlation to assess the association among genera. We used the *Hmisc* package for calculating correlation coefficients and p values. Correlation coefficients greater than 0.7 with a corresponding p-value less than 0.001 were considered statistically significant. Eligible correlations

1 https://portal.gdc.cancer.gov

TABLE 1 Basic characteristics of study participants.

Characteristics	All ($n = 308$)	RM ($n = 174$)	Non-RM ($n = 134$)
Gender (F/M)	140/168	77/97	63/71
T stage (T1/T2/T3/T4/TX)	77/173/49/8/1	36/92/40/5/1	41/81/9/3/0
N stage (N0/N1/N2/N3/NX)	180/83/39/2/4	97/49/25/1/2	83/34/14/1/2
M stage (M0/M1/MX)	234/7/67	118/6/50	116/1/17
Stage (I/II/III/IV/Unknown)	140/98/60/7/3	64/63/40/6/1	76/35/20/1/2
Age (Mean ± SD)	65.44 ± 9.55	65.66 ± 9.27	65.15 ± 9.88
Histology (LUAD/LUSC)	173/135	114/60	59/75
Smoking history (Never/Reformed smoker≤15 years/ Reformed smoker>15 years/ Current smoker/ Unknown)	30/136/57/71/14	20/68/35/43/8	10/68/22/28/6

LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

are used to generate the networks. The undirected networks were explored and visualized with the interactive platform Gephi (Bastian et al., 2009), using the *Fruchterman-Reingold* layout. Some important topological parameters and node scores of the resulting network are obtained through Gephi (Newman, 2006). In our networks, nodes represented the genera, and the edges represented Spearman's rank correlations. The average degree is the number of edges on each node. Path length and diameter, respectively, represent the nearest distance and the largest distance between two nodes in a network. Clustering coefficient indicates the extent a node is connected to its neighbors. We used high degree to statistically identify the keystone taxa (Banerjee et al., 2019).

Machine-learning classification model and biomarkers identification

We used the microbiome at the genus level as a feature to predict the possibility of recurrence or metastasis of patients in the future. We labeled the patients of RM as "0," and the patients of non-RM as "1." Thus, this problem can be considered a binary classification task. We selected Random Forest (RF) to complete our classification task. RF is used for classification purposes and it had a good performance in recent years. This model was implemented by Python's Sklearn module. We estimated the performance of the classification algorithms using the 5-fold cross-validation (5-fold-cv) procedure. We evaluated the predicted goodness at each abundance level in steps of one hundred. The performance of the classification algorithm was estimated by averaging the area under the curve (AUC) in the 5 test datasets. To ensure comparability, the division of the datasets on each abundance was consistent.

We calculated the variable importance of the top 100 bacteria in relative abundance for identifying RM and non-RM using the Random Forest algorithm. We identified 15 bacterial biomarkers that best discriminated between RM and non-RM based on two variable importance metrics from Random Forest, mean decrease accuracy (MDA) and mean decrease gini (MDGini). Further, Wilcoxon rank-sum test was used to compare the differences of these 15 biomarkers between RM and non-RM. p-value <0.05 was considered statistically significant.

Validation of predictions on survival

From the 15 bacterial biomarkers that discriminate between RM and non-RM, we used the bacteria with the top 6 variables in importance to predict all samples into two groups, recurrence or metastasis (Pred_lable = RM) and without recurrence or metastasis (Pred_lable = non-RM), respectively. Then, overall survival time and status were used to evaluate the prognosis of lung cancer patients in the two groups. The survival curve was performed by using the Kaplan–Meier method and the log-rank test was used to compare the difference in survival probability with R package "survival." p-value <0.05 was considered statistically significant.

Analysis of tissue microbes in different stages of lung cancer

The TNM staging system was first proposed by the French Pierre Denoix between 1943 and 1952(Asare et al., 2019), and later the American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC) gradually began to establish an international system. In 1968, the first edition of the '*TNM Classification of Malignant Tumors*' manual was officially published. It has become the standard method for staging malignant tumors by clinicians and medical scientists.

In the TNM staging system: (1) T refers to the condition of the primary tumor. With the increase in tumor volume and the increase in the extent of adjacent tissue involvement, it is represented by T1 ~ T4 in turn. (2) N refers to the involvement of regional lymph nodes. When the lymph nodes are not involved, it is indicated by N0. With the increase in the degree and scope of lymph node involvement, it is represented by N1 ~ N3 in turn. (3) M refers to distant metastasis (usually blood duct metastasis), M0 is used for those without distant metastasis, and M1 is used for those with distant metastasis. On this basis, use the combination of the three indicators of TNM to draw a specific stage. To

investigate changes in bacterial composition at different stages of lung cancer, we compared the relative abundances of 15 bacterial biomarkers capable of distinguishing metastatic and non-metastatic between different stages. Wilcoxon rank-sum test was used to compare the differences between groups. Besides, taking the T stage as an example, we constructed five-fold cross-validation random forest models with features from the combinations of three bacterial biomarkers (*Dickeya*, *Lactococcus*, and *Pseudogulbenkiania*) to validate the performance of bacterial biomarkers in predicting the tumor stage.

Identification of the patient's smoking history

Based on the smoking history information of patients provided by TCGA, we divided all patients into four groups: smoking history >15 years, smoking history ≤15 years, current smokers, and unknown. Among them, the smoking age of current smoker is not clear, so we focused on comparing the relative abundance of bacterial biomarkers between the two groups of samples with a smoking history of more than 15 years and less than 15 years. Wilcoxon rank-sum test was used to compare the differences between groups.

Statistical analysis

All analyses were implemented with R version 4.1.3[2] and its appropriate packages. Principal coordinate analysis (PCoA) was performed with R package 'ape' based on the Bray-Curtis distance matrix. The Shannon index and Bray–Curtis dissimilarity index were calculated by using the R package "vegan." Non-metric multidimensional scaling (NMDS) was performed with the R package "vegan." Comparison between groups was conducted utilizing Wilcoxon rank-sum test. p-value <0.05 was set as the threshold.

Results

Characteristics of the lung cancer datasets in meta-analysis

A total of 308 lung tissue samples from 298 patients with lung cancer were obtained. We determined the recurrence or metastasis of patients based on the follow-up information provided by TCGA. Detailedly, we defined patients without recurrence or metastasis within three years after the initial diagnosis of lung cancer as non-RM samples, and patients with recurrence, metastasis, and simultaneous recurrence and metastasis within 3 years as RM samples. The demographics and clinical characteristics are provided in Table 1.

Bacterial profile of the lung cancer microbiome is dominated by proteobacteria

Previous microbial studies of lung cancer have shown that bacterial composition of cancerous lungs shifts compared to non-cancerous lungs (Huang et al., 2011); however, these compositional changes have not been examined in distant metastatic lung cancer. To obtain a comprehensive characteristic of the bacterial community of metastatic lung cancer, we thoroughly compared the bacterial compositions of RM and non-RM. As shown in Figure 1, Proteobacteria dominated the tissue microbiome of lung cancer with an average relative abundance of 52.3%, followed by Firmicutes (21.8%) and Actinobacteria (16.0%). Importantly, Proteobacteria was generally more dominant in non-RM (Wilcoxon $p = 0.041$), indicating that this is a recurrent phenomenon in lung cancer (Woerner et al., 2022).

Bacterial composition carry information on recurrence or metastasis in lung cancer

We next computed the firth two principal coordinates based on the Bray–Curtis dissimilarity and the PCoA plot showed two distinct clusters (Figure 1B). The two groups (RM and non-RM) were not randomly dispersed between the two clusters. Instead, enrichment was observed in specific clusters for certain groups, providing further evidence that the bacterial composition may carry RM/non-RM information in lung cancer. Then, we examined the alpha diversity (Shannon) and richness of the microbiome within samples of RM and non-RM. Specifically, there was no significant difference in the Shannon index between RM and non-RM; however, we observed a significant increase in richness in non-RM as compared to RM tissue (Figures 1C,D). Further, we calculated the Bray–Curtis dissimilarity for each pair of samples to measure how different each pair is regarding bacterial composition. Non-RM samples were far more similar to one another than to the RM samples (Wilcoxon $p < 0.0001$; Figure 1E). We detected 1,509 and 1,500 genera in non-RM and RM, respectively, indicating that the vast majority of genera were shared in lung cancer tissues regardless of recurrence or metastasis (Figure 1F).

Co-occurrence networks and keystone taxa of RM and non-RM

We know that bacterial composition in the tissues of lung cancer patients with and without metastasis is different; however, the interaction pattern of bacterial communities in lung cancer

2 http://www.R-project.org

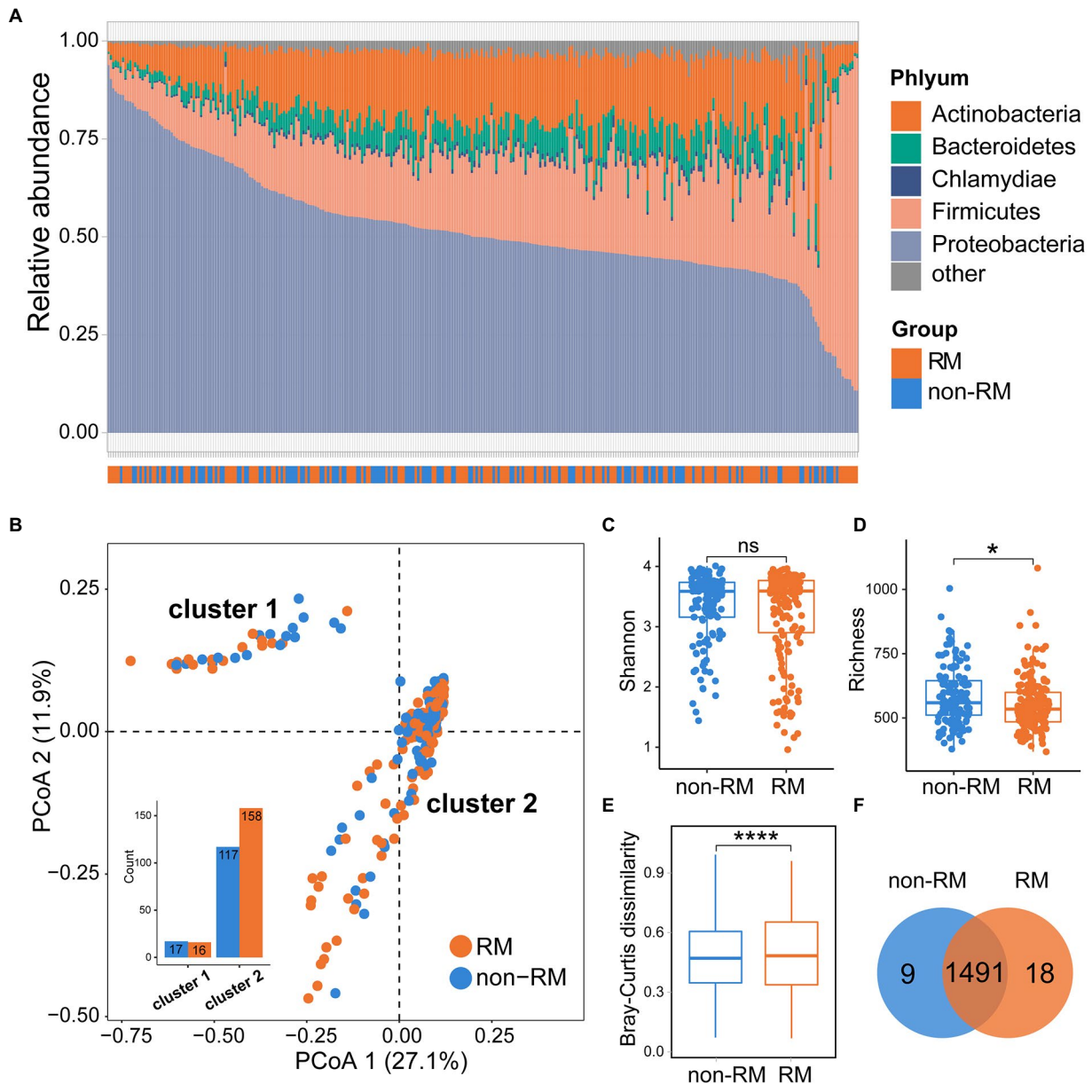
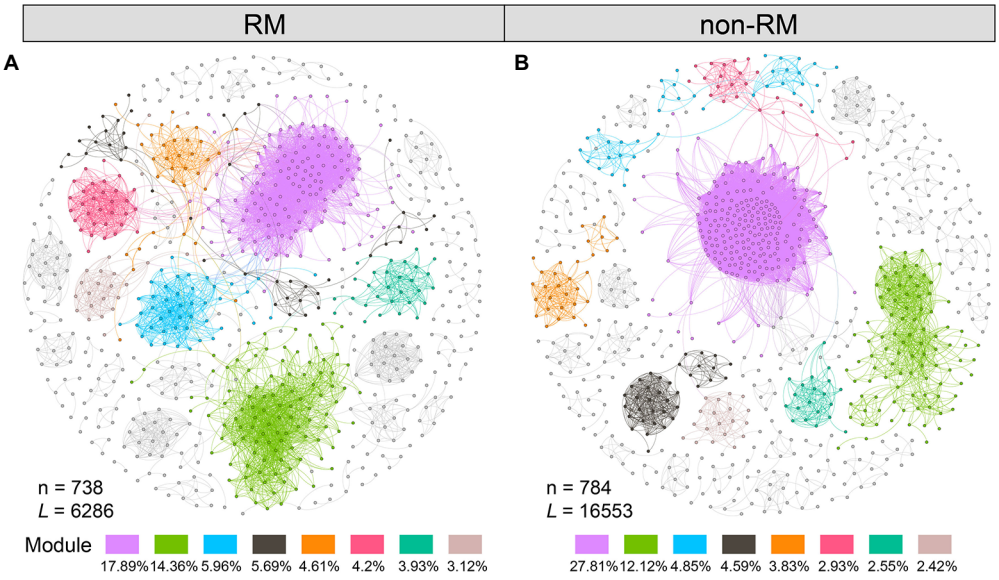


FIGURE 1
Bacterial community structures of RM and non-RM patients with lung cancer. **(A)** Bacteria composition at phylum level for all samples; **(B)** The PCoA plot, on the genus level, colored by group as in panel; Comparisons of **(C)** Shannon index and **(D)** richness of bacterial communities between RM and non-RM; **(E)** Bray–Curtis dissimilarity measures, on the genus level, based for all pairs of samples; **(F)** Common and unique genera between RM and non-RM. Wilcoxon test, * $p<0.05$, **** $p<0.0001$.

tissues has not been disclosed. To reveal the underlying patterns, based on genus pairs with significant positive correlations screened by thresholds, we mapped co-occurrence networks for RM and non-RM, respectively (Figures 2A,B). Network complexity varied considerably between the two groups. Specifically, compared to non-RM, microbial communities in RM had a less complex network with fewer edges (6286), fewer nodes (738), a lower average degree (17.04), and a lower average clustering coefficient (0.78, Figure 2C). In the RM, the keystone genera we detected was *Azoarcus*, while in the non-RM, it was *Variovorax*, *Ramlibacter*, and *Sphaerotilus*. Although limited studies have directly linked these genera to lung cancer metastasis, the difference in keystone certainly implies a divergence in bacterial interactions between RM and non-RM.

Further, we drilled down into the largest module in the network, i.e., with the most nodes, which we called Module 1 in this study. Module1 of RM and non-RM contained 135 and 218 genera, respectively, of which 90 were shared (Figure 2D). NMDS analysis showed that the bacterial composition in Module 1 of the two groups was significantly different (Figure 2E; stress = 0.0889).



C

Network parameters and keystone genera of RM and non-RM

Network	Nodes/edges	Avg. Degree	Path length	Diameter	Clust. Coeff.	Keystone genera
RM	738/6286	17.04	6.14	18	0.78	*Azoarcus*
non-RM	784/16553	42.23	2.56	10	0.8	*Variovorax* *Ramlibacter* *Sphaerotilus*

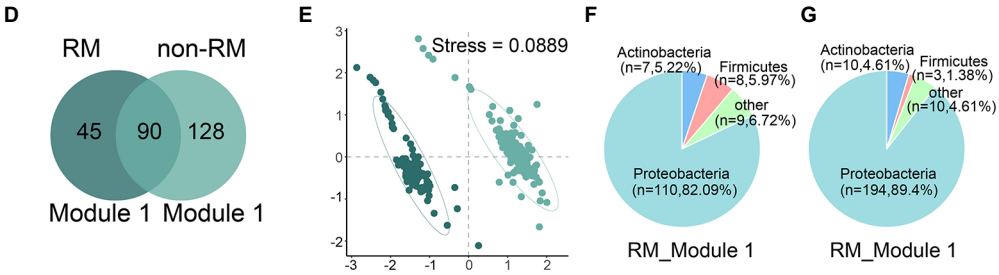


FIGURE 2
Network analysis reveals distinct bacterial community interaction patterns between RM and non-RM. Network of co-occurring bacteria of **(A)** RM and **(B)** non-RM. Only Spearman's correlation coefficient (r>0.7 significant at p<0.001) is shown. The nodes are colored according to module. The percentage indicates the ratio of the number of nodes in the module to the total number; **(C)** Network parameters and the potential keystone genera of RM and non-RM. Average degree is the number of edges on each node. Path length and diameter, respectively, represent the nearest distance and the largest distance between two nodes in a network. Clustering coefficient indicates the extent a node is connected to its neighbors; **(D)** Common and unique genera in the largest modules of RM and non-RM; **(E)** NMDS analysis shows significant differences in the largest modules of RM and non-RM; Phylum-level bacterial composition in the largest modules of **(F)** RM and **(G)** non-RM.

Detailedly, Proteobacteria were the core phylum in these two modules, accounting for 82.1 and 89.4%, respectively (Figures 2F,G), further indicating that Proteobacteria dominated the lung cancer tissue bacterial community.

Bacterial biomarkers for differentiating RM and non-RM are associated with patient outcomes in lung cancer

Given the observed differences in bacterial content between RM and non-RM (Figures 2A,B,3), we reasoned that bacteria might be able to classify recurrence or metastasis of patients with lung cancer. To this end, we constructed a machine-learning classifier to identify recurrence or metastasis from tissue bacteria. The top 500 genera in relative abundance were chosen as a compromise between reduced resolution with more genera and decreased representation of bacterial community with fewer genera. The average AUC of the classifier using bacterial content reached 0.72 (Figure 3A). The performance of our machine-learning classifier provides evidence that bacterial composition contains a signal that tracks recurrence or metastasis.

In addition to providing an algorithm to assign classes (RM/non-RM in our case) based on features (relative abundance of

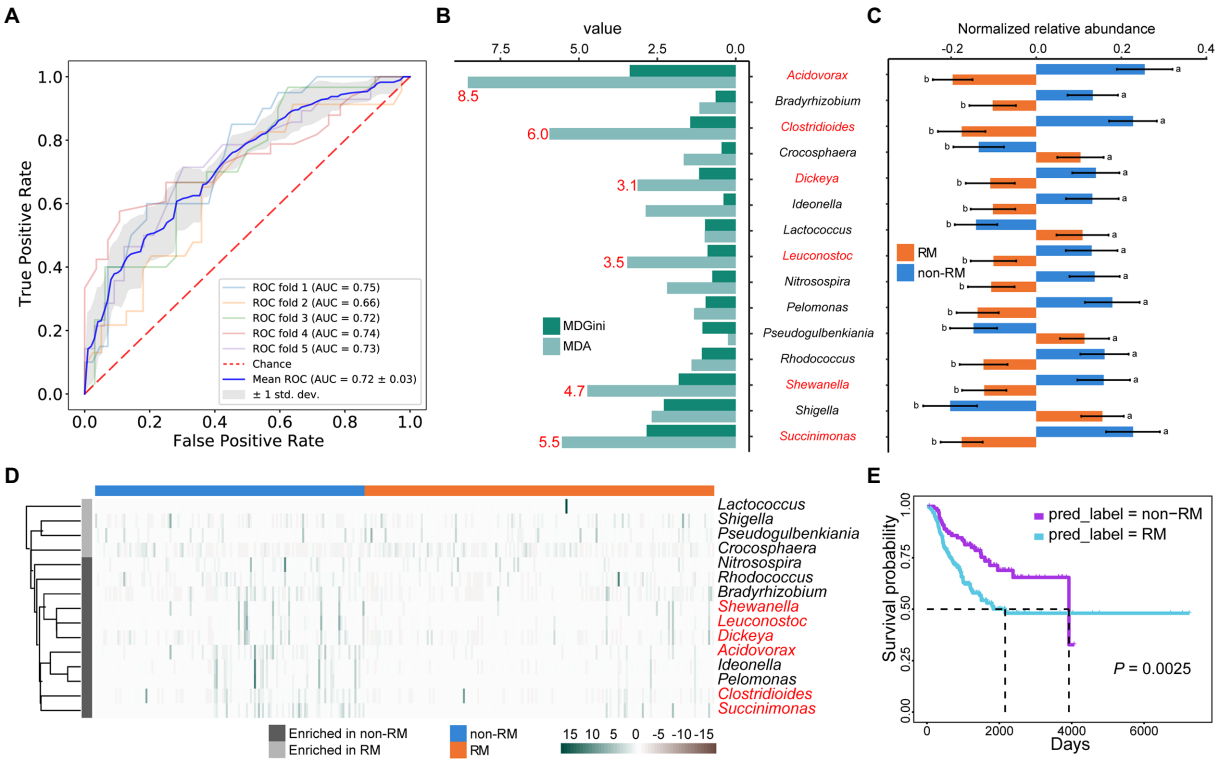


FIGURE 3
Metastasis and survival in lung cancer patients share some bacterial biomarkers. **(A)** Fivefold cross-validation random forest models with feature from top 500 genera in relative abundance; **(B)** The top 15 genera of variable importance predicted by random forest. The top 6 genera of MDA are shown in red; **(C)** The relative abundances of these 15 genera were significantly different between RM and non-RM; **(D)** Heatmap shows the enrichment of these 15 genera in RM and non-RM; **(E)** Kaplan–Meier survival curve shows that non-RM showed a significant overall survival benefit as compared with RM.

bacterial genera). Random Forest also assigns variable importance to each categorical feature. Based on the MDA value, we identified the top 15 genera with variable importance and significant differences between RM and non-RM (Figures 3B,C). *Acidovorax* has been reported to develop as a panel of sputum biomarkers that could diagnose lung squamous cell carcinoma (Leng et al., 2021). It is currently known that patients with lung cancer are at high risk of developing *Clostridium difficile* infection (CDI) due to continued chemotherapy, prolonged hospital stay, and general debility (Hwang et al., 2013). However, we detected a reduced relative abundance of *Clostridium* in recurrence or metastasis lung cancer patients, suggesting that its mechanism in recurrence or metastasis remains to be elucidated. Many other genera that contribute to discrimination between RM and non-RM (Figure 3D) are known to be associated with lung disease or lung cancer chemotherapy outcomes, e.g., *Leuconostoc* (Zhao Z. et al., 2021), *Shigella* (Zhang et al., 2018), *Rhodococcus* (Haramati and Jenny-Avital, 1998), and *Bradyrhizobium* (Jin et al., 2019).

Given the significance of predicting the prognosis of lung cancer patients and the relationship between bacteria and lung cancer patient survival demonstrated by multiple studies (Salazar et al., 2020; Tomita et al., 2020; Zhao Y. et al., 2021), we tried to correlate these bacterial biomarkers with patient survival. First, we selected the top 6 genera of variable importance as biomarkers (Figure 3D), then used these biomarkers to predict the recurrence or metastasis of all lung cancer patients, and finally performed survival analysis on the predicted two groups. As shown in Figure 3E, non-RM showed a significant overall survival benefit as compared with RM ($p = 0.0025$). Our results further prove the accuracy and clinical significance of the bacterial biomarkers we identified, as well as the fact that the patients with lung cancer recurrence or metastasis have reduced survival.

Smoking history influences bacteria that distinguish recurrence or metastasis in lung cancer patients

Smoking is the greatest risk factor for lung cancer, up to 90% of lung cancers can be attributable to smoking (de Groot et al., 2018). Previous studies have demonstrated that nicotine-induced N2-neutrophils have a pro-metastatic role in lung cancer cell colonization (Tyagi et al., 2021). However, whether bacteria are mediators linking smoking and lung cancer recurrence or metastasis is still unknown.

Thus, we associated smoking history with recurrence or metastasis-related bacterial biomarkers (Figure 3B) in lung cancer patients. Coincidentally, we found that the relative abundance of most bacterial biomarkers was reduced in patients with a longer smoking history (Figure 4A). The relative abundances of the genera *Acidovorax*, *Clostridioides*, and *Lactococcus* varied with smoking history (Figures 4B–D). In particular, the relative abundance of the genus *Acidovorax* was significantly higher in patients with a smoking history of less than 15 years than in patients with a smoking history of more than 15 years (Figure 4B). Similarly, we also detected a reduced relative abundance of *Acidovorax* in RM compared to non-RM (Figure 3C). Naturally, we speculate that excessive smoking can cause changes in the content of certain bacteria, which, in turn, promotes the recurrence or metastasis of lung cancer patients.

Bacterial biomarkers of disease stage and lung cancer recurrence or metastasis intersect

It has long been recognized that regional and metastatic cancers have a worse prognosis, and many cancers can be traced back to this gradual progression (Cserni et al., 2018). This has become the basis for cancer staging, including lung cancer. The Tumor-Node-Metastasis (TNM) system established by the Union for International Cancer Control (UICC) has become a worldwide means of describing the anatomical extent of cancer and determining its stage.

We have known that bacterial biomarkers that can distinguish recurrence or metastasis of lung cancer are related to patient survival (Figure 3E), and then we wondered whether these biomarkers also carry disease stage information. Interestingly, we found that the relative abundances of some of these 15 bacterial markers (Figure 2B) varied significantly between stages (Figures 5A–I). For example, the relative abundance of *Dickeya* was significantly lower in T4 patients compared to T2 patients (Figure 5A). *Rhodococcus* has the lowest relative abundance in N3 stage patients compared to other stages (Figure 5F). Then, taking the T stage as an example, we constructed five-fold cross-validation Random Forest models with features from these three biomarkers (Figures 5A–C). As expected, features from the combination of these three bacteria showed capabilities for identifying the T stage for patients with lung cancer (Figure 5J). The genera *Dickeya*, *Lactococcus*, and *Pseudogulbenkiania* displayed the strongest ability to identify the T stage with an average AUC of 0.84.

Wu and his colleagues found that mannan exopolysaccharides (EPS) produced by a subsp. of *Lactococcus lactis* affected the production of inflammatory cytokine (Wu et al., 2016). Similarly, we also detected a decrease in the relative abundance of *Lactococcus* in advanced patients (Figure 5B). In general, our results demonstrate that bacteria capable of discriminating recurrence or metastasis from lung cancer also carry disease stage information, thereby assisting clinicians and medical scientists in staging malignancies.

Discussion

Genetic and environmental factors have long been recognized as contributors to cancer recurrence or metastasis (Bhujwalla et al., 2001; Rosell and Karachaliou, 2015; Song et al., 2020). Recently, histopathological images are also been found capable of predicting cancer recurrence or metastasis (Yang J. et al., 2022; Ye et al., 2022); however, little is known about the tissue microbiome that promotes cancer recurrence or metastasis. We demonstrate that recurrence or metastasis in lung cancer patients is associated with specific bacteria and that smoking significantly affects the relative abundance of these bacteria. In-depth, by building machine-learning classifiers, we found that six recurrence- or metastasis-distinguishing bacterial biomarkers (*Acidovorax*, *Clostridioides*, *Succinimonas*, *Shewanella*, *Leuconostoc*, and *Dickeya*) were associated with survival in lung cancer patients. Further, we verified that bacteria capable of discriminating recurrence or metastasis also carry information on tumor stage in lung cancer, and three genera, *Dickeya*, *Lactococcus*, and *Pseudogulbenkiania*, can accurately predict tumor T stage. Collectively, the above results support our proposal that smoking can lead to changes in the bacterial community in lung cancer tissue, which, in turn, affects tumor metastasis in patients, and the bacteria closely associated with recurrence or metastasis are inseparable from patient prognosis and tumor stage.

The number one risk factor for lung cancer development is tobacco exposure, which outweighs all other factors that lead to lung cancer (Bade and Dela Cruz, 2020). Tobacco smoke contains many potential carcinogens and bacterial products, which can induce epithelial cells to secrete inflammatory cytokines, cause barrier function impairment, and even alter the microbiome to influence lung carcinogenesis (Sapkota et al., 2010; Pauly and Paszkiewicz, 2011; Heijink et al., 2012; Checa et al., 2016). We observed significantly lower relative abundances of certain bacteria, such as *Acidovorax*, in patients with a smoking history of more than 15 years compared with patients with a smoking history of less than 15 years. Similarly, in a study of tumor ($n = 143$) and non-tumor adjacent tissues ($n = 144$), they observed a significant difference in the relative abundance of *Acidovorax* among smokers as compared to non-smokers (Greathouse et al., 2018). In addition, a study of non-malignant lung tissue showed that a greater abundance of *Acidovorax* was specifically found in the extracellular vesicles of smokers (Kim et al., 2017). Innovatively, we are the first to suggest that the relative abundance of *Acidovorax* is reduced in recurrence or metastasis patients compared to without recurrence or metastasis patients, echoing the reduced relative abundance of *Acidovorax* in patients with a longer smoking history. Nevertheless, future studies should mechanistically elucidate the role of *Acidovorax* between tobacco exposure and lung cancer metastasis.

FIGURE 4
Relative abundances of bacterial biomarkers vary in patients with different smoking histories. **(A)** Heatmap shows the relative abundance of 15 bacterial biomarkers in patients with different smoking histories. The relative abundances of **(B)** *Acidovorax*, **(C)** *Clostridioides*, and **(D)** *Lactococcus* are significantly different among patients with different smoking histories.

Lung cancer patients face severe mortality even when detected in the early stages of cancer. Different from other types of cancers that are detected early and have obvious survival advantages, about 35–45% of patients with stage I lung cancer will die due to recurrence within 5 years even if the operation is successful (Molina et al., 2008; Zhao et al., 2017). We have verified that lung cancer patients with recurrence or metastasis are associated with lower survival rates. Further, our machine-learning classifier with features of *Acidovorax*, *Clostridioides*, *Succinimonas*, *Shewanella*, *Leuconostoc*, and *Dickeya* predicted recurrence or metastasis information in lung cancer patients, and patients predicted to be recurrence or metastasis had lower survival rates. Although the mechanism remains to be elucidated by more evidence, the current findings undoubtedly provide guidance for clinicians to preliminarily judge patient survival.

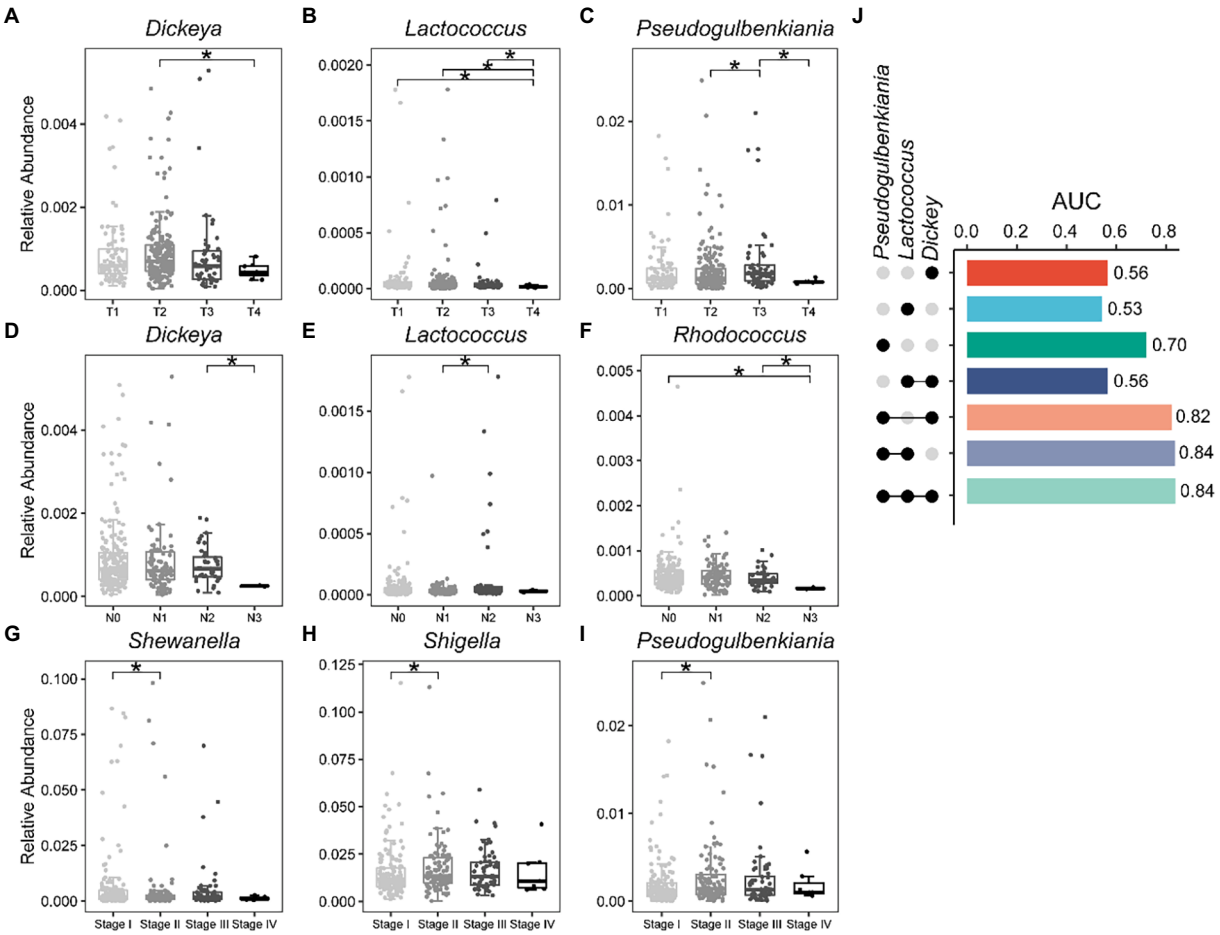


FIGURE 5
Bacterial biomarkers can identify disease stage in lung cancer patients. Relative abundance of specific genera varies significantly between different stages of lung cancer. **(A–C)** T stage; **(D–F)** N stage; **(G–I)** Stage; **(J)** Fivefold cross-validation random forest models with features from *Dickeya*, *Lactococcus,* and *Pseudogulbenkiania* to predict the T stage of lung cancer patients. Wilcoxon test, *$p<0.05$.

Our study showed that bacteria capable of distinguishing recurrence or metastasis can predict tumor stage in patients. In a study of 156 incident lung cancer cases and 156 individually matched controls, they found that species *Lactococcus lactis* was associated with decreased lung cancer risk (Shi et al., 2021). We also detected a decrease in the relative abundance of *Lactococcus* from the T1–T4 stages. *Dickeya*, *Lactococcus*, and *Pseudogulbenkiania* outperformed in predicting T4 and T1 stages in lung cancer patients.

The strength of our findings includes two accurately divided lung cancer cohorts with and without recurrence or metastasis within 3 years, the microbiome at the site of initial cancer development, and detailed follow-up information for nearly all patients. But our research still has some limitations. The distribution of samples in different stages is not uniform; for example, the number of samples in the T4 stage is much smaller than that in the T2 stage, we admit that this may skew the results. Although we comprehensively compared the tissue microbiome of patients without and those with recurrence or metastasis, the absence of healthy controls is a pity. Functional experiments are needed in the future to determine if and how bacteria influence the progression of lung cancer. Such experiments will reveal the potential of bacteria as biomarkers in lung cancer recurrence or metastasis and may provide treatment options for patients. Functional experiments to further provide treatment assistance for lung cancer patients is the focus of our future work.

Conclusion

Through a comprehensive comparison of tissue microbes in recurrence or metastasis and without recurrence or metastasis lung cancer patients, we identified 15 bacterial biomarkers that differentiate between RM and non-RM lung cancer, with the relative abundance of most bacteria decreasing in recurrence or metastasis patients. Besides, six recurrence or metastasis-distinguishing bacterial biomarkers (*Acidovorax*, *Clostridioides*, *Succinimonas*, *Shewanella*, *Leuconostoc*, and *Dickeya*) were associated with survival

in lung cancer patients. Further, we found that patients with longer smoking history were associated with lower abundances of these biomarkers, such as the genus *Acidovorax*. Finally, these bacterial biomarkers (*Dickeya*, *Lactococcus*, and *Pseudogulbenkiania*) accurately predicted the tumor T stage in lung cancer patients. We propose that smoking induces tissue microbial changes in lung cancer patients, which, in turn, promotes recurrence or metastasis in lung cancer patients, and the altered bacteria are associated with patient prognosis and tumor progression. With these results, we foresee a new avenue for mechanistic studies to address the role of microbes in the recurrence or metastasis of lung cancer patients, patient prognosis, and tissue tumor progression monitoring.

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.

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

JY, LJ, and CL conceived this study. XY and ZW performed the study and experiments. KL, GT, and MT collected the data. XY and LJ wrote the manuscript. All authors contributed to the interpretation of data and to the revision of the manuscript.

Conflict of interest

GT, LJ, and JY were employed by Geneis Beijing 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.

References

Asare, E. A., Grubbs, E. G., Gershenwald, J. E., Greene, F. L., and Aloia, T. A. (2019). Setting the "stage" for surgical oncology fellows: Pierre Denoix and TNM staging. *J. Surg. Oncol.* 119:823. doi: 10.1002/jso.25404

Bade, B. C., and Dela Cruz, C. S. (2020). Lung cancer 2020: epidemiology, etiology, and prevention. *Clin. Chest Med.* 41, 1–24. doi: 10.1016/j.ccm.2019.10.001

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

Bastian, M., Heymann, S., and Jacomy, M. (2009). "Gephi: an open source software for exploring and manipulating networks." in *Proceedings of the International AAAI Conference on Web and Social Media*, San Jose, CA, 361–362.

Bertocchi, A., Carloni, S., Ravenda, P. S., Bertalot, G., Spadoni, I., Lo Cascio, A., et al. (2021). Gut vascular barrier impairment leads to intestinal bacteria dissemination and colorectal cancer metastasis to liver. *Cancer Cell* 39, 708.e11–724.e11. doi: 10.1016/j.ccell.2021.03.004

Bhujwalla, Z. M., Artemov, D., Aboagye, E., Ackerstaff, E., Gillies, R. J., Natarajan, K., et al. (2001). The physiological environment in cancer vascularization, invasion and metastasis. *Novartis Found. Symp.* 240:23–38; discussion 38–45, 152–153. doi: 10.1002/0470868716.ch3

Boleij, A., Hechenbleikner, E. M., Goodwin, A. C., Badani, R., Stein, E. M., Lazarev, M. G., et al. (2015). The Bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clin. Infect. Dis.* 60, 208–215. doi: 10.1093/cid/ciu787

Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., et al. (2017). Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. *Science* 358, 1443–1448. doi: 10.1126/science.aal5240

Checa, M., Hagood, J. S., Velazquez-Cruz, R., Ruiz, V., Garcia-De-Alba, C., Rangel-Escareno, C., et al. (2016). Cigarette smoke enhances the expression of Profibrotic molecules in alveolar epithelial cells. *PLoS One* 11:e0150383. doi: 10.1371/journal.pone.0150383

Cheng, W. T., Kantilal, H. K., and Davamani, F. (2020). The mechanism of Bacteroides fragilis toxin contributes to colon cancer formation. *Malays J. Med. Sci.* 27, 9–21. doi: 10.21315/mjms2020.27.4.2

Cserni, G., Chmielik, E., Cserni, B., and Tot, T. (2018). The new TNM-based staging of breast cancer. *Virchows Arch.* 472, 697–703. doi: 10.1007/s00428-018-2301-9

Cullin, N., Azevedo Antunes, C., Straussman, R., Stein-Thoeringer, C. K., and Elinav, E. (2021). Microbiome and cancer. *Cancer Cell* 39, 1317–1341. doi: 10.1016/j.ccell.2021.08.006

de Groot, P. M., Wu, C. C., Carter, B. W., and Munden, R. F. (2018). The epidemiology of lung cancer. *Transl. Lung. Cancer Res.* 7, 220–233. doi: 10.21037/tlcr.2018.05.06

Erdman, S. E., and Poutahidis, T. (2015). Gut bacteria and cancer. *Biochim. Biophys. Acta* 1856, 86–90. doi: 10.1016/j.bbcan.2015.05.007

Greathouse, K. L., White, J. R., Vargas, A. J., Bliskovsky, V. V., Beck, J. A., von Muhlinen, N., et al. (2018). Interaction between the microbiome and TP53 in human lung cancer. *Genome Biol.* 19:123. doi: 10.1186/s13059-018-1501-6

Haramati, L. B., and Jenny-Avital, E. R. (1998). Approach to the diagnosis of pulmonary disease in patients infected with the human immunodeficiency virus. *J. Thorac. Imaging* 13, 247–260. doi: 10.1097/00005382-199810000-00005

Heijink, I. H., Brandenburg, S. M., Postma, D. S., and van Oosterhout, A. J. (2012). Cigarette smoke impairs airway epithelial barrier function and cell-cell contact recovery. *Eur. Respir. J.* 39, 419–428. doi: 10.1183/09031936.00193810

Herbst, R. S., Morgensztern, D., and Boshoff, C. (2018). The biology and management of non-small cell lung cancer. *Nature* 553, 446–454. doi: 10.1038/nature25183

Huang, Y. J., Nelson, C. E., Brodie, E. L., DeSantis, T. Z., Baek, M. S., Liu, J. N., et al. (2011). Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma. *J. Allergy Clin. Immun.* 127, 372.e3–381.e3. doi: 10.1016/j.jaci.2010.10.048

Hwang, K. E., Hwang, Y. R., Seol, C. H., Chul-Park, P. S. H., Yoon, K. H., Park, D. S., et al. (2013). Clostridium difficile infection in lung cancer patients. *Jpn. J. Infect. Dis.* 66, 379–382. doi: 10.7883/yoken.66.379

Jin, J., Gan, Y. C., Liu, H. Y., Wang, Z. R., Yuan, J. Y., Deng, T. B., et al. (2019). Diminishing microbiome richness and distinction in the lower respiratory tract of lung cancer patients: a multiple comparative study design with independent validation. *Lung Cancer* 136, 129–135. doi: 10.1016/j.lungcan.2019.08.022

Kim, H. J., Kim, Y. S., Kim, K. H., Choi, J. P., Kim, Y. K., Yun, S., et al. (2017). The microbiome of the lung and its extracellular vesicles in nonsmokers, healthy smokers and COPD patients. *Exp. Mol. Med.* 49:e316. doi: 10.1038/emm.2017.7

Kwa, M., Plottel, C. S., Blaser, M. J., and Adams, S. (2016). The intestinal microbiome and estrogen receptor-positive female breast cancer. *J. Natl. Cancer Inst.* 108:djw029. doi: 10.1093/jnci/djw029

Leng, Q. X., Holden, V. K., Deepak, J., Todd, N. W., and Jiang, F. (2021). Microbiota biomarkers for lung cancer. *Diagnostics* 11:407. doi: 10.3390/diagnostics11030407

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. *Front. Cell Develop. Biol.* 9:619330. doi: 10.3389/fcell.2021.619330

Lu, H., Gao, N. L., Tong, F., Wang, J. J., Li, H. H., Zhang, R. G., et al. (2021). Alterations of the human lung and gut microbiomes in non-small cell lung carcinomas and distant metastasis. *Microbiol. Spectr.* 9:e0080221. doi: 10.1128/Spectrum.00802-21

Lu, Y., Zheng, Y. S., Wang, Y. H., Gu, D. M., Zhang, J., Liu, F., et al. (2021). FlowCell-enriched circulating tumor cells as a predictor of lung cancer metastasis. *Hum. Cell* 34, 945–951. doi: 10.1007/s13577-021-00500-8

Molina, J. R., Yang, P. G., Cassivi, S. D., Schild, S. E., and Adjei, A. A. (2008). Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin. Proc.* 83, 584–594. doi: 10.1016/S0025-6196(11)60735-0

Nasim, F., Sabath, B. F., and Eapen, G. A. (2019). Lung cancer. *Med. Clin. N. Am.* 103:463. doi: 10.1016/j.mcna.2018.12.006

Newman, M. E. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Pauly, J. L., and Paszkiewicz, G. (2011). Cigarette smoke, bacteria, mold, microbial toxins, and chronic lung inflammation. *J. Oncol.* 2011:819129. doi: 10.1155/2011/819129

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

Rosell, R., and Karachaliou, N. (2015). Relationship between gene mutation and lung cancer metastasis. *Cancer Metast Rev.* 34, 243–248. doi: 10.1007/s10555-015-9557-1

Salazar, Y., Zheng, X., Brunn, D., Raifer, H., Picard, F., Zhang, Y. J., et al. (2020). Microenvironmental Th9 and Th17 lymphocytes induce metastatic spreading in lung cancer. *J. Clin. Invest.* 130, 3560–3575. doi: 10.1172/JCI124037

Sapkota, A. R., Berger, S., and Vogel, T. M. (2010). Human pathogens abundant in the bacterial metagenome of cigarettes. *Environ. Health Perspect.* 118, 351–356. doi: 10.1289/ehp.0901201

Seyfried, T. N., and Huysentruyt, L. C. (2013). On the origin of cancer metastasis. *Crit. Rev. Oncog.* 18, 43–73. doi: 10.1615/CritRevOncog.v18.i1-2.40

Shi, J. J., Yang, Y. H., Xie, H., Wang, X. F., Wu, J., Long, J. R., et al. (2021). Association of oral microbiota with lung cancer risk in a low-income population in the southeastern USA. *Cancer Cause Control* 32, 1423–1432. doi: 10.1007/s10552-021-01490-6

Song, Z., Chen, X., Shi, Y., Huang, R., Wang, W., Zhu, K., et al. (2020). Evaluating the potential of T cell receptor repertoires in predicting the prognosis of resectable non-small cell lung cancers. *Mol. Ther. Methods Clin. Dev.* 18, 73–83. doi: 10.1016/j.omtm.2020.05.020

Tomita, Y., Ikeda, T., Sakata, S., Saruwatari, K., Sato, R., Iyama, S., et al. (2020). Association of Probiotic Clostridium butyricum therapy with survival and response to immune checkpoint blockade in patients with lung cancer. *Cancer Immunol. Res.* 8, 1236–1242. doi: 10.1158/2326-6066.CIR-20-0051

Tyagi, A., Sharma, S., Wu, K., Wu, S. Y., Xing, F., Liu, Y., et al. (2021). Nicotine promotes breast cancer metastasis by stimulating N2 neutrophils and generating pre-metastatic niche in lung. *Nat. Commun.* 12:474. doi: 10.1038/s41467-020-20733-9

Woerner, J., Huang, Y. D., Hutter, S., Gurnari, C., Sanchez, J. M. H., Wang, J., et al. (2022). Circulating microbial content in myeloid malignancy patients is associated with disease subtypes and patient outcomes. *Nat. Commun.* 13:1038. doi: 10.1038/s41467-022-28678-x

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46

Wu, Z., Wang, G., Pan, D., Guo, Y., Zeng, X., Sun, Y., et al. (2016). Inflammation-related pro-apoptotic activity of exopolysaccharides isolated from Lactococcus lactis subsp lactis. *Benef Microbes* 7, 761–768. doi: 10.3920/BM2015.0192

Yang, J., Hui, Y., Zhang, Y., Zhang, M., Ji, B., Tian, G., et al. (2021). Application of circulating tumor DNA as a biomarker for non-small cell lung cancer. *Front. Oncol.* 11:725938. doi: 10.3389/fonc.2021.725938

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., et al. (2022). Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput. Struct. Biotechnol. J.* 20, 333–342. doi: 10.1016/j.csbj.2021.12.028

Yang, M., Yang, H., Ji, L., Hu, X., Tian, G., Wang, B., et al. (2022). A multi-omics machine learning framework in predicting the survival of colorectal cancer patients. *Comput. Biol. Med.* 146:105516. doi: 10.1016/j.compbiomed.2022.105516

Ye, Z., Zhang, Y., Liang, Y., Lang, J., Zhang, X., Zang, G., et al. (2022). Cervical cancer metastasis and recurrence risk prediction based on deep convolutional neural network. *Curr. Bioinforma.* 17, 164–173. doi: 10.2174/1574893616666210708143556

Zhang, W. Q., Zhao, S. K., Luo, J. W., Dong, X. P., Hao, Y. T., Li, H., et al. (2018). Alterations of fecal bacterial communities in patients with lung cancer. *Am. J. Transl. Res.* 10, 3171–3185.

Zhao, F., An, R., Wang, L. Q., Shan, J. K., and Wang, X. J. (2021). Specific gut microbiome and serum Metabolome changes in lung cancer patients. *Front. Cell Infect. Microbiol.* 11:725284. doi: 10.3389/fcimb.2021.725284

Zhao, Z., Fei, K. L., Bai, H., Wang, Z. J., Duan, J. C., and Wang, J. (2021). Metagenome association study of the gut microbiome revealed biomarkers linked to chemotherapy outcomes in locally advanced and advanced lung cancer. *Thorac. Cancer* 12, 66–78. doi: 10.1111/1759-7714.13711

Zhao, Y., Liu, Y. X., Li, S., Peng, Z. Y., Liu, X. T., Chen, J., et al. (2021). Role of lung and gut microbiota on lung cancer pathogenesis. *J. Cancer Res. Clin.* 147, 2177–2186. doi: 10.1007/s00432-021-03644-0

Zhao, J., Zhang, X. L., Gong, C. J., and Zhang, J. L. (2017). Targeted therapy with apatinib in a patient with relapsed small cell lung cancer a case report and literature review. *Medicine* 96:e9259. doi: 10.1097/MD.0000000000009259

Zheng, Y. J., Fang, Z. Y., Xue, Y., Zhang, J., Zhu, J. J., Gao, R. Y., et al. (2020). Specific gut microbiome signature predicts the early-stage lung cancer. *Gut. Microbes* 11, 1030–1042. doi: 10.1080/19490976.2020.1737487

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 23 September 2022
DOI 10.3389/fmicb.2022.1007295

OPEN ACCESS

EDITED BY
Fei Ma,
Chinese Academy of Medical Sciences and Peking Union Medical College, China

REVIEWED BY
Martina Barchitta,
University of Catania,
Italy
Jing Yang,
ShanghaiTech University,
China
Meijing Li,
Shanghai Maritime University,
China

*CORRESPONDENCE
Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cai_yud@126.com

†These authors have contributed equally to this work

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

RECEIVED 30 July 2022
ACCEPTED 01 September 2022
PUBLISHED 23 September 2022

CITATION
Liu Z, Meng M, Ding S, Zhou X, Feng K, Huang T and Cai Y-D (2022) Identification of methylation signatures and rules for predicting the severity of SARS-CoV-2 infection with machine learning methods. *Front. Microbiol.* 13:1007295. doi: 10.3389/fmicb.2022.1007295

Identification of methylation signatures and rules for predicting the severity of SARS-CoV-2 infection with machine learning methods

Zhiyang Liu[1†], Mei Meng[2†], ShiJian Ding[3†], XiaoChao Zhou[2], KaiYan Feng[4], Tao Huang[5,6*] and Yu-Dong Cai[3*]

[1]School of Life Sciences, Changchun Sci-Tech University, Changchun, China, [2]State Key Laboratory of Oncogenes and Related Genes, Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [3]School of Life Sciences, Shanghai University, Shanghai, China, [4]Department of Computer Science, Guangdong AIB Polytechnic College, Guangzhou, China, [5]Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, [6]CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

Patients infected with SARS-CoV-2 at various severities have different clinical manifestations and treatments. Mild or moderate patients usually recover with conventional medical treatment, but severe patients require prompt professional treatment. Thus, stratifying infected patients for targeted treatment is meaningful. A computational workflow was designed in this study to identify key blood methylation features and rules that can distinguish the severity of SARS-CoV-2 infection. First, the methylation features in the expression profile were deeply analyzed by a Monte Carlo feature selection method. A feature list was generated. Next, this ranked feature list was fed into the incremental feature selection method to determine the optimal features for different classification algorithms, thereby further building optimal classifiers. These selected key features were analyzed by functional enrichment to detect their biofunctional information. Furthermore, a set of rules were set up by a white-box algorithm, decision tree, to uncover different methylation patterns on various severity of SARS-CoV-2 infection. Some genes (PARP9, MX1, IRF7), corresponding to essential methylation sites, and rules were validated by published academic literature. Overall, this study contributes to revealing potential expression features and provides a reference for patient stratification. The physicians can prioritize and allocate health and medical resources for COVID-19 patients based on their predicted severe clinical outcomes.

KEYWORDS

SARS-CoV-2, severity, methylation, machine learning, classification rule

Introduction

Since its outbreak in late 2019, COVID-19, which is caused by SARS-CoV-2, has resulted in more than 5 million deaths. SARS-CoV-2 binds to the spike (S) protein primarily through its functional receptor ACE2, an 805-amino acid type I transmembrane protein, allowing the virus to attach to the host cell membrane. This process results in alteration of the extracellular domain of ACE2 and internalization of the transmembrane domain, leading to further fusion of the viral particle with the host cell (Schulte-Schrepping et al., 2020). The SARS-CoV-2 infection progresses to different severities, including discharge from the emergency department, hospitalization, transfer to the ICU, and death, due to a variety of factors, such as age, gender, and other underlying diseases (Konigsberg et al., 2021). Therefore, rapidly determining the severity of the patient and taking corresponding treatment measures for timely and effective diagnosis and treatment is crucial.

Viruses can escape the immune clearance of the body through a variety of ways, among which epigenetic modification is an important way for respiratory viruses to resist the immune response of the body. DNA methylation, mainly of CpG islands, is a crucial reversible epigenetic regulation process (Fan et al., 2017; Benhamida et al., 2020). The regulation of the activity of a variety of DNA/RNA viruses, including HIV, HBV, and HPV, is related to changes in DNA methylation (Castro De Moura et al., 2021). Studies have shown that MERS-CoV and H5N1 influenza virus infection leads to methylation of antigen-presenting gene promoters in infected cells, which eliminates the expression of related genes, thereby antagonizing antigen presentation, resulting in impaired T-lymphocyte function during acute infection and aggravating the degree of virus infection in the body (Hatta et al., 2010; Menachery et al., 2018). Similarly, as a respiratory virus, SARS-CoV infection also results in DNA methylation in host cells (Menachery et al., 2018). Among them, the hypermethylation of the IFN pathway and inflammation-related genes is an important feature of severe COVID-19 (Corley et al., 2021). The study of ACE2 revealed that the DNA in the CpG island of the ACE2 promoter in lung epithelial cells is hypomethylated, indicating its high expression in the lung. Moreover, its methylation status was significantly correlated with age and gender, explaining the effect of age and gender on the severity of COVID-19 (Kianmehr et al., 2021). In addition, ACE2 mRNA is highly expressed in various diseases, especially cancer, which may be an important reason for the severe COVID-19 caused by the underlying disease in SARS-CoV infection (Sen et al., 2021). RNA modification, namely N 6-methylation of adenosine (m6A), also plays an important role in evading the innate immune recognition of exogenous RNA of the host, affecting virus structure and replication (Eberle et al., 2021). The study of human metapneumovirus showed that m6A-binding protein can label viral RNA as the RNA of the host after binding to m6A, thereby evading the antiviral response of the host (Durbin et al., 2016; Chen et al., 2019). In addition, some studies have found that the N region of the SARS-CoV-2 virus genome is rich in m6A modification and is regulated by the host cell methyltransferase METTL3. The reduced expression level of METTL3 will lead to a decrease in the level of SARS-CoV-2 m6A and correspondingly increased expression of inflammatory genes (Li S. et al., 2021). This process is more pronounced in severely infected patients than that in moderately infected patients. These findings suggest the possibility of using methylation to characterize disease states, and numerous studies have demonstrated the feasibility of this approach.

This study conducted a computational investigation on the blood methylation profile on severity of SARS-CoV-2 infection. Several advanced machine learning methods were adopted. First, the profile was analyzed by the Monte Carlo feature selection (MCFS) method (Dramiński et al., 2007) to analyze the importance of methylation features. One feature list was produced, which was further analyzed by incremental feature selection (IFS) (Liu and Setiono, 1998) method. Four classification algorithms were adopted in the IFS method to discover their optimal features, and build the optimal classifiers and classification rules. For the essential methylation features, their corresponding genes were picked up for gene ontology (GO) and KEGG enrichment analysis. Some results, including essential methylation sites, classification rules, and enrichment analysis results, were extensively discussed and can be validated by existing literature. The results reported in this study are helpful for the stratification of clinical patients and provide an effective reference for clinical diagnosis and treatment.

Materials and methods

Methylation dataset

The blood DNA methylation dataset investigated in this study was obtained from the Gene Expression Omnibus (GEO) database with the accession ID of GSE167202 (Konigsberg et al., 2021). This dataset comprised 164 SARS-CoV-2-positive samples, 296 SARS-CoV-2-negative infection samples, and 65 other infection samples. In addition, the positive samples were classified into four categories based on severity score. The severity score is determined primarily by discharge from emergency, admission to inpatient care, progression to the ICU, and death. The above four categories, negative infection, and other infections were termed as six classes in this study. The methylation dataset was deeply analyzed by modeling a classification problem on the dataset. The sample size of each class is listed in Table 1. Each sample was represented by 655,010 methylation features. This dataset would be analyzed in the following steps.

Monte Carlo feature selection

A large number of methylation features were used to represent each sample. However, only a few of them were highly related to the severity of SARS-CoV-2 infection. It was necessary to reveal

TABLE 1 Sample size of each class for the methylation profile.

Class name	Sample size
Negative infection	296
Other infection	65
Discharged from emergency department	34
Admitted to inpatient care	84
Progressed to ICU	35
Death	11

essential methylation features with advanced computer techniques. Here, MCFS method was employed (Dramiński et al., 2007).

MCFS is a tree-based feature selection method that is widely used in methylation profiling analysis as it is deemed to be good at dealing with datasets containing small number of samples and huge number of features. It randomly constructs several decision trees (DTs) from the original training dataset and uses these DTs to evaluate the importance of features. More specifically, *s* subsets with *m* features are randomly selected from the original training dataset. *t* trees for each subset are then constructed based on samples randomly sampled from the original dataset. The performance of each tree is evaluated on test samples that are not selected as training samples. Overall, $s \times t$ DTs are built in this process. The overall position of a feature on the tree node partition is used to estimate a measurement, called relative importance (RI). A high RI score of a feature indicates the importance of a feature. The RI score is defined as follows:

$$RI = \sum_{\tau=1}^{st} (wACC)^u \sum_{n_g(\tau)} IG\left(n_g(\tau)\right) \left(\frac{no.in\, n_g(\tau)}{no.in\, \tau} \right)^v, \quad (1)$$

where $IG\left(n_g(\tau)\right)$ indicates the information gain of tree node $n_g(\tau)$, $no.in\, n_g(\tau)$ and $no.in\, \tau$ represent the number of samples in node $n_g(\tau)$ and tree τ, respectively, and $wAcc$ indicates the weighted accuracy of the DT τ. In addition, *u* and *v* are the two parameters for RI calculation.

The MCFS program developed by Dramiński et al. was applied in this study, which can be downloaded at https://home.ipipan.waw.pl/m.draminski/mcfs.html, to rank the methylation features. Default parameters were used, where *u* and *v* were set to 1. By applying the MCFS program on the methylation dataset, a ranked feature list was obtained.

Incremental feature selection

Based on the MCFS method, the methylation features were ranked in a list. However, the threshold was difficult to determine, that is, which features were selected for further analysis. In view of this, we further employed the IFS method (Liu and Setiono, 1998).

The IFS method is always used to determine the optimal number of features in a ranked feature list combined with one supervised classification algorithm, such as random forest (RF). More specifically, IFS first generates a series of feature subsets based on a step size. For example, the first and second subsets, respectively, comprise the top 5 and 10 features when the step size is five. Next, on each feature subset, the samples represented by features in such subset are learned by the given classification algorithm, thereby building a classifier. Its performance is evaluated by the 10-fold cross-validation (Kohavi, 1995). After the evaluation metrics of all classifiers are obtained, the classifier with the highest performance is easy to find. Such classifier is called the optimal classifier. The corresponding feature subset is picked up and features in this subset are termed as the optimal features for the used classification algorithm.

Synthetic minority oversampling technique

As shown in Table 1, the sample sizes under six classes were quite different. The largest class contained samples about 17 times as many as those in the smallest class. This may lead to biased performance of the established classifiers. Therefore, the synthetic minority oversampling technique (SMOTE) algorithm (Chawla et al., 2002; Ding et al., 2022; Zhou et al., 2022), an oversampling method, was applied to solve the problem. The core idea of SMOTE is to generate new samples to each minor class for enlarging its size. For each minor class, SMOTE randomly selects one sample, say *x*, from this class and finds its k-nearest neighbor samples in the same class. One sample, say *y*, is randomly selected from these k-nearest neighbor samples. One new sample is synthesized by the linear combination of *x* and *y*. As such new sample is highly related to *x* and *y*, it belongs to the same class with a high probability. Thus, it is put into the minor class. Such procedures execute several times until the size of the minor class is equal to that of the major class.

In this study, the SMOTE program from the imblearn package[1] was used to process the methylation data for solving the imbalanced problem when constructing classifiers in the IFS method.

Classification algorithm

As the execution of IFS method needs one classification algorithm, four classic classification algorithms were attempted in this study to fully assess each constructed feature subset. They were k-nearest neighbor (kNN; Cover and Hart, 1967), RF (Breiman, 2001), support vector machine (SVM; Cortes and Vapnik, 1995), and DT (Safavian and Landgrebe, 1991). Their brief descriptions were as follows.

1 https://imbalanced-learn.org/stable/

k-Nearest neighbor

k-Nearest neighbor is one of the most classic classification algorithms. It determines the class of a sample based on measuring the distance between samples. Given a training dataset, for a new test sample, the *k* neighbors closest to such sample are found in the training dataset. By counting the classes of its *k* neighbors, the class of the test sample can be determined. Generally, the class that occurs most for its *k* neighbors is assigned to the test sample.

Random forest

RF is an ensemble algorithm that contains several DTs. Each DT is constructed by randomly selecting samples from the original dataset and features from all features. RF provides the final prediction result using the voting strategy on predictions yielded by DTs. RF is generally much more powerful than its component DT, and few parameters are involved in this algorithm.

Support vector machine,

SVM is an excellent classification algorithm in machine learning. The original SVM can only tackle binary classification. It separates samples into two classes by constructing a hyperplane, which can separate samples into two classes with the maximum interval. However, such hyperplane does not always exist or is not easy to find out. SVM maps samples into a high-dimensional space using one kernel function. In the new space, the hyperplane can be easily constructed. For a test sample, it is also mapped into the high-dimensional space. Its class is determined by the side it lies. The "one-versus-rest" or "one-versus-one" can be adopted to generalize the original SVM so that it can tackle multi-class classification problems.

Decision tree

Different from the above algorithms, which are deemed as black-box algorithms, DT can make the classification procedures interpretable. By learning the distributions of samples under each feature, a tree-like structure is built by DT. In this structure, each internal node indicates a decision on an attribute, outputting a judgment result, and each leaf node denotes a classification outcome. Besides, DT can also be represented by a set of rules. Each rule is obtained by a path from the root node to one leaf node in the tree. In terms of these rules, the class of a test sample can be determined. This operation also makes the classification procedures completely open, giving more chances for us to understand the procedures. In this case, more meaningful and hidden information in the dataset can be mined.

Above classification algorithms have wide applications in many fields. They are always important candidates for building classifiers in tackling various biological and medical problems (Zhou et al., 2020a,b, 2022; Chen et al., 2021, 2022; Onesime et al., 2021; Zhang Y. et al., 2021; Ding et al., 2022; Li et al., 2022; Ran et al., 2022; Tang and Chen, 2022; Wang and Chen, 2022; Wu and Chen, 2022; Yang and Chen, 2022). These algorithms were implemented in this study through the scikit-learn (Pedregosa et al., 2011) program in Python and run with default parameters.

Performance measurement

The prediction performance of each classifier was mainly evaluated with the weighted F1. Its calculation is based on the F1 score on each class. The F1 score for one class can be computed by

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{3}$$

$$\text{F1 score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{4}$$

where *TP*, *FP*, and *FN* represent true-positive, false-positive, and false-negative for the class, respectively. The weighted F1 is defined as the weighed mean of F1 scores on all classes. The direct mean of F1 scores on all classes was also provided, which was called macro F1.

To fully evaluate the performance of classifiers in the IFS method, we also adopted overall accuracy (ACC) and Matthews correlation coefficients (MCC; Matthews, 1975; Gorodkin, 2004). ACC is defined as the ratio of correctly predicted samples and all samples, which is the most accepted measurement. However, it is not perfect when the class sizes are of great differences. In view of this, MCC was proposed, which is deemed as a balanced measurement. For computing MCC, two binary matrices *X* and *Y* should be constructed first, where *X* stands for the true class of each sample and *Y* represents the predicted class of each sample. Then, MCC can be computed by

$$\text{MCC} = \frac{\text{cov}(\text{X},\text{Y})}{\sqrt{\text{cov}(\text{X},\text{X})\text{cov}(\text{Y},\text{Y})}} \tag{5}$$

Enrichment analysis

According to the IFS results, the essential methylation features for severity of SARS-CoV-2 infection can be obtained. Their corresponding genes can be picked up for further analysis. GO and KEGG enrichment analysis is a common method for uncovering biological meanings behind a set of genes. Here, it was applied to discover the biofunctional information of the genes corresponding to essential methylation features. Such analysis was performed by using the R package clusterProfiler 4.0 (Wu et al., 2021) with a threshold of 0.05.

Results

This study conducted a deep computational investigation on the blood methylation profile with six severity types from the GEO database. The entire procedures are illustrated in Figure 1.

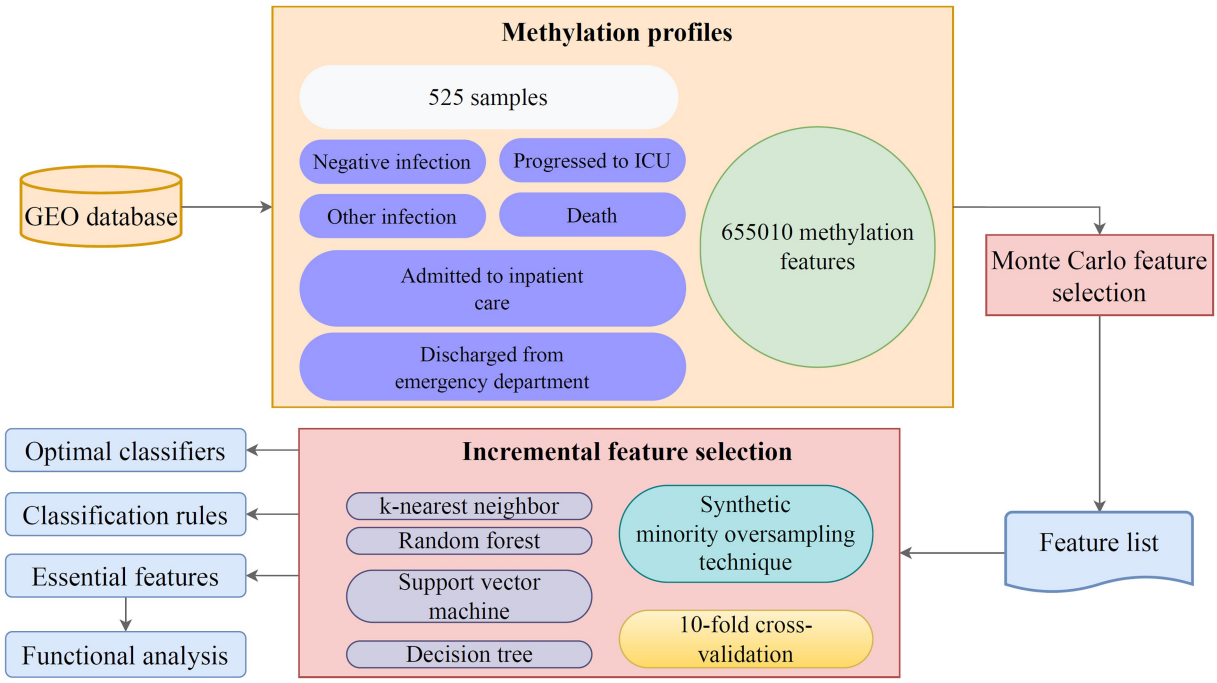


FIGURE 1
Workflow of this study. First, the Monte Carlo feature selection (MCFS) method was used to rank methylation signatures based on their importance, and a ranked feature list was generated. This list was then fed into the incremental feature selection (IFS) method with different classification algorithms to determine the optimal features for each classification algorithm. Optimal classifiers were set up. Classification rules generated by the optimal decision tree (DT) classifier were used to analyze the methylation expression pattern. The genes corresponding to essential methylation sites were subjected to functional enrichment analysis.

The MCFS method was first used to rank methylation features based on their importance, and a ranked feature list was generated. This list was then fed into the IFS method with different classification algorithms to determine the optimal features for each classification algorithm and construct optimal classifiers. Classification rules generated by the optimal DT classifier were used to analyze the expression pattern of key methylation features.

Results of methylation feature ranking by the MCFS method

Initially, the MCFS method was used to rank 655,010 methylation features contained in blood the methylation profile. Each feature was assigned a RI score. A ranked feature list in descending order based on RI scores was generated. As some features were assigned RI scores of 0, they were removed. Thus, the final list contained 654,081 features with RI scores larger than 0, which is provided in Supplementary Table S1. The top 10 features alone with their RI score are plotted in Figure 2.

Identification of the optimal number of methylation features with IFS

The IFS method was applied to determine the optimal features in the ranked feature list for each classification algorithm. To save time, we only considered top 2000 features in the list because of the huge number of features. The step size is set to five in the IFS method, thereby generating 400 feature subsets. The sample dataset comprising these feature subsets was learned by each of four classification algorithms, namely DT, kNN, RF, and SVM. Lots of classifiers were built, which were evaluated by 10-fold cross-validation. The evaluation metrics for each classifier are provided in Supplementary Table S2. To clearly display the performance of classifiers under different feature subsets, an IFS curve is plotted for each classification algorithm, which is provided in Figure 3. For SVM, the highest weighted F1 was 0.921 when top 1,025 features were adopted. These features constituted the optimal features for SVM and an optimal SVM classifier was built based on these features. As for kNN and RF, their highest weighted F1 values were 0.790 and 0.895, respectively. Their optimal features were top 10 and 35 features in the list. Furthermore, the optimal kNN and RF classifiers were set up with their optimal features, respectively. For DT, its highest weighted F1 was 0.780, which was obtained by using top 590 features. Such features comprised the optimal features for DT and the optimal DT classifier was built using these optimal features. According to the weighted F1 values of above optimal classifiers, the optimal SVM classifier was best, followed by the optimal RF and kNN classifiers, whereas the optimal DT classifier provided the lowest performance. Table 2 further lists other overall measurements for four optimal classifiers. It can be observed that on each measurement, the optimal SVM classifier always provided the

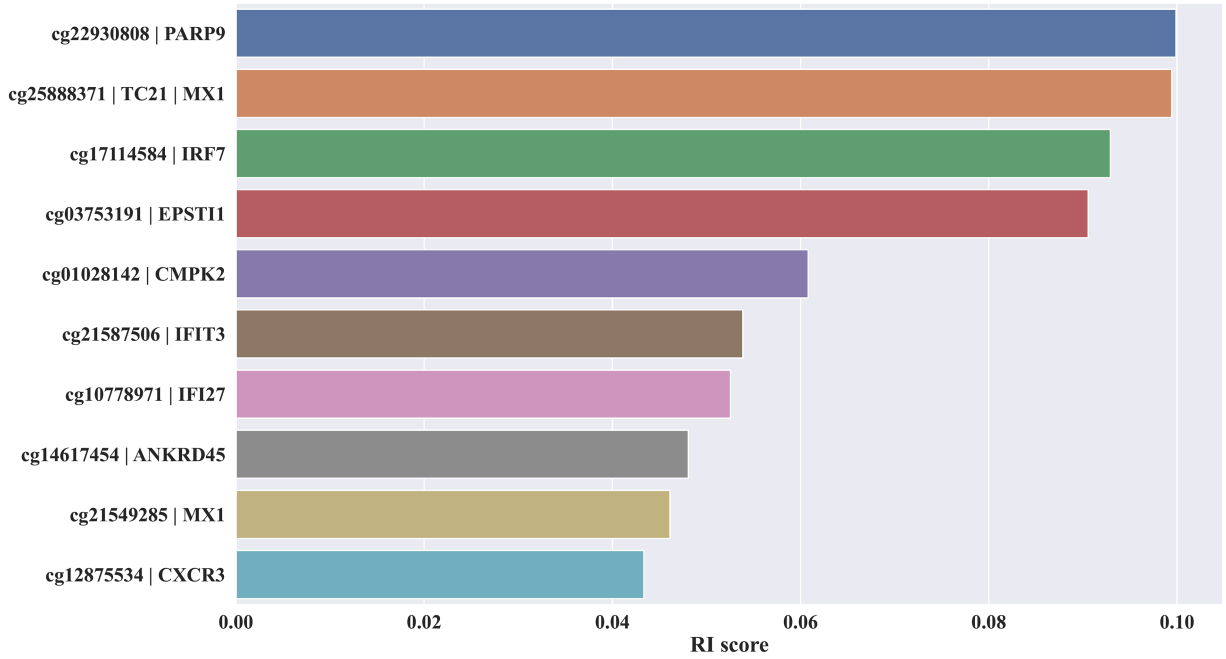


FIGURE 2
Bar chart to show top 10 key methylation features and their relative importance scores.

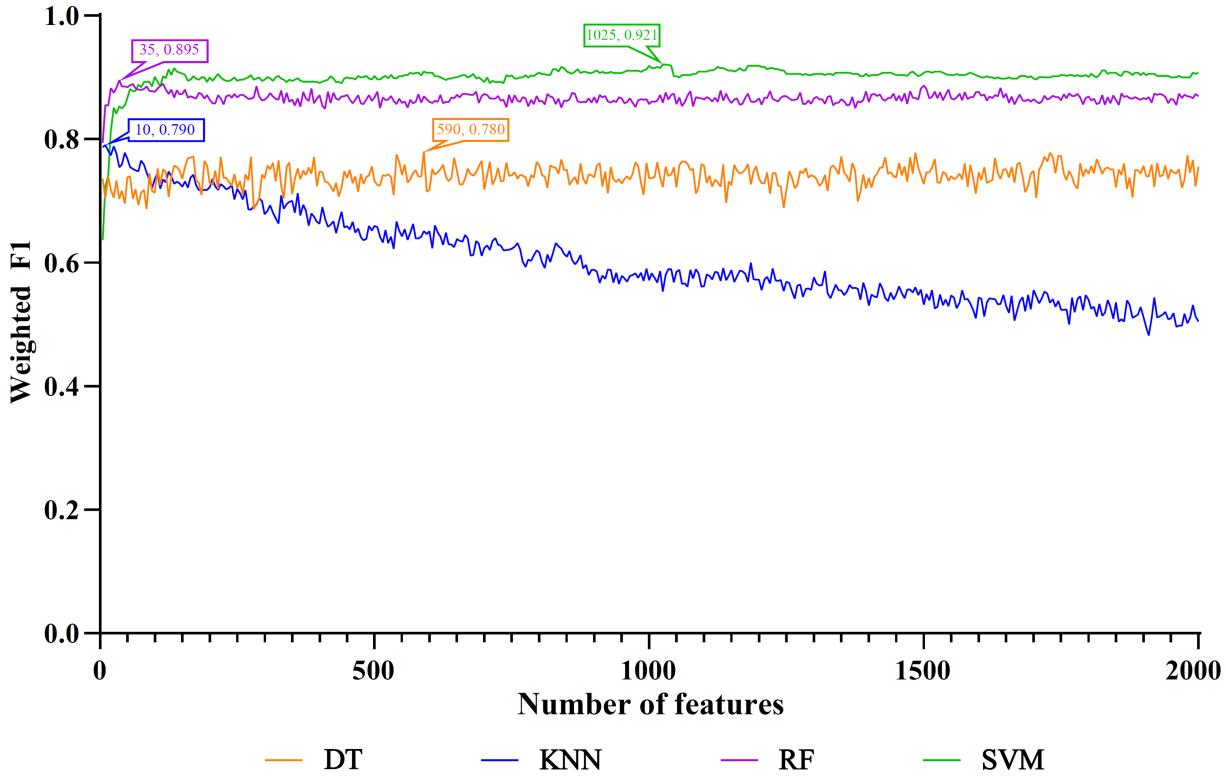


FIGURE 3
IFS curves to show the performance of different classification algorithms under different feature subsets. The highest weighted F1 for each classification algorithm was marked on the corresponding IFS curve. The SVM yielded the highest weighted F1 of 0.921 when top 1,025 features were used.

highest performance, and the optimal RF classifier yielded slightly lower performance than the optimal SVM classifier. The performance of the other two optimal classifiers was much lower. The optimal DT classifier was a little inferior to the optimal kNN classifier. As for the performance of the above four optimal classifiers on six classes, it is illustrated in Figure 4. Clearly, the optimal SVM classifier generated the highest performance on all classes. On most classes, the optimal RF classifier occupied the second places. The optimal kNN and DT classifiers gave an almost equal performance. These results conformed to the overall performance of four optimal classifiers mentioned above.

With the above arguments, the optimal SVM classifier was best. It can be an efficient tool to determine the severity of SARS-CoV-2 infection. The optimal RF classifier was inferior to the optimal SVM classifier. However, its efficiency was much higher than that of the optimal SVM classifier as much less features were used. This classifier can be used to conduct large-scale tests.

TABLE 2 Overall performance of the optimal classifiers.

Classification algorithm	Number of features	ACC	MCC	Macro F1	Weighted F1
k-nearest neighbor	10	0.784	0.730	0.793	0.790
Random forest	35	0.893	0.842	0.873	0.895
Support vector machine	1,025	0.920	0.881	0.926	0.921
Decision tree	590	0.771	0.686	0.749	0.780

Classification rules generated by the optimal DT classifier

Although the optimal DT classifier provided lower performance than the other three optimal classifiers, it can provide much more explicable information than other classifiers. As the optimal DT classifier adopted top 590 features in the list, a DT classifier trained with all samples comprising these features was built. Classification rules were extracted from the tree, resulting in 77 rules. These rules are provided in Supplementary Table S3. The number of rules for each class is displayed in Figure 5. The rules

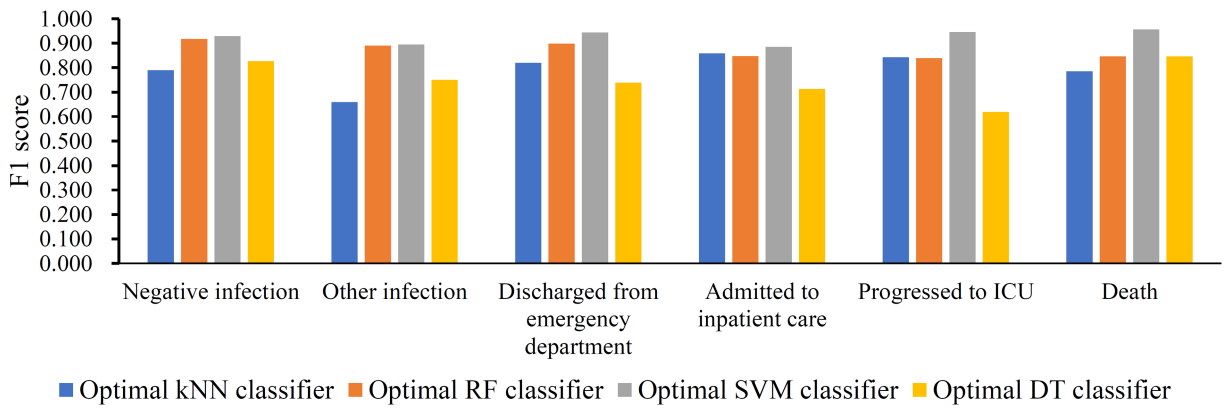


FIGURE 4
Performance of four optimal classifiers on six classes. The optimal SVM classifier produced best performance on all classes.

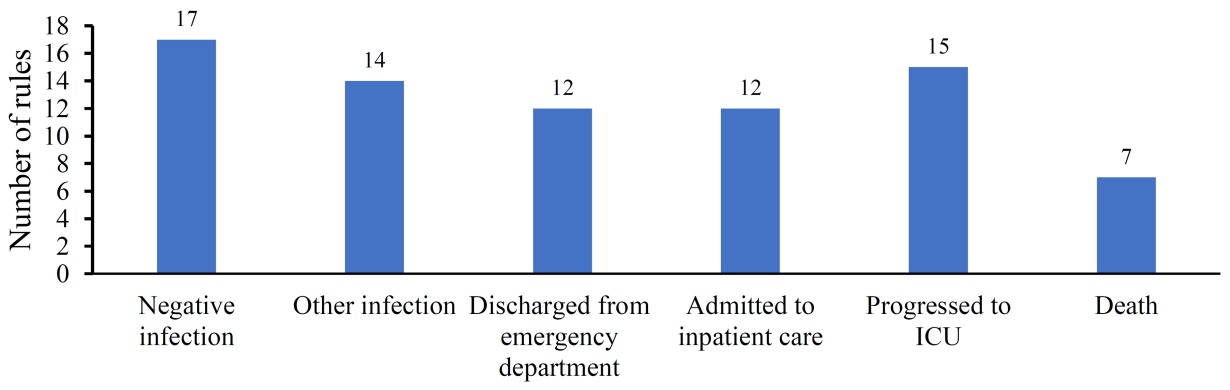


FIGURE 5
Distribution of classification rules on six classes.

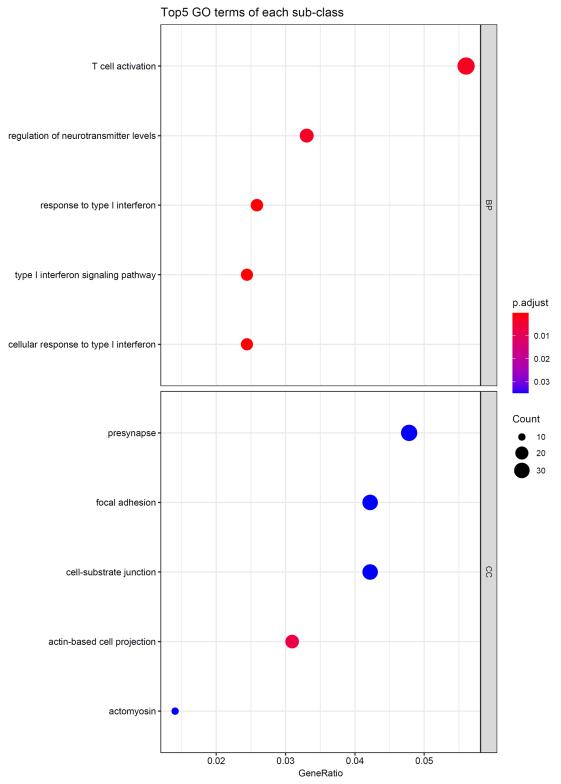


FIGURE 6
Top five GO terms enriched by the genes converted by the top 1,025 methylation features.

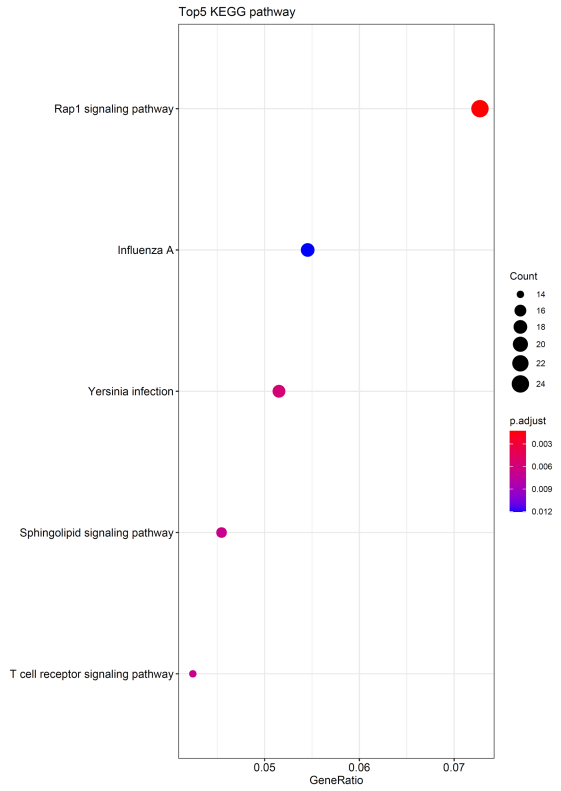


FIGURE 7
Top five KEGG pathways enriched by the genes converted by the top 1,025 methylation features.

for "negative infection" were most, whereas those for "Death" were least. In section "Analysis of rules for different classes", some rules would be discussed.

Results of functional enrichment analysis

As the optimal SVM classifier gave the best performance. This meant that features used in this classifier, that is the optimal features for SVM, were essential for determining the severity of SARS-CoV-2 infection. The corresponding genes of these features were picked up and the GO and KEGG enrichment analyses were performed on these genes, uncovering the biological meaning behind these genes. The detailed results are listed in Supplementary Table S4. Figures 6, 7 reveal that these genes were mainly enriched in biological processes, such as T-cell activation, regulation of neurotransmitter levels, and type I interferon signaling and KEGG pathways (e.g., Rap1 signaling pathway, Yersinia infection, and T-cell receptor signaling pathway). The role of these biological functions in SARS-CoV-2 infection will be verified in the section "Functional analysis based on GO and KEGG pathway".

Discussion

Most studies only distinguish COVID-19-positive and negative samples. In this study, based on blood methylation biomarkers, we can not only classify COVID-19 from negative controls and other infections, but also accurately predict the clinical outcome of COVID-positive patients in detail. In practice, the physicians can prioritize and allocate health and medical resources for COVID-19 patients based on their predicted severe clinical outcomes. For the least severe patient, they can be discharged from hospital and avoid medical resource overstretch. For the second least severe patient, they can be hospitalized but without intensive health care. For the severe patient, intensive health care should be prepared. For the most severe patient who may die from COVID-19, life support system should be prepared.

A variety of machine learning methods were used to investigate the methylation profile on severity of SARS-CoV-2 infection. Some essential methylation features that can characterize the severity of SARS-CoV-2 infection were identified. Furthermore, a set of rules were also set up, which can not only classify SARS-CoV-2 infection samples, but also depict the methylation patterns for different severity of SARS-CoV-2 infection. These methylation features and rules would then be discussed below.

TABLE 3 Essential methylation sites and their corresponding genes for distinguishing severity of SARS-CoV-2 infection.

Methylation sites	Gene symbol	Description
cg22930808	PARP9	Poly (ADP-Ribose) Polymerase Family Member 9
cg25888371	MX1	MX Dynamin Like GTPase 1
cg17114584	IRF7	Interferon Regulatory Factor 7

Analysis of essential features

Key methylation signatures that can be used to distinguish severity of SARS-CoV-2 infection were obtained by using a set of machine learning methods. The genes corresponding to the top-ranked methylation signatures, listed in Table 3, were analyzed to demonstrate the reliability of the results.

As a type I IFN regulatory gene, high expressions of polyadenosine diphosphate ribose polymerase 9 (PARP9, cg22930808) accompanied by hypomethylation at relevant sites can enhance IFN signaling (Zhu et al., 2019), thereby playing a role in solid tumors, macrophage regulation, and antiviral immunity (Xing et al., 2021). PARP9 mediates the production of type I interferon after binding to viral RNA by activating the PI3K/AKT3 signaling pathway, thereby protecting against viral infection (Zhang et al., 2015). In addition, PARP9 is involved in the activation of anti-inflammatory M2 macrophages. This condition showed that the SARS-CoV-2 Nsp3 protein is similar to PARP9 and can inhibit PARP9 through molecular mimicry, depleting M2 macrophages, and weakening interferon signaling, which then weakens the ability of the host to resist viral infection (Da Silva et al., 2020; Fehr et al., 2020). The reduction of PARP9 combined with the reduction of NK and CD8+ cells leads to a weak viral response of the host, which may be an important reason for the life-threatening severe infections in patients.

Similar to PARP9, as an important host interferon-stimulated gene in antiviral infection (Anderson et al., 2021), MX1 (cg25888371) is hypomethylated in CpG after viral infection (Luo et al., 2021) and then participates in regulating the defense response of the host to infection. The study found that the expression of MX1 was significantly increased in COVID-19 patients compared with non-COVID-19 patients and increased with the viral load (Bizzotto et al., 2020). In addition, the methylation of CpG in MX1 is associated with the severity of HIV patients using cocaine in HIV infection studies (Shu et al., 2020), suggesting that MX1 methylation levels may be a reliable predictor of COVID-19 severity.

IRF7 (cg17114584), a member of the interferon regulatory factor (IRF) family, can regulate the response of type I IFN to viral infection. Phosphorylation of IRF7 upon pathogen stimulation followed by nuclear translocation induces the expression of IFN-α (Puthia et al., 2016). The methylation level of its promoter region affects the clinical manifestations of diseases (Konigsberg et al., 2021). Studies have shown that the expression level of IRF7 is increased in patients with mild/moderate COVID-19 (Li N. et al., 2021), while those with reduced IRF7 expression due to hypermethylation of the IRF7 promoter gene are likely to develop severe infection after SARS-CoV-2 infection (Liu and Hill, 2020).

Overall, the obtained genes showed differential expression of methylation in different infection groups, suggesting that the methylation status of different genes may be an important feature to distinguish different SARS-CoV2 infection severities.

Analysis of rules for different classes

The decision rules (Supplementary Table S3) revealed the importance of IRF7 (cg17114584) in predicting the clinical outcome of SARS-CoV-2 infection. IRF7 is markedly hypermethylated in patients with poor clinical response (progressed to ICU or death) compared with patients with mild clinical response (discharged from the emergency department or admitted to inpatient care). This finding is consistent with a previous result, in which IRF7 can regulate the response of type I IFN to viral infection and the expression level is negatively correlated with clinical manifestations. Recent studies show that methylation levels of IRF7 correlate with COVID-19 severity (Barturen et al., 2021), which is also consistent with the conclusions in the data source literature (Konigsberg et al., 2021).

The decision rule for distinguishing between other infections and non-COVID-19/COVID-19 infections indicated that FHL1 (cg00012680) was highly methylated in patients with other infections. As a member of the FHL protein family, FHL1 is mainly expressed in the heart and skeletal muscles (Shathasivam et al., 2010). As a tumor suppressor gene, FHL1 is downregulated in a variety of tumors (Wang et al., 2017). Studies have also shown that FHL1 is associated with viral infections (e.g., acting as a host factor to promote chikungunya virus infection; Meertens et al., 2019). Conversely, patients in the "death" cohort had low levels of FHL1 methylation. A study has shown that in COVID-19 patients, FHL1 is associated with the JAK–STAT pathway, which can indirectly activate STATs and induce various inflammatory responses (Bass et al., 2021). Another key criterion in distinguishing patients from other infections is the methylation level of TGFB3 (cg06958766), which is hypomethylated in COVID-19 patients (especially ICU and death patients). Existing studies have demonstrated that TGFB3 is a gene related to immune dysregulation in cardiovascular disease, and its expression is also dysregulated in COVID-19 (Lee et al., 2021). The association of the methylation level of TGFB3 with the clinical outcome of COVID-19 infection has not been revealed, and such level is speculated to be possibly associated with poor clinical response to COVID-19.

The result also indicated that the methylation level of the interferon type I pathway-related gene RSAD2 (cg10549986) for COVID-19 patients was negatively correlated to the severity of COVID-19, and the expression of RSAD2 is reported to have reached the highest level in the early stage compared with the late stage of COVID-19 (Zhang C. et al., 2021). This finding may

be related to the decrease in IFN activity in patients with severe infections. In COVID-19 patients, RSAD2 can enhance antiviral and immunomodulatory functions after viral infection, and patients discharged from the emergency department in the current results had lower levels of RSAD2 methylation, possibly related to high RSAD2 expression levels and enhanced antiviral immunity (Zhu et al., 2020).

Functional analysis based on go and KEGG pathway

T-cell activation is the most significantly enriched pathway. Studies have shown that RNA m6A methylation is crucial for controlling the activation and differentiation of T lymphocytes (Qiu et al., 2021). m6A with T-cell activation function mainly mediates the activation and proliferation of T cells by increasing TGF-β and PI3K-AKT signaling necessary for T-cell differentiation and plays an anti-COVID-19 role (Li et al., 2017). Increased m6A regulator expression in COVID-19 patients results in the high expression of activated CD4 memory T cells (Yao et al., 2021). As a crucial immune cell in SARS-CoV-2 infection, T cells have dual roles in patients with COVID-19. The expression level of T cells is increased in patients with mild infection; among which, $CD8^+$ T cells highly express cytotoxic molecules, such as granzyme A, which play an antiviral immune effect (Liao et al., 2020). Meanwhile, the expression levels of cytotoxic molecules and Tregs in severe patients are reduced (De Biasi et al., 2020; Toor et al., 2021). Studies have shown the presence of a complete memory T-cell response in asymptomatic or mildly infected COVID-19 patients (Sekine et al., 2020) and detected SARS-CoV-2-related T-cell responses in healthy blood samples, which may be due to seasonal coronavirus-induced T-cell responses and may further prevent serious infections (Braun et al., 2020; Mateus et al., 2020).

The current study also observed enrichment of pathways that regulate the level of neurotransmitters, suggesting the role of methylation of neurotransmitter-related genes in immunity to virus infection. Studies have shown that in addition to macrophages, viral infection also activates mast cells to release histamine, arachidonic acid, and other neurotransmitters, and histamine can strongly raise the level of IL-1, which, in turn, increases lung inflammation in SARS-CoV-2 infection (Conti et al., 2020). Furthermore, SARS-CoV2 infection will reduce the synthesis of dopamine and acetylcholine, resulting in the weakened immune function of the body (Blum et al., 2020; Alexandris et al., 2021).

Type I interferon plays a crucial role in antiviral immunity, and studies have shown that hypermethylation of IFN-related genes is a unique methylation signature of severe COVID-19. Moreover, three of the significant enrichment pathways are related to type I interferon response, further confirming the important role of IFN-related methylation in determining the severity of COVID-19 and the reliability of the current study. *In vivo*, IFN can bind to IFN receptors in an autocrine and paracrine manner to activate the JAK/STAT signaling pathway, thus demonstrating antiviral effects (Liu et al., 2012). IFN activity was also lower in patients with severe infection than mild infection patients, and impaired IFN-α production is an important sign of severe infection (Hadjadj et al., 2020), which may be related to the hypermethylation of IFN-related genes and the inhibition of the expression of related genes. In addition, studies have shown that IFN expression is delayed in SARS-CoV-2 infection (Kim et al., 2016). Such a delay leads to high levels of interferon expression in severely infected patients but does not reduce viral load; meanwhile, IFN pretreatment can significantly reduce viral infection levels, suggesting that drugs that can boost IFN production may be an effective option for early treatment of SARS-CoV-2 (Park and Iwasaki, 2020).

The enrichment of cellular components of differentially methylated genes mainly focused on the virus infection process of cells, including cell junction and migration. Synapses mainly mediate information transmission between neurons; they can also transmit large particles and mediate virus particles into the central nervous system, thus reflecting the neuroinvasiveness of coronaviruses (Li et al., 2020). As the main site of cell adhesion, focal adhesions help viral particles enter cells, and its functional integrity is critical to the infection and spread of SARS-CoV-2 (Sulzmaier et al., 2014).

This study also found enrichment of differentially methylated genes in the RAP1 pathway. RAP1 pathway plays an important role in processes, such as cell adhesion, junction, and polarity, and promotes tumor cell invasion and migration (Looi et al., 2020). Pulmonary vascular barrier integrity defection is a fatal factor in severe COVID-19 patients (Yamamoto et al., 2021). Meanwhile, studies have found that RAP1 can enhance endothelial cell–cell junctions mediated by VE-cadherin and regulate vascular permeability (Rho et al., 2017), suggesting that the RAP1 signaling pathway may serve as a potential therapeutic target for COVID-19.

The analysis of key features and related decision rules verified the effectiveness of methylation status in distinguishing different states of SARS-CoV-2 infection, which will provide a reference for studying the stratification of patients and help develop new treatment strategies.

Conclusion

A computational workflow containing several machine learning methods was designed to identify the blood methylation features and their expression rules, which can distinguish the severity of SARS-CoV-2 infection. First, the methylation features in the expression profile were analyzed by the MCFS algorithm, producing a ranked feature list. Next, this list was introduced into the IFS method to generate a series of feature subsets. Different classification algorithms were used to train samples comprising these feature subsets to build classifiers. After evaluating their performance, the optimal features were determined. The

classification rules were extracted by the optimal DT classifier. The essential features were analyzed by functional enrichment to detect their biofunctional information. Some key features and rules are justified by recently published academic literature, which provides a reference for further related research.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE167202.

Author contributions

TH and Y-DC designed the study. SD and KF performed the experiments. ZL, MM, and XZ analyzed the results. ZL, MM, and SD wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDA26040304 and XDB38050200), National Key R&D Program of China (2018YFC0910403), and the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

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.2022.1007295/full#supplementary-material

References

Alexandris, N., Lagoumintzis, G., Chasapis, C. T., Leonidas, D. D., Papadopoulos, G. E., Tzartos, S. J., et al. (2021). Nicotinic cholinergic system and COVID-19: in silico evaluation of nicotinic acetylcholine receptor agonists as potential therapeutic interventions. *Toxicol. Rep.* 8, 73–83. doi: 10.1016/j.toxrep.2020.12.013

Anderson, D., Neri, J., Souza, C. R. M., Valverde, J. G., De Araújo, J. M. G., Nascimento, M., et al. (2021). Zika virus changes methylation of genes involved in immune response and neural development in Brazilian babies born with congenital microcephaly. *J. Infect. Dis.* 223, 435–440. doi: 10.1093/infdis/jiaa383

Barturen, G., Carnero-Montoro, E., Martínez-Bueno, M., Rojo-Rello, S., Sobrino, B., Alcántara-Domínguez, C., et al (2021). Whole-blood DNA methylation analysis reveals respiratory environmental traits involved in COVID-19 severity following SARS-CoV-2 infection. medRXiv.

Bass, A., Liu, Y., and Dakshanamurthy, S. (2021). Single-cell and bulk RNASeq profiling of COVID-19 patients reveal immune and inflammatory mechanisms of infection-induced organ damage. *Viruses* 13:2418. doi: 10.3390/v13122418

Benhamida, J. K., Hechtman, J. F., Nafa, K., Villafania, L., Sadowska, J., Wang, J., et al. (2020). Reliable clinical MLH1 promoter Hypermethylation assessment using a high-throughput genome-wide methylation Array platform. *J. Mol. Diagn.* 22, 368–375. doi: 10.1016/j.jmoldx.2019.11.005

Bizzotto, J., Sanchis, P., Abbate, M., Lage-Vickers, S., Lavignolle, R., Toro, A., et al. (2020). SARS-CoV-2 infection boosts MX1 antiviral effector in COVID-19 patients. *iScience* 23:101585. doi: 10.1016/j.isci.2020.101585

Blum, K., Cadet, J. L., Baron, D., Badgaiyan, R. D., Brewer, R., Modestino, E. J., et al. (2020). Putative COVID-19 induction of reward deficiency syndrome (RDS) and associated behavioral addictions with potential concomitant dopamine depletion: is COVID-19 social distancing a double edged sword? *Subst. Use Misuse* 55, 2438–2442. doi: 10.1080/10826084.2020.1817086

Braun, J., Loyal, L., Frentsch, M., Wendisch, D., Georg, P., Kurth, F., et al. (2020). SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. *Nature* 587, 270–274. doi: 10.1038/s41586-020-2598-9

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Castro De Moura, M., Davalos, V., Planas-Serra, L., Alvarez-Errico, D., Arribas, C., Ruiz, M., et al. (2021). Epigenome-wide association study of COVID-19 severity with respiratory failure. *EBioMedicine* 66:103339. doi: 10.1016/j.ebiom.2021.103339

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. doi: 10.1613/jair.953

Chen, Y. G., Chen, R., Ahmad, S., Verma, R., Kasturi, S. P., Amaya, L., et al. (2019). N6-Methyladenosine modification controls circular RNA immunity. *Mol. Cell* 76, 96–109.e9. doi: 10.1016/j.molcel.2019.07.016

Chen, W., Chen, L., and Dai, Q. (2021). iMPT-FDNPL: identification of membrane protein types with functional domains and a natural language processing approach. *Comput. Math. Methods Med.* 2021, 1–10. doi: 10.1155/2021/7681497

Chen, L., Li, Z., Zhang, S., Zhang, Y.-H., Huang, T., and Cai, Y.-D. (2022). Predicting RNA 5-methylcytosine sites by using essential sequence features and distributions. *Biomed. Res. Int.* 2022, 1–11. doi: 10.1155/2022/4035462

Conti, P., Caraffa, A., Tetè, G., Gallenga, C. E., Ross, R., Kritas, S. K., et al. (2020). Mast cells activated by SARS-CoV-2 release histamine which increases IL-1 levels causing cytokine storm and inflammatory reaction in COVID-19. *J. Biol. Regul. Homeost. Agents* 34, 1629–1632. doi: 10.23812/20-2EDIT

Corley, M. J., Pang, A. P. S., Dody, K., Mudd, P. A., Patterson, B. K., Seethamraju, H., et al. (2021). Genome-wide DNA methylation profiling of peripheral blood reveals an epigenetic signature associated with severe COVID-19. *J. Leukoc. Biol.* 110, 21–26. doi: 10.1002/JLB.5HI0720-466R

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

Da Silva, S. J. R., Alves Da Silva, C. T., Mendes, R. P. G., and Pena, L. (2020). Role of nonstructural proteins in the pathogenesis of SARS-CoV-2. *J. Med. Virol.* 92, 1427–1429. doi: 10.1002/jmv.25858

De Biasi, S., Meschiari, M., Gibellini, L., Bellinazzi, C., Borella, R., Fidanza, L., et al. (2020). Marked T cell activation, senescence, exhaustion and skewing towards TH17 in patients with COVID-19 pneumonia. *Nat. Commun.* 11:3434. doi: 10.1038/s41467-020-17292-4

Ding, S., Wang, D., Zhou, X., Chen, L., Feng, K., Xu, X., et al. (2022). Predicting heart cell types by using Transcriptome profiles and a machine learning method. *Life* 12:228. doi: 10.3390/life12020228

Dramiński, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2007). Monte Carlo feature selection for supervised classification. *Bioinformatics* 24, 110–117. doi: 10.1093/bioinformatics/btm486

Durbin, A. F., Wang, C., Marcotrigiano, J., and Gehrke, L. (2016). RNAs containing modified nucleotides fail to trigger RIG-I conformational changes for innate immune signaling. *mBio* 7, e00833–e00816. doi: 10.1128/mBio.00833-16

Eberle, C., James-Todd, T., and Stichling, S. (2021). SARS-CoV-2 in diabetic pregnancies: a systematic scoping review. *BMC Pregnancy Childbirth* 21:573. doi: 10.1186/s12884-021-03975-3

Fan, R., Mao, S. Q., Gu, T. L., Zhong, F. D., Gong, M. L., Hao, L. M., et al. (2017). Preliminary analysis of the association between methylation of the ACE2 promoter and essential hypertension. *Mol. Med. Rep.* 15, 3905–3911. doi: 10.3892/mmr.2017.6460

Fehr, A. R., Singh, S. A., Kerr, C. M., Mukai, S., Higashi, H., and Aikawa, M. (2020). The impact of PARPs and ADP-ribosylation on inflammation and host-pathogen interactions. *Genes Dev.* 34, 341–359. doi: 10.1101/gad.334425.119

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* 28, 367–374. doi: 10.1016/j.compbiolchem.2004.09.006

Hadjadj, J., Yatim, N., Barnabei, L., Corneau, A., Boussier, J., Smith, N., et al. (2020). Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science* 369, 718–724. doi: 10.1126/science.abc6027

Hatta, Y., Hershberger, K., Shinya, K., Proll, S. C., Dubielzig, R. R., Hatta, M., et al. (2010). Viral replication rate regulates clinical outcome and CD8 T cell responses during highly pathogenic H5N1 influenza virus infection in mice. *PLoS Pathog.* 6:e1001139. doi: 10.1371/journal.ppat.1001139

Kianmehr, A., Faraoni, I., Kucuk, O., and Mahrooz, A. (2021). Epigenetic alterations and genetic variations of angiotensin-converting enzyme 2 (ACE2) as a functional receptor for SARS-CoV-2: potential clinical implications. *Eur. J. Clin. Microbiol. Infect. Dis.* 40, 1587–1598. doi: 10.1007/s10096-021-04264-9

Kim, E. S., Choe, P. G., Park, W. B., Oh, H. S., Kim, E. J., Nam, E. Y., et al. (2016). Clinical progression and cytokine profiles of Middle East respiratory syndrome coronavirus infection. *J. Korean Med. Sci.* 31, 1717–1725. doi: 10.3346/jkms.2016.31.11.1717

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, *Vol. 2* (Montreal, QC: Morgan Kaufmann Publishers Inc.).

Konigsberg, I. R., Barnes, B., Campbell, M., Davidson, E., Zhen, Y., Pallisard, O., et al. (2021). Host methylation predicts SARS-CoV-2 infection and clinical outcome. *Commun Med (London)* 1:42. doi: 10.1038/s43856-021-00042-y

Lee, A. C., Castaneda, G., Li, W. T., Chen, C., Shende, N., Chakladar, J., et al. (2021). COVID-19 severity potentially modulated by cardiovascular-disease-associated immune dysregulation. *Viruses* 13:1018. doi: 10.3390/v13061018

Li, Y. C., Bai, W. Z., and Hashikawa, T. (2020). The neuroinvasive potential of SARS-CoV2 may play a role in the respiratory failure of COVID-19 patients. *J. Med. Virol.* 92, 552–555. doi: 10.1002/jmv.25728

Li, S., Duan, X., Li, Y., Li, M., Gao, Y., Li, T., et al. (2021). Differentially expressed immune response genes in COVID-19 patients based on disease severity. *Aging (Albany NY)* 13, 9265–9276. doi: 10.18632/aging.202877

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

Li, X., Lu, L., and Chen, L. (2022). Identification of protein functions in mouse with a label space partition method. *Math. Biosci. Eng.* 19, 3820–3842. doi: 10.3934/mbe.2022176

Li, H. B., Tong, J., Zhu, S., Batista, P. J., Duffy, E. E., Zhao, J., et al. (2017). M(6)a mRNA methylation controls T cell homeostasis by targeting the IL-7/STAT5/SOCS pathways. *Nature* 548, 338–342. doi: 10.1038/nature23450

Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., et al. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* 26, 842–844. doi: 10.1038/s41591-020-0901-9

Liu, B. M., and Hill, H. R. (2020). Role of host immune and inflammatory responses in COVID-19 cases with underlying primary immunodeficiency: a review. *J. Interf. Cytokine Res.* 40, 549–554. doi: 10.1089/jir.2020.0210

Liu, S. Y., Sanchez, D. J., Aliyari, R., Lu, S., and Cheng, G. (2012). Systematic identification of type I and type II interferon-induced antiviral factors. *Proc. Natl. Acad. Sci. U. S. A.* 109, 4239–4244. doi: 10.1073/pnas.1114981109

Liu, H., and Setiono, R. (1998). Incremental feature selection. *Appl. Intell.* 9, 217–230. doi: 10.1023/A:1008363719778

Looi, C. K., Hii, L. W., Ngai, S. C., Leong, C. O., and Mai, C. W. (2020). The role of Ras-associated protein 1 (Rap1) in cancer: bad actor or good player? *Biomedicine* 8:334. doi: 10.3390/biomedicines8090334

Luo, X., Peng, Y., Chen, Y. Y., Wang, A. Q., Deng, C. W., Peng, L. Y., et al. (2021). Genome-wide DNA methylation patterns in monocytes derived from patients with primary Sjogren syndrome. *Chin. Med. J.* 134, 1310–1316. doi: 10.1097/CM9.0000000000001451

Mateus, J., Grifoni, A., Tarke, A., Sidney, J., Ramirez, S. I., Dan, J. M., et al. (2020). Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. *Science* 370, 89–94. doi: 10.1126/science.abd3871

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-protein. *Structure* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

Meertens, L., Hafirassou, M. L., Couderc, T., Bonnet-Madin, L., Kril, V., Kümmerer, B. M., et al. (2019). FHL1 is a major host factor for chikungunya virus infection. *Nature* 574, 259–263. doi: 10.1038/s41586-019-1578-4

Menachery, V. D., Schäfer, A., Burnum-Johnson, K. E., Mitchell, H. D., Eisfeld, A. J., Walters, K. B., et al. (2018). MERS-CoV and H5N1 influenza virus antagonize antigen presentation by altering the epigenetic landscape. *Proc. Natl. Acad. Sci. U. S. A.* 115, E1012–e1021. doi: 10.1073/pnas.1706928115

Onesime, M., Yang, Z., and Dai, Q. (2021). Genomic Island prediction via Chi-Square test and random Forest algorithm. *Comput. Math. Methods Med.* 2021, 1–9. doi: 10.1155/2021/9969751

Park, A., and Iwasaki, A. (2020). Type I and type III Interferons – induction, signaling, evasion, and application to combat COVID-19. *Cell Host Microbe* 27, 870–878. doi: 10.1016/j.chom.2020.05.008

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.

Puthia, M., Ambite, I., Cafaro, C., Butler, D., Huang, Y., Lutay, N., et al. (2016). IRF7 inhibition prevents destructive innate immunity-a target for nonantibiotic therapy of bacterial infections. *Sci. Transl. Med.* 8:336ra359. doi: 10.1126/scitranslmed.aaf1156

Qiu, X., Hua, X., Li, Q., Zhou, Q., and Chen, J. (2021). M(6)a regulator-mediated methylation modification patterns and characteristics of immunity in blood leukocytes of COVID-19 patients. *Front. Immunol.* 12:774776. doi: 10.3389/fimmu.2021.774776

Ran, B., Chen, L., Li, M., Han, Y., and Dai, Q. (2022). Drug-drug interactions prediction using fingerprint only. *Comput. Math. Methods Med.* 2022, 1–14. doi: 10.1155/2022/7818480

Rho, S. S., Ando, K., and Fukuhara, S. (2017). Dynamic regulation of vascular permeability by vascular endothelial cadherin-mediated endothelial cell-cell junctions. *J. Nippon Med. Sch.* 84, 148–159. doi: 10.1272/jnms.84.148

Safavian, S. R., and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* 21, 660–674. doi: 10.1109/21.97458

Schulte-Schrepping, J., Reusch, N., Paclik, D., Baßler, K., Schlickeiser, S., Zhang, B., et al. (2020). Severe COVID-19 is marked by a Dysregulated myeloid cell compartment. *Cells* 182, 1419–1440.e23. doi: 10.1016/j.cell.2020.08.001

Sekine, T., Perez-Potti, A., Rivera-Ballesteros, O., Strålin, K., Gorin, J. B., Olsson, A., et al. (2020). Robust T cell immunity in convalescent individuals with asymptomatic or mild COVID-19. *Cells* 183, 158–168.e14. doi: 10.1016/j.cell.2020.08.017

Sen, R., Garbati, M., Bryant, K., and Lu, Y. (2021). Epigenetic mechanisms influencing COVID-19. *Genome* 64, 372–385. doi: 10.1139/gen-2020-0135

Shathasivam, T., Kislinger, T., and Gramolini, A. O. (2010). Genes, proteins and complexes: the multifaceted nature of FHL family proteins in diverse tissues. *J. Cell. Mol. Med.* 14, 2702–2720. doi: 10.1111/j.1582-4934.2010.01176.x

Shu, C., Justice, A. C., Zhang, X., Wang, Z., Hancock, D. B., Johnson, E. O., et al. (2020). DNA methylation mediates the effect of cocaine use on HIV severity. *Clin. Epigenetics* 12:140. doi: 10.1186/s13148-020-00934-1

Sulzmaier, F. J., Jean, C., and Schlaepfer, D. D. (2014). FAK in cancer: mechanistic findings and clinical applications. *Nat. Rev. Cancer* 14, 598–610. doi: 10.1038/nrc3792

Tang, S., and Chen, L. (2022). iATC-NFMLP: identifying classes of anatomical therapeutic chemicals based on drug networks, fingerprints and multilayer perceptron. *Curr. Bioinforma.* 17. doi: 10.2174/1574893617666220318093000 [Epub ahead of print].

Toor, S. M., Saleh, R., Sasidharan Nair, V., Taha, R. Z., and Elkord, E. (2021). T-cell responses and therapies against SARS-CoV-2 infection. *Immunology* 162, 30–43. doi: 10.1111/imm.13262

Wang, R., and Chen, L. (2022). Identification of human protein subcellular location with multiple networks. Current. *Proteomics* 19, 344–356. doi: 10.2174/1570164619666220531113704

Wang, J., Huang, F., Huang, J., Kong, J., Liu, S., and Jin, J. (2017). Epigenetic analysis of FHL1 tumor suppressor gene in human liver cancer. *Oncol. Lett.* 14, 6109–6116. doi: 10.3892/ol.2017.6950

Wu, Z., and Chen, L. (2022). Similarity-based method with multiple-feature sampling for predicting drug side effects. *Comput. Math. Methods Med.* 2022, 1–13. doi: 10.1155/2022/9547317

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovations* 2:100141. doi: 10.1016/j.xinn.2021.100141

Xing, J., Zhang, A., Du, Y., Fang, M., Minze, L. J., Liu, Y. J., et al. (2021). Identification of poly(ADP-ribose) polymerase 9 (PARP9) as a noncanonical sensor for RNA virus in dendritic cells. *Nat. Commun.* 12:2681. doi: 10.1038/s41467-021-23003-4

Yamamoto, K., Takagi, Y., Ando, K., and Fukuhara, S. (2021). Rap1 small GTPase regulates vascular endothelial-cadherin-mediated endothelial cell-cell junctions and vascular permeability. *Biol. Pharm. Bull.* 44, 1371–1379. doi: 10.1248/bpb.b21-00504

Yang, Y., and Chen, L. (2022). Identification of drug–disease associations by using multiple drug and disease networks. *Curr. Bioinforma.* 17, 48–59. doi: 10.2174/1574893616666210825115406

Yao, Y., Yang, Y., Guo, W., Xu, L., You, M., Zhang, Y. C., et al. (2021). METTL3-dependent m(6)a modification programs T follicular helper cell differentiation. *Nat. Commun.* 12:1333. doi: 10.1038/s41467-021-21594-6

Zhang, C., Feng, Y. G., Tam, C., Wang, N., and Feng, Y. (2021). Transcriptional profiling and machine learning unveil a concordant biosignature of type I interferon-inducible host response across nasal swab and pulmonary tissue for COVID-19 diagnosis. *Front. Immunol.* 12:733171. doi: 10.3389/fimmu.2021.733171

Zhang, Y., Mao, D., Roswit, W. T., Jin, X., Patel, A. C., Patel, D. A., et al. (2015). PARP9-DTX3L ubiquitin ligase targets host histone H2BJ and viral 3C protease to enhance interferon signaling and control viral infection. *Nat. Immunol.* 16, 1215–1227. doi: 10.1038/ni.3279

Zhang, Y.-H., Zeng, T., Chen, L., Huang, T., and Cai, Y.-D. (2021). Determining protein–protein functional associations by functional rules based on gene ontology and KEGG pathway. *Biochim. Biophys. Acta Proteins Proteom.* 1869:140621. doi: 10.1016/j.bbapap.2021.140621

Zhou, J.-P., Chen, L., and Guo, Z.-H. (2020a). iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs. *Bioinformatics* 36, 1391–1396. doi: 10.1093/bioinformatics/btz757

Zhou, J.-P., Chen, L., Wang, T., and Liu, M. (2020b). iATC-FRAKEL: a simple multi-label web-server for recognizing anatomical therapeutic chemical classes of drugs with their fingerprints only. *Bioinformatics* 36, 3568–3569. doi: 10.1093/bioinformatics/btaa166

Zhou, X., Ding, S., Wang, D., Chen, L., Feng, K., Huang, T., et al. (2022). Identification of cell markers and their expression patterns in skin based on single-cell RNA-sequencing profiles. *Life* 12:550. doi: 10.3390/life12040550

Zhu, H., Wu, L. F., Mo, X. B., Lu, X., Tang, H., Zhu, X. W., et al. (2019). Rheumatoid arthritis-associated DNA methylation sites in peripheral blood mononuclear cells. *Ann. Rheum. Dis.* 78, 36–42. doi: 10.1136/annrheumdis-2018-213970

Zhu, L., Yang, P., Zhao, Y., Zhuang, Z., Wang, Z., Song, R., et al. (2020). Single-cell sequencing of peripheral mononuclear cells reveals distinct immune response landscapes of COVID-19 and influenza patients. *Immunity* 53:e683, 685–696.e3. doi: 10.1016/j.immuni.2020.07.009

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 06 October 2022
DOI 10.3389/fmicb.2022.1004664

OPEN ACCESS

EDITED BY
Fei Ma,
Chinese Academy of Medical Sciences and Peking Union Medical College, China

REVIEWED BY
Nazan Yurtcu,
Sivas Cumhuriyet University Faculty of Medicine, Turkey
Ana Afonso,
Universidade NOVA de Lisboa, Portugal
Yi Zhao,
Institute of Computing Technology (CAS), China

*CORRESPONDENCE
Bo Meng
bo0328@qq.com
Geng Tian
Tiang@geneis.cn

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

RECEIVED 27 July 2022
ACCEPTED 20 September 2022
PUBLISHED 06 October 2022

CITATION
Hu J, Wu Y, Quan L, Yang W, Lang J, Tian G and Meng B (2022) Research of cervical microbiota alterations with human papillomavirus infection status and women age in Sanmenxia area of China.
Front. Microbiol. 13:1004664.
doi: 10.3389/fmicb.2022.1004664

Research of cervical microbiota alterations with human papillomavirus infection status and women age in Sanmenxia area of China

Jintao Hu[1,2], Yuhan Wu[3], Lili Quan[4], Wenjuan Yang[2], Jidong Lang[2], Geng Tian[2]* and Bo Meng[2]*

[1]Faculty of Engineering and Information Technology, The University of Melbourne, Parkville, VIC, Australia, [2]Genesis (Beijing) Co., Ltd., Beijing, China, [3]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China, [4]Department of Gynecology, Sanmenxia Central Hospital of Henan University of Science and Technology, Sanmenxia, Henan, China

Background: Human papillomavirus (HPV) infection is the leading cause of cervical cancer. More and more studies discovered that cervical microbiota (CM) composition correlated with HPV infection and the development of cervical cancer. However, more studies need to be implemented to clarify the complex interaction between microbiota and the mechanism of disease development, especially in a specific area of China.

Materials and methods: In this study, 16S rDNA sequencing was applied on 276 Thin-prep Cytologic Test (TCT) samples of patients from the Sanmenxia area. Systematical analysis of the microbiota structure, diversity, group, and functional differences between different HPV infection groups and age groups, and co-occurrence relationships of the microbiota was carried out.

Results: The major microbiota compositions of all patients include *Lactobacillus iners*, *Escherichia coli*, *Enterococcus faecalis,* and *Atopobium vaginae* at species level, and *Staphylococcus*, *Lactobacillus*, *Gardnerella*, *Bosea*, *Streptococcus,* and *Sneathia* in genus level. Microbiota diversity was found significantly different between HPV-positive (Chao1 index: 98.8869, $p < 0.01$), unique-268 infected (infections with one of the HPV genotype 52, 56, or 58, 107.3885, $p < 0.01$), multi-268 infected (infections with two or more of HPV genotype 52, 56, and 58, 97.5337, $p = 0.1012$), other1 (94.9619, $p < 0.05$) groups and HPV-negative group (83.5299). Women older than 60 years old have higher microbiota diversity (108.8851, $p < 0.01$, $n = 255$) than younger women (87.0171, $n = 21$). The abundance of *Gardnerella* and *Atopobium vaginae* was significantly higher in the HPV-positive group than in the HPV-negative group, while *Burkholderiaceae* and *Mycoplasma* were more abundant in the unique-268 group compared to the negative group. *Gamma-proteobacteria* and *Pseudomonas* were found more abundant in older than 60 patients than younger groups. Kyoto Encyclopedia of Genes and Genomes (KEGG) and Clusters of Orthologous Groups (COG) analysis

revealed the effects on metabolism by microbiota that the metabolism of cells, proteins, and genetic information-related pathways significantly differed between HPV-negative and positive groups. In contrast, lipid metabolism, signal transduction, and cell cycle metabolism pathway significantly differed between multi-268 and negative groups.

Conclusion: The HPV infection status and age of women were related to CM's diversity and function pathways. The complex CM co-occurrent relationships and their mechanism in disease development need to be further investigated.

KEYWORDS

cervical cancer, TCT, HPV - human papillomavirus, microbiota, diversity

Introduction

Cervical cancer is one of the most common malignant tumours among women. Clinical and epidemiological studies have determined that persistent Human Papillomavirus (HPV) infection is the leading risk factor for developing cervical cancer (Khan et al., 2020). The average time interval from carcinogenic HPV infection to cervical cancer progression is 25–30 years. The Thin-prep Cytologic Test (TCT) and HPV DNA were recommended to be used for HPV infection and cervical cancer status determination (Zhang et al., 2020). TCT detects the morphology of the cells and analyses the bacterial population in the sample through 16S rDNA (Deoxyribonucleic Acid) (Liu et al., 2020), which makes it possible to perform large-scale testing of samples.

Recent research revealed that microbiota might be a significant factor in the relationship between HPV and cervical cancer. Klein et al. found that changes in the cervical microbiota (CM) are related to cervical cancer (Kunene and Mahlangu, 2017). Some studies have shown that cervical/vaginal *Lactobacilli* can produce lactic acid that inhibits the growth of bacteria associated with bacterial vaginosis (BV) and viral infections (Polatti, 2012). The change in the proportion of microorganisms is related to pathological changes in the reproductive tract. In addition, recent studies have shown a clear correlation between microbiota and HPV infection (Libby et al., 2008; Anahtar et al., 2015). A report also explained a positive correlation between cervical HPV infection and BV-related microbiota (Gillet et al., 2011; Lee et al., 2013). Therefore, microbiota may play an important role in between, which implies that the reveal of the mechanism microbiota play is beneficial to comprehend the HPV infection and cancer evolution. However, the current research results have not clarified the mutual influence (Libby et al., 2008; Anahtar et al., 2015). So far, there are relatively few studies on the association amongst CM, cervical cancer and HPV infection, especially in China, which prompted this research to be conducted.

Therefore, samples from 276 patients were obtained to conduct microbiota research for further analysis. This research aimed to explore the relationship between the HPV infection and CM changes by analysing microbiota changes and the HPV infection concerning HPV infection group, different genotype HPV groups, and the impact of microbiota on cell and metabolic functions, as well as to explore the microbiota changes amongst the group divided by ages in a cohort of populations in Sanmenxia, Henan Province. Aside from that, this research also provides a new reference basis for further understanding the CM's overall characteristics.

Materials and methods

Study population and specimen collection

A total of 276 cervical lesion samples were collected from patients at Sanmenxia Central Hospital for high-grade squamous intraepithelial lesion (HSIL) screening. The hospital's medical ethics committee approved this study, and all experiments were carried out following the relevant guidelines and regulations. A fluorescent HPV genotyping kit (Bioperfectus Technologies, Jiangsu, China) was used to analyse the samples for confirmation and subsequent HPV typing. Women who came to Sanmenxia central hospital to do cervical tests in 2019, including TCT test and HPV genotype test, were enrolled in this study. Those samples were not qualified for further study, or low-quality sequencing results were excluded from the study.

DNA extraction

After Pap Smear preparation, 1-ml of the remaining fluid sample was used for DNA isolation. According to the manufacturer's instructions, the total Genomic DNA sample was

extracted using TIANamp Micro DNA Kit (TIANGEN, Beijing, China). The double-stranded (ds) DNA was quantified using a Nanodrop 2000 and Qubit dsDNA HS assay kit (Thermo Fisher Scientific, Inc., Waltham, MA, USA). The average fragment size of DNA (> 5 Kbp) was measured (identified by comparison to a DL2000 PLUS DNA Ladder, Life Technologies, Carlsbad, CA, USA) on a 1.0% agarose gel in 1×TAE buffer on a Bio-Rad CHEF DRII system.

Sequencing and bioinformatic processing

To build a sequencing library, use PCR primers to amplify the V3–V4 hypervariable region of the 16S rDNA gene. This area provides sufficient information for the taxonomic classification of microbial communities in specimens related to human microbiota research and is used by the Human Microbiota Project.

Then use Agencourt AMPure XP (Beckman Coulter, Indianapolis, Indiana) to select the product size in a ratio of 0.9 and group in equal moles. Then, a Qubit 2.0 fluorometer (Life Technologies) was used to quantify the pool of the selected size and loaded into the Illumina HiSeq flow cell (Illumina, Inc., San Diego, CA, USA) 2 × 250. Mix the library with the Illumina-generated PhiX control library and our genomic library, and use fresh NaOH for denaturation. Perform image analysis, base calls, and data quality assessment on the MiSeq instrument.

Data analysis

Paired-end sequencing (2 × 250) was performed on Illumina HiSeq. The FASTQ conversion of the original data file is completed after demultiplexing with MiSeq Reporter. The quality assessment of FASTQ files is carried out using FASTQC,[1] then quality filtering is performed using the FASTX toolkit.[2] The high-quality reads used for analysis (where 80% of the base Q scores > 20 reads) and reads with unknown bases ("N") are discarded. The remaining steps are performed using the Quantitative Analysis of Microbial Ecology (QIIME) software package version 1.8. Use UCHIME to filter chimeric sequences and use UCLUST to group sequences into the Operational Taxa Unit (OTU) with a similarity threshold of 97%. The Ribosomal Database Program (RDP) classifier trained using Greengenes 16S rDNA database (v13.8) assigns all OTUs to all OTUs with a confidence threshold of 80%. OTUs with an average abundance of less than 0.005% are eliminated. Use PyNAST v1.2 for multiple sequence alignment and FastTree v2.1 to construct a phylogenetic tree. The alpha and beta diversity indicators are calculated according to the method implemented in QIIME. A Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) is used to predict the orthologs from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Clusters of Orthologous Groups (COG) (i.e., the count of functional genes) for each sample and inferred genes. The count is allocated to the KEGG and COG channels.

Statistic analysis

The numeric values representing the relative abundance of the OTU were analysed for statistical significance by performing a *t*-test. The statistical software in Sigma Plot version 11.0 was used (Systat Software Inc., San Jose, CA, USA). A total of 95% confidence intervals were estimated for sensitivity and specificity with a binomial test. Differences in means with *p*-values less than 0.05 were considered statistically significant. When tests for normality failed, the non-parametric data were analysed using the Mann–Whitney Rank Sum test and median values were determined.

Results

Patient cohort sociodemographic and characteristics

Complete data were available for 276 cervical smear samples taken from eligible women. The patients' characteristics of the study population were summarised in **Table 1**, and detailed information about each patient was shown in **Supplementary Table 1**. The mean age of the patient's cohort was 44.68 ± 10.94, from 17 to 80 years old. There were 83 HPV-infected patients in the 255 candidates under 60 years old and 11 HPV-infected patients in the 21 candidates over 60. The infection status of the subjects, including HPV single-type infection and multi-type infection. Five majors HPV types were detected, including HPV16 (n = 10), HPV39 (n = 7), HPV52 (n = 8), HPV56 (n = 7), and HPV58 (n = 12), and 36 subjects were infected by other HPV genotypes. Several cases were detected as unique-268 infections (52, 56, or 58 genotype infections, n = 27) among the 80 single infections and other related co-infections were multi-268 infections (52, 56, and 58 genotypes infecting two or more, n = 13) among the 14 multiple infections. Hence, the number of infections excluding 52, 56, and 58 were 54 infections, denoting as other1. Among the 21 subjects diagnosed with Cervical Intraepithelial Neoplasms (CIN), thirteen patients were infected by HPV and eight were not. While 81 HPV-positive and 174 (255 totally) HPV-negative subjects were identified in subjects without CIN status.

1 http://www.Bioinformatics.babraham.ac.uk/projects/fastqc/

2 http://hannonlab.cshl.edu/fastx_toolkit/

TABLE 1 Patients age and infection status.

Characteristics	Positive	Negative
Age		
A	83	172
A60	11	10
HPV infection situation		
Single-type infection	80	–
HPV16	10	–
HPV39	7	–
HPV52	8	–
HPV56	7	–
HPV58	12	–
Other	36	–
Unique-268	27	–
Multi-type infection	14	–
Muiti-268	13	–
Other1	54	–
CIN suffering		
CIN	13	8
Normal	81	174

A refers to age under 60, A60 refers to age over 60; Unique-268 refers to HPV single infection with one of 52, 56, or 58 subtypes, multi-268 refers to HPV multiple infections with at least two of 52, 56, or 58 subtypes, other1 refers to HPV infections besides 52, 56, and 58 subtypes. *Denotes no data.

Baseline composition similar but the structure of the cervical microbiota variant between samples

The average length of the PCR product was about 465 bps from V3 to V4 segments of the 16S rDNA genes. After sequencing, the data amount and quality of each sample were evaluated by GC content (averagely 52.8%), Q20 value (averagely 96.2%), Q30 value (averagely 92%), and effectiveness (averagely 73.3). The detailed information of each sample was summarised in **Supplementary Table 1**. After removing the low-quality sequencing reads, a total number of 14,435,817 clean tags and an average of 79,095 tags of each specimen (each specimen generating at least 8,807 clean tags) were obtained. After the removal of singletons and rare OTUs (species abundance less than 0.005%), a total of 11 phyla, 17 classes, 30 orders, 53 families, 99 genera, and 132 species (**Supplementary Table 2**) were identified from the study cohort sequencing data.

The abundance of the ten most abundant bacteria families was summarised and listed in **Supplementary Table 3**. *Lactobacillaceae*, *Enterobacteriaceae*, *Staphylococcaceae*, *Enterococcaceae*, *Bifidobacteriaceae*, *Beijerinckiaceae*, *Streptococcaceae*, *Leptotrichiaceae*, *Burkholderiaceae,* and *Corynebacteriaceae* were the top 10 most abundant bacteria families in all the samples. The corresponding distribution figure is shown in **Supplementary Figure 1**. Results showed that the microbiota structure variant obviously between samples, but there were still pattern similarities between parts of the samples. As previously published papers mentioned, the CM was classified into five clusters (Zhou et al., 2020). Our results showed sample clusters with highly abundant *Escherichia-Shigella*, *Lactobacillus*, *Enterococcus*, *Staphylococcus,* and *Lactobacillus* in each cluster, respectively (**Supplementary Figure 1**). Moreover, some samples showed different characteristics with more bacterium types in the structure. Sample alpha diversity, including Chao1, ACE, Shannon, and Simpson indexes of all the samples, was calculated and summarised in **Supplementary Table 4**, indicating that the diverse samples differentiated significantly between samples.

Differences in human papillomavirus infection status or age are highly relevant to the change of microbiota structure and diversity

This study explored the difference in microbiota composition and diversity between sample groups. Two group pairs showed significant differences in HPV infection status and age. The relative abundance results of different HPV infection status groups (**Figure 1A** and **Supplementary Table 4**) showed the following conclusions: (1) Both normal sample groups with HPV infection (n = 78) and non-infection (n = 187) were predominated by the following types of bacteria, including *Lactobacillus iners* (HPV-positive/HPV-negative: 0.185/0.104), *Escherichia coli* (0.112/0.143), *Enterococcus faecalis* (0.071/0.108), and *Atopobium vaginae* (0.033/0.013) in species level, and *Staphylococcus* (0.116/0.117), *Lactobacillus* (excluding *Lactobacillus iners AB1*, 0.078/0.069), *Gardnerella* (0.076/0.048), *Bosea* (0.026/0.049), *Streptococcus* (0.015/0.043), and *Sneathia* (0.031/0.020) in genus level; (2) In the microbiota structure (relative abundance), which may illustrate a structure change before and after the HPV infection (**Figure 1A** and **Supplementary Table 4**). Further study on the unique-268 (sum of 52, 56, and 58 genotype HPV infection cases, n = 27) and multi-268 (co-infections with one or two HPV 52, 56, and 58 genotypes, n = 13) types of HPV infected samples, the microbiota detected was almost the same as the HPV-positive except for the composition proportion (**Figure 1B** and **Supplementary Table 4**). Additionally, the age group under 60 (n = 255) and above 60 (n = 31) detected a similar microbiota with a different proportion of composition (**Figure 1C** and **Supplementary Table 4**).

Alpha diversity was applied to analyse the complexity of species diversity. Regarding the analysis, the HPV-positive group had higher diversity (Chao1 index: 98.8869, $p < 0.01$) compared to the negative group (Chao1 index: 83.5299, **Figures 1C,D** and **Supplementary Table 5**). This disparity in microbiota diversity resulted from the significant differences

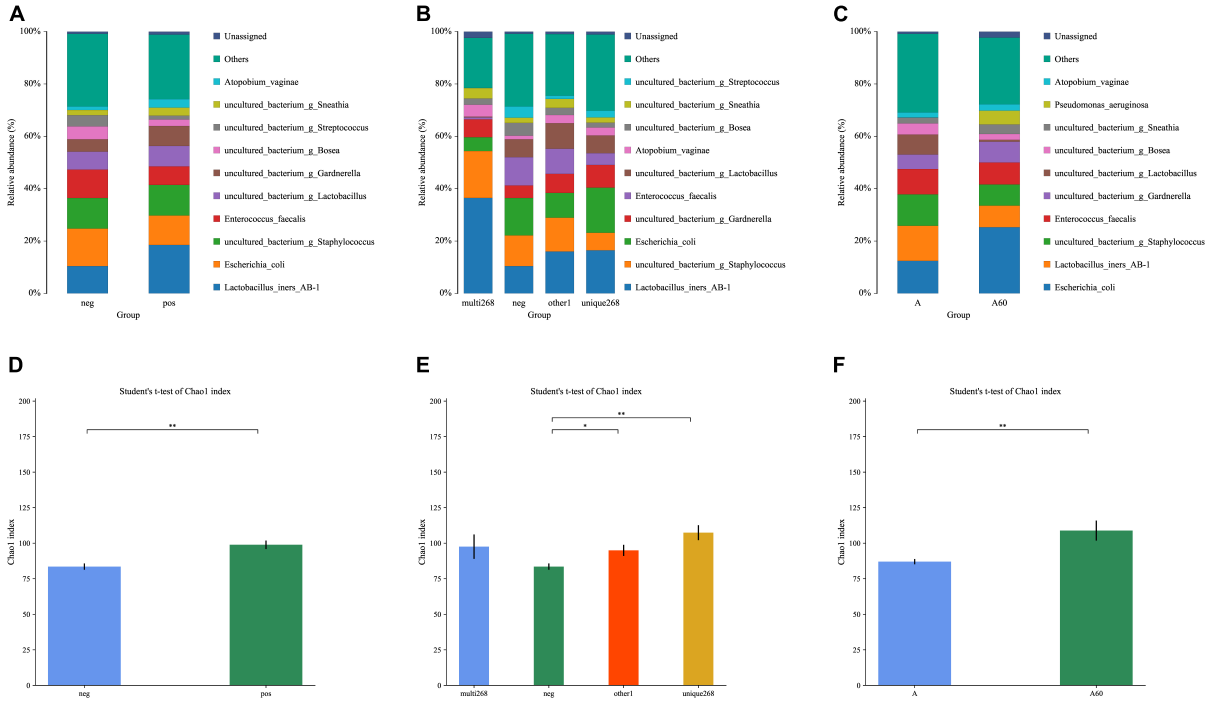


FIGURE 1
Relative abundance and alpha-diversity analysis of this study. **(A)** Microbiota relative abundance distribution of HPV-positive and HPV-negative patient groups; **(B)** microbiota relative abundance distribution of HPV-positive unique-268 infection and multiple 268 infection, and HPV-negative patient groups; **(C)** microbiota relative abundance analysis of the elder (age > 60) and the younger (age ≤ 60) patient groups; alpha-diversity analysis; a-diversity comparison bar diagram, **(D)** Chao1 index for normal (without CIN) HPV-positive and HPV-negative patient groups, **(E)** Chao1 index for HPV-positive unique-268, and multiple 268 infections, HPV-negative patient groups, **(F)** Chao1 index for the elder (age > 60) and the younger (age ≤ 60) patient groups. neg: HPV-negative without CIN, pos: HPV-positive without CIN, multi-268: HPV 52, 56, and 58 genotypes multiple infection (regardless of CIN), unique268: HPV 52, 56, or 58 genotypes unique (single) infection (regardless of CIN), other1: other HPV infection situations (regardless of CIN). A: age under 60, A60: age above 60, *p-value < 0.05, **p-value < 0.01.

between the two cases and evidenced structural change. Unique-268 (Chao1 index: 107.3885, $p < 0.01$), multi-268 (Chao1 index: 97.53) and other1 (Chao1 index: 94.9619, $p < 0.05$) had a higher microbiota diversity compared to the HPV-negative groups (Figure 1E and Supplementary Table 5). In addition, compared to younger patients, the elder group (age > 60, $n = 31$) has a higher diversity with statistical significance (Chao1 index: 108.8851, $p < 0.01$) than the younger group (age ≤ 60, $n = 255$, Chao1 index: 87.0171, Figure 1F and Supplementary Table 5), which also demonstrates the difference in microbiota structure. Hence, this diversity analysis indicates the following conclusions: (1) The normal HPV-positive groups and (2) unique-268 HPV and other1 infections were more diverse in microbiota than the HPV-negative groups, while (3) the age group over 60 had higher diversity concerning the microbiota.

Bacteria biomarkers were identified in different subject groups

Linear discriminant analysis (LDA) score was used to compare the different bacteria of each group. The results showed that *Bifidobacteriales* (order), *Bifidobacteriaceae* (family), *Gardnerella* (genus), *Coriobacteriia* (class), *Atopobium vaginae* (species), and *Clostridia* (class) were higher in HPV-infected group compared with the negative (Figure 2A). Between multi-268 and unique-268 groups, *Betaproteobacteriales* (order), *Burkholderiaceae* (family), *Weeksellaceae* (family), *Flavobacteriales* (family), *Gardnerella* (genus), *Pseudomonas aeruginosa* (species), and *Mycoplasma* (genus) were found with higher relative abundance in unique-268. In contrast, *Saccharimonadales* (order), *Saccharimonadia* (class), *Patescibacteria* (phylum), *Bifidobacteriales* (order), and *Bifidobacteriaceae* (family) were higher in the multi-268 group (Figure 2B). Among them, *Bifidobacteriaceae* was the most significantly different between the two groups, indicating its strong association with multi-268 infection (Figure 2B). *Corynebacterium* (genus), *Lactobacillus iners AB-1* (species), *Bacilli* (class), and *Firmicutes* (phylum) were identified higher in the group with age younger than 60 and *Gamma-proteobacteria* (class) and *Pseudomonas* (genus) higher in the group older than 60 years old (Figure 2C).

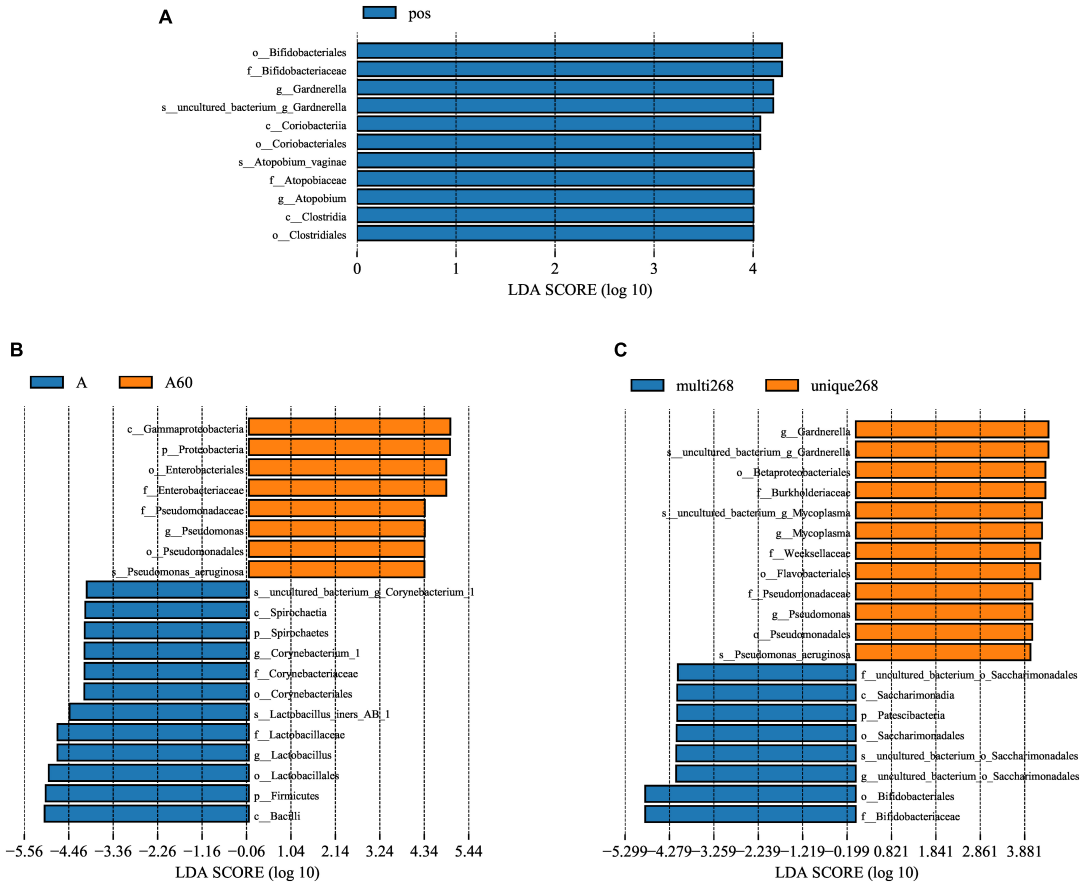


FIGURE 2
Microbiota significant difference analysed by LEfSe. **(A)** HPV-positive and HPV-negative patient groups, **(B)** HPV-positive single infection and multiple infection patient groups, and **(C)** age above and below 60 years old patient groups. LDA score threshold set as 4, above 4 will be shown in charts. LDA, linear discriminant analysis; LEfSe, LDA effective size. HPV, Human Papillomavirus. k_: kingdom, p_: phylum, c_: class_, o_: order, f_: family_, g_: genus.

Microbiota function difference of subjects group was identified

Kyoto Encyclopedia of Genes and Genomes and COG analysis were applied, and functional difference between groups was explored. Among the three group comparisons, two group pairs were found significantly different and they are HPV-negative/positive group pair and the multi-268/negative group pair. No significant function difference was identified between Age groups. Between HPV-positive and negative groups, KEGG pathways, including Cell growth and death, Excretory system, Folding, sorting, and degradation, Endocrine and metabolic diseases, Nucleotide metabolism, Replication and repair, Immune system, and Transcription, were significantly different (Figure 3A). Moreover, the COG categories of Amino acid transport and metabolism, Cell cycle control, cell division, chromosome partitioning, Inorganic ion transport and metabolism, Translation, ribosomal structure and biogenesis and Defence mechanisms between the two groups were significantly different (Figure 3B). Comparison analysis results of multi-268 and the negative group showed that KEGG pathway Excretory system, Lipid metabolism, Signal transduction and Folding, sorting and degradation and COG categories of Cell cycle control, cell division, chromosome partitioning were significantly different (Figures 3C,D).

Complexed cervical microbiota network relationships existed in the cervical microbiota system

Co-occurrent analysis of identified microbiota in cervical samples and results are shown in Figure 4. There were 80 genera of bacteria identified with more than seven relationships with other bacteria, and the correlation was higher than 10%, with a *p*-value less than 0.05. A network relationship was identified between them (Figure 4). The top 50 bacteria with high correlation were shown in the

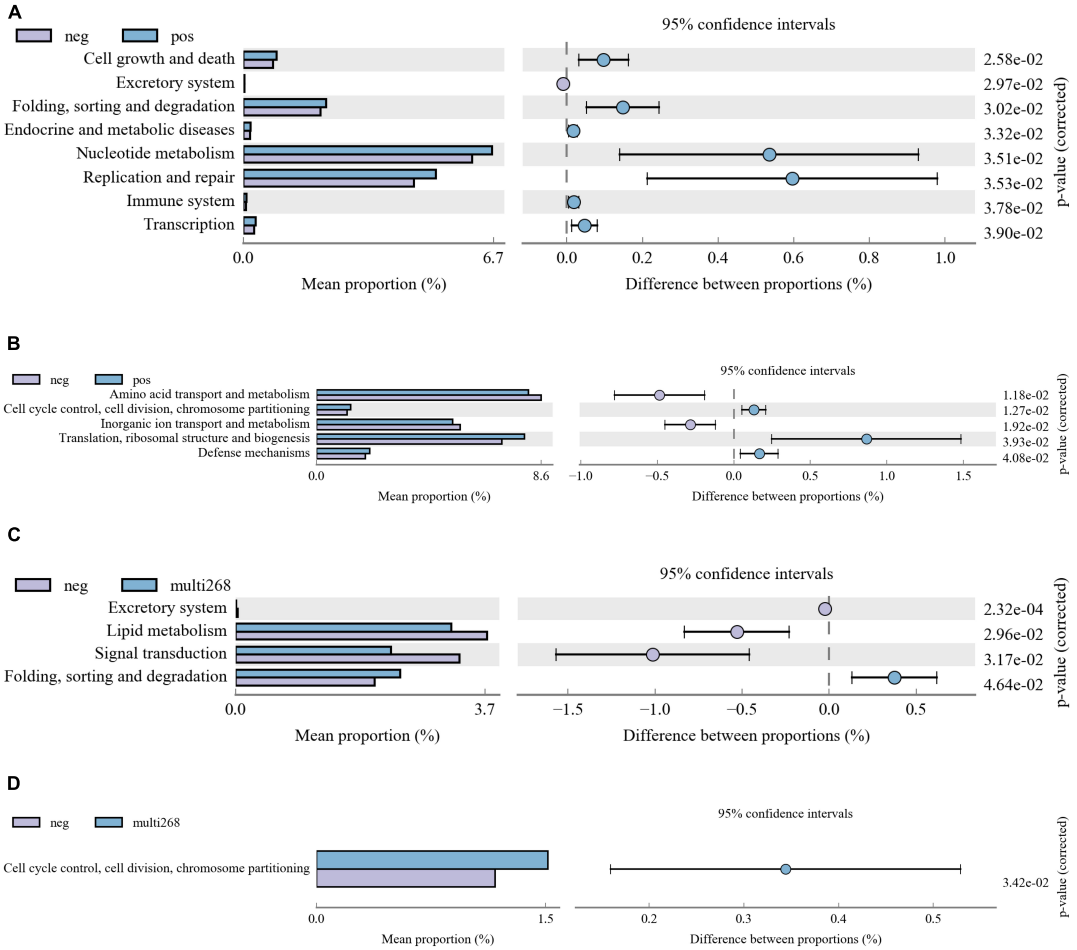


FIGURE 3
Microbiota KEGG and COG function difference diagrams. The picture shows the difference analysis diagram of the KEGG (and COG) metabolic pathway under the second level (also can be analysed for the third or first level): the different colours in the picture represent different groups. The figure shows the abundance ratio of different functions in the two sets of sample groups, the middle shows the difference ratio of the function abundance within the 95% confidence interval, and the rightmost value is the *p*-value. **(A)** KEGG pathway difference between HPV-positive and HPV-negative patient groups, **(B)** COG pathway difference between HPV-positive and HPV-negative patient groups. **(C)** KEGG pathway difference between multi-268 HPV infections and HPV-negative patient groups, **(D)** COG pathway difference between multi-268 HPV infections and HPV-negative patient groups. KEGG, Kyoto Encyclopedia of Genes and Genomes; COG, Clusters of Orthologous Groups.

figure (**Figure 4**), 24 of which were correlated with two or more other genera. The abundance of the genera was different, ranging from 12.0 to 15463.4, and no significant correlation was observed between genera abundance and its correlation with other genera. The abundance of *Lactobacillus* (abundance of 15463.4), *Escherichia-Shigella* (10609.8), and *Staphylococcus* (9270.3) were high, but the correlation with other bacteria is relatively low, less than 0.34, 0.32, and 0.27, respectively. On the contrary, some genera's abundance was relatively low, but the correlations with others were quite high (**Supplementary Table 6**). For example, *Atopobium* (abundance: 1510.4) is highly correlated to *Dialister* (correlation: 0.675, abundance: 236.2), *Prevotella* (0.656, 1259.3) and *Fastidiosipila* (0.573, 1121.7); *Achromobacter* (483.5) is tensely correlated to *Stenotrophomonas* (0.793, 305.4), *Sphingobium* (0.759, 241.9), and *Herbaspirillum* (0.648, 21.8); *Gardnerella (4349.4)* is highly correlated to *Atopobium* (0.659, 1510.4), *Aerococcus* (0.527, 246.2) and *Sneathia* (0.0.498, 2389.7); *Sneathia* (2389.7) is highly correlated to *Fastidiosipila* (0.648, 1121.7), DNF00809 (0.604, 242.4), *Parvimonas* (0.572, 278.8), and *Atopobium* (0.546, 1510.4). In summary, a complexed bacteria network relationship was existing in cervical system and the interactions between genera was not correlated with its abundance.

Discussion

Data analysis showed that changes in the cervical microbiome, especially anaerobic bacteria, were significantly correlated with HPV infection status. *Gardnerella*, *Atopobium*

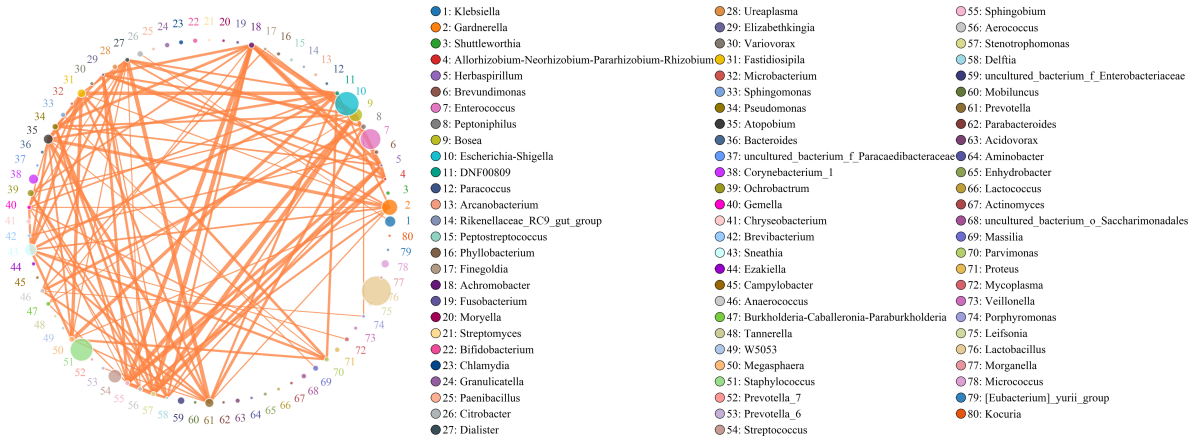


FIGURE 4
Microbiota network diagram. The top 50 most relevant genera were shown here. The circle in the figure represents the species, and the size of the circle represents the abundance; the line represents the relation between the two species, the thickness of the line represents the strength of the relation, and the colour of the line: orange represents positive correlation, and the green represents negative correlation.

vaginae, and *Sneathia* were the three most increased amongst microbiota, and these microorganisms form pathogenic biofilms through close cooperation. *Gardnerella* acts as a "scaffold" for biofilms (Harwich et al., 2010; Fethers et al., 2012; Curty et al., 2017; Khan et al., 2020), promoting the growth of *Atopobium vaginae* (Libby et al., 2008; Anahtar et al., 2015), *Sneathia* and other related pathogens by altering the microenvironment (Lee et al., 2013; Zhou et al., 2020). The growth of these pathogens has led to a rise in microbial diversity. Not only that, the cellular pro-inflammatory responses that these pathogenic microorganisms elicited will affect cellular metabolisms (Kacerovsky et al., 2015; Mitra et al., 2015, 2020; Onderdonk et al., 2016; Gosmann et al., 2017; Khan et al., 2020), such as amino acid transport (Mitra et al., 2020) and inorganic ion transport, and even cell shedding (Harwich et al., 2012; Africa et al., 2014). This deteriorates the immune response and leads to a defection of the microenvironment (Anahtar et al., 2015), making cervix HPV susceptible and possibly leading to the cervical cancer. Therefore, these three microorganisms are of great significance as biomarkers in the clinical identification of HPV infection.

In addition to the correlation with HPV, there is also a relationship between CM and age status. Lee et al. (2013) suggested that the decline in oestrogen and progesterone levels in the female reproductive tract after menopause is associated with an increased proportion of anaerobic bacteria. Several studies have also confirmed that older women have a higher proportion of anaerobic bacteria, and the biofilm produced by them is an important factor in HPV susceptibility (Singh et al., 2015; Zhou et al., 2020). Besides, Lee's experiment also evidenced that due to the presence of oestrogen and progesterone, the proportion of *Lactobacillus* in young women is higher to maintain the homeostasis of the reproductive tract microenvironment (Lee et al., 2013; Oh et al., 2015). In contrast, the proportion of *Lactobacillus* in older women is lower, so it insufficiently maintains the homeostasis of the internal environment, making the proportion of γ-proteobacteria and Pseudomonas species increase as biomarkers.

However, this study also showed some results that differed from the prevailing view. In this study, the proportion of *Lactobacillus* in the HPV-positive group was higher than that in the normal group. It contradicts the mainstream ideas that the presence of *Lactobacillus* can maintain pH stability (Larsson et al., 1991; Brotman et al., 2014) and homeostasis in the reproductive tract (Mitra et al., 2016; Borgogna et al., 2020) and is therefore reduced in the HPV-positive group. However, a report from Iran showed the same results as this research and concluded that the proportion change of *Lactobacillus* was not strongly correlated with HPV infection status but did not rule out the influence of factors such as customs on sample interference (Ghaniabadi et al., 2020). Therefore, the influence of other factors, such as personal habits, could not be ruled out for the presence of interference with the sample microbiota. The more detailed mechanisms still need more experiments to verify and analyse.

Aside from that, this study also presented some new findings. A higher abundance of *Bifidobacterium* was also found in the multi-268 group than in the unique-268 group when comparing the two case samples. Under such circumstances, the same bacteria have different environmental adaptations in different case samples. Besides, by analysing the low-abundance flora of different groups, it was found that there were differences in the microbial composition of different HPV infection states. Comparing the LEfSe analysis of unique-268 and multi-268, it could be seen that some low-abundance bacterial groups play important roles in different HPV-infected samples. In the

unique-268 group of patients, *Burkholderiaceae*, a pathogenic bacteria, could sensitise cells to HPV (Brenner et al., 2005). *Mycoplasma* could promote HPV penetration, survival and persistence, and it is frequently present in high-risk HPV patients (Biernat Sudolska et al., 2011; Ye et al., 2018; Wei et al., 2021). *Pseudomonas aeruginosa* is a prevalent factor in high-risk HPV samples, especially in the cancerous cervix (Werner et al., 2012; Di Paola et al., 2017; Zhang et al., 2021). The low-abundance species *Saccharimonadales*, *Saccharimonadia,* and *Patescibacteria* in the multi-268 group were all related to the synthesis of compound elements (Herrmann et al., 2019; Lemos et al., 2019; Tian et al., 2020; Hosokawa et al., 2021; Mason et al., 2021; Zhou et al., 2021; Wang et al., 2022). The results of the above microbiota under different HPV infection statuses have clinical implications for biomarkers for identifying cases.

In addition to the above microorganisms, this study also found that the microbiota (especially pathogenic microorganisms) significantly impacted metabolic function. Apart from the abnormal cellular metabolism mentioned above, differences in genetic metabolism, lipid metabolism, signal transduction and cell cycle metabolism were also detected between the HPV-positive group and the multi-268 group. Abnormalities in these functions are likely associated with increased microbial diversity and an increased proportion of pathogenic microorganisms (Mitra et al., 2020). However, there are few studies in this regard, so further experiments are needed to explore their relationship.

This study analysed the possible effects of cervical microbiome changes from different aspects. However, due to the limited number of statistical samples in the research process, we could not perform significant statistics for some more refined HPV genotypes. In addition, the lack of clinical information about patients (such as smoking, eating and other behaviours that may cause cervical cancer) also interfered with the experiment to a certain extent. However, the data analysis of this experiment still provides a sufficient factual basis and data support for clinical examination. Meanwhile, the microbial changes of single-infection and multi-infection case samples and the differences in metabolic functions under different HPV infection conditions were compared from a new perspective.

Conclusion

Overall, the characteristics of cervical samples microbiota were explored in this study. *Escherichia coli, Enterococcus faecalis,* and *Atopobium vaginae* in species level, *Staphylococcus, Lactobacillus* (excluding *Lactobacillus iners AB1*), *Gardnerella, Bosea, Streptococcus,* and *Sneathia* in genus level were found as high abundant bacteria in studied samples. Microbiota composition was related to HPV infection status and age, which further influenced the diversity. Specific bacteria were identified with significantly different abundance between groups. For instance, compared with unique-268, *Bifidobacteriaceae* impacted more on the multi-268 group. Moreover, some low abundance bacteria also play a vital role in specific HPV infections, such as *Saccharimonadales*, *Saccharimonadia,* and *Patescibacteria* in multi-268, *Burkholderiaceae Mycoplasma,* and *Pseudomonas aeruginosa* in unique-268. Besides, the different composition of microbiota also affected the disparities of function pathways to the metabolism of the cell, protein and genetic information between HPV infection and HPV-negative groups, and the metabolism of lipid, signal transduction and cell cycle between multi-268 infection and HPV-negative groups. In summary, our study descriptively explored the microbiota characteristics of cervical samples from Sanmenxia area patients. The analysis of single infections was not developed due to the sample size. The research concerns specific single infections, and CIN could be further investigated into their microbiota in future works.

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/, PRJNA795603.

Ethics statement

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

Author contributions

JH contributed to the manuscript writing, revision, and partial statistical data analysis. YW contributed to the initial draft discussing, editing and the general ideas of the manuscript, and collection of literature. LQ contributed to all the sample collection, ethic letter application, and results discussion. JL contributed to the part of the data analysis and data subscribing. WY did the experiments of 16S rDNA sequencing. GT contributed to the conceptualization of the study, study fund support, and results discussion. BM contributed to the conceptualization of the study, data generation and analysis, and draft revision. All authors read and revised the manuscript.

Funding

This work was supported by Geneis (Beijing) Co., Ltd.

Acknowledgments

We would like to acknowledge Department of Gynecology, Sanmenxia Central Hospital of Henan University of Science and Technology, Sanmenxia, Henan, China.

Conflict of interest

Author LQ was employed by Sanmenxia Central Hospital. Authors WY, JL, GT, and BM were employed by Geneis (Beijing) 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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.1004664/full#supplementary-material

References

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

Anahtar, M. N., Byrne, E. H., Doherty, K. E., Bowman, B. A., Yamamoto, H. S., Soumillon, M., et al. (2015). Cervicovaginal bacteria are a major modulator of host inflammatory responses in the female genital tract. *Immunity* 42, 965–976. doi: 10.1016/j.immuni.2015.04.019

Biernat Sudolska, M., Szostek, S., Rojek Zakrzewska, D., Klimek, M., and Kosz Vnenchak, M. (2011). Concomitant infections with human papillomavirus and various mycoplasma and ureaplsasma species in women with abnormal cervical cytology. *Adv. Med. Sci.* 56, 299–303. doi: 10.2478/v10039-011-0028-9

Borgogna, J., Shardell, M., Santori, E., Nelson, T., Rath, J., Glover, E., et al. (2020). The vaginal metabolome and microbiota of cervical HPV-positive and HPV-negative women: A cross-sectional analysis. *BJOG* 127, 182–192. doi: 10.1111/1471-0528.15981

Brenner, D. J., Boone, D. R., Garrity, G. M., Goodfellow, M., Krieg, N. R., Rainey, F. A., et al. (2005). *Bergey's Manual of Systematic Bacteriology, 2nd Edition, Vol. 2 (The Proteobacteria), part C (The Alpha-, Beta-, Delta-, and Epsilonproteobacteria)*. New York, NY: Springer.

Brotman, R. M., Shardell, M. D., Gajer, P., Tracy, J. K., Zenilman, J. M., Ravel, J., et al. (2014). Interplay between the temporal dynamics of the vaginal microbiota and human papillomavirus detection. *J. Infect. Dis.* 210, 1723–1733. doi: 10.1093/infdis/jiu330

Curty, G., Costa, R. L., Siqueira, J. D., Meyrelles, A. I., Machado, E. S., Soares, E. A., et al. (2017). Analysis of the cervical microbiome and potential biomarkers from postpartum HIV-positive women displaying cervical intraepithelial lesions. *Sci. Rep.* 7:17364. doi: 10.1038/s41598-017-17351-9

Di Paola, M., Sani, C., Clemente, A. M., Iossa, A., Perissi, E., Castronovo, G., et al. (2017). Characterization of cervico-vaginal microbiota in women developing persistent high-risk Human Papillomavirus infection. *Sci. Rep.* 7:10200. doi: 10.1038/s41598-017-09842-6

Fethers, K., Twin, J., Fairley, C. K., Fowkes, F. J., Garland, S. M., Fehler, G., et al. (2012). Bacterial vaginosis (BV) candidate bacteria: Associations with BV and behavioural practices in sexually-experienced and inexperienced women. *PLoS One* 7:e30633. doi: 10.1371/journal.pone.0030633

Ghaniabadi, R., Hashemi, S., Bajgiran, M. S., Javadi, S., Mohammadzadeh, N., and Masjedian, F. (2020). Distribution of Lactobacillus species in Iranian women with both human papillomavirus (HPV) infection and bacterial vaginosis (BV). *Meta Gene* 26:100791. doi: 10.1016/j.mgene.2020.100791

Gillet, E., Meys, J. F., Verstraelen, H., Bosire, C., De Sutter, P., Temmerman, M., et al. (2011). Bacterial vaginosis is associated with uterine cervical human papillomavirus infection: A meta-analysis. *BMC Infect. Dis.* 11:10. doi: 10.1186/1471-2334-11-10

Gosmann, C., Anahtar, M. N., Handley, S. A., Farcasanu, M., Abu-Ali, G., Bowman, B. A., et al. (2017). Lactobacillus-deficient cervicovaginal bacterial communities are associated with increased HIV acquisition in young South African women. *Immunity* 46, 29–37. doi: 10.1016/j.immuni.2016.12.013

Harwich, M. D., Alves, J. M., Buck, G. A., Strauss, J. F., Patterson, J. L., Oki, A. T., et al. (2010). Drawing the line between commensal and pathogenic Gardnerella vaginalis through genome analysis and virulence studies. *BMC Genom.* 11:375. doi: 10.1186/1471-2164-11-375

Harwich, M. D., Serrano, M. G., Fettweis, J. M., Alves, J. M., Reimers, M. A., Buck, G. A., et al. (2012). Genomic sequence analysis and characterisation of Sneathia amnii sp. nov. *BMC Genom.* 13:S4. doi: 10.1186/1471-2164-13-S8-S4

Herrmann, M., Wegner, C.-E., Taubert, M., Geesink, P., Lehmann, K., Yan, L., et al. (2019). Predominance of Cand. Patescibacteria in groundwater is caused by their preferential mobilisation from soils and flourishing under oligotrophic conditions. *Front. Microbiol.* 10:1407. doi: 10.3389/fmicb.2019.01407

Hosokawa, S., Kuroda, K., Narihiro, T., Aoi, Y., Ozaki, N., Ohashi, A., et al. (2021). Cometabolism of the Superphylum Patescibacteria with Anammox Bacteria in a Long-Term Freshwater Anammox Column Reactor. *Water* 13:208.

Kacerovsky, M., Vrbacky, F., Kutova, R., Pliskova, L., Andrys, C., Musilova, I., et al. (2015). Cervical microbiota in women with preterm prelabor rupture of membranes. *PLoS One* 10:e0126884. doi: 10.1371/journal.pone.0126884

Khan, S., Vancuren, S. J., and Hill, J. E. (2020). A generalist lifestyle allows rare Gardnerella spp. to persist at low levels in the vaginal microbiome. *Microb. Ecol.* 82, 1048–1060. doi: 10.1007/s00248-020-01643-1

Kunene, V., and Mahlangu, J. (2017). "Access to Systemic Anticancer Treatment and Radiotherapy Services in Sub-Saharan Africa," in *Cancer in Sub-Saharan Africa*, ed. O. Adedeji (Berlin: Springer), 175–190.

Larsson, P., Platz-Christensen, J., and Sundström, E. (1991). Is bacterial vaginosis a sexually transmitted disease? *Int. J. STD AIDS* 2, 362–364.

Lee, J. E., Lee, S., Lee, H., Song, Y.-M., Lee, K., Han, M. J., et al. (2013). Association of the vaginal microbiota with human papillomavirus infection in a Korean twin cohort. *PLoS One* 8:e63514. doi: 10.1371/journal.pone.0063514

Lemos, L. N., Medeiros, J. D., Dini-Andreote, F., Fernandes, G. R., Varani, A. M., Oliveira, G., et al. (2019). Genomic signatures and co-occurrence patterns of the

ultra-small Saccharimonadia (*phylum CPR/Patescibacteria*) suggest a symbiotic lifestyle. *Mol. Ecol.* 28, 4259–4271. doi: 10.1111/mec.15208

Libby, E. K., Pascal, K. E., Mordechai, E., Adelson, M. E., and Trama, J. P. (2008). Atopobium vaginae triggers an innate immune response in an *in vitro* model of bacterial vaginosis. *Microbes Infect.* 10, 439–446. doi: 10.1016/j.micinf.2008.01.004

Liu, S., Guo, Y., Li, B., Zhang, H., Zhang, R., and Zheng, S. (2020). Analysis of Clinicopathological Features of Cervical Mucinous Adenocarcinoma with a Solitary Ovarian Metastatic Mass as the First Manifestation. *Cancer Manag. Res.* 12:8965. doi: 10.2147/CMAR.S270675

Mason, L., Eagar, A., Patel, P., Blackwood, C., and Deforest, J. (2021). Potential microbial bioindicators of phosphorus mining in a temperate deciduous forest. *J. Appl. Microbiol.* 130, 109–122. doi: 10.1111/jam.14761

Mitra, A., Macintyre, D. A., Marchesi, J. R., Lee, Y. S., Bennett, P. R., and Kyrgiou, M. (2016). The vaginal microbiota, human papillomavirus infection and cervical intraepithelial neoplasia: What do we know and where are we going next? *Microbiome* 4:58. doi: 10.1186/s40168-016-0203-0

Mitra, A., Macintyre, D. A., Paraskevaidi, M., Moscicki, A.-B., Mahajan, V., Smith, A., et al. (2020). The Vaginal Microbiota and Innate Immunity After Local Excisional Treatment for Cervical Intraepithelial Neoplasia. *Genome Med. Actions* 13:176. doi: 10.1186/s13073-021-00977-w

Mitra, A., Macintyre, D., Lee, Y., Smith, A., Marchesi, J. R., Lehne, B., et al. (2015). Cervical intraepithelial neoplasia disease progression is associated with increased vaginal microbiome diversity. *Sci. Rep.* 5:16865.

Oh, H., Kim, B.-S., Seo, S.-S., Kong, J.-S., Lee, J.-K., Park, S.-Y., et al. (2015). The association of uterine cervical microbiota with an increased risk for cervical intraepithelial neoplasia in Korea. *Clin. Microbiol. Infect.* 21: 674.e1–9.

Onderdonk, A. B., Delaney, M. L., and Fichorova, R. N. (2016). The human microbiome during bacterial vaginosis. *Clin. Microbiol. Rev.* 29, 223–238.

Polatti, F. (2012). Bacterial vaginosis, Atopobium vaginae and nifuratel. *Curr. Clin. Pharmacol.* 7, 36–40.

Singh, S., Zhou, Q., Yu, Y., Xu, X., Huang, X., Zhao, J., et al. (2015). Distribution of HPV genotypes in Shanghai women. *Int. J. Clin. Exp. Pathol.* 8:11901.

Tian, R., Ning, D., He, Z., Zhang, P., Spencer, S. J., Gao, S., et al. (2020). Small and mighty: Adaptation of superphylum Patescibacteria to groundwater environment drives their genome simplicity. *Microbiome* 8:51. doi: 10.1186/s40168-020-00825-w

Wang, G., Jin, Z., Wang, X., George, T. S., Feng, G., and Zhang, L. (2022). Simulated root exudates stimulate the abundance of Saccharimon to improve the alkaline phosphatase activity in maise rhizosphere. *Appl. Soil Ecol.* 170:104274. doi: 10.1016/j.apsoil.2021.104274

Wei, Z.-T., Chen, H.-L., Wang, C.-F., Yang, G.-L., Han, S.-M., and Zhang, S.-L. (2021). Depiction of vaginal microbiota in women with high-risk human papillomavirus infection. *Front. Public Health* 8:587298. doi: 10.3389/fpubh.2020.587298

Werner, J., Decarlo, C. A., Escott, N., Zehbe, I., and Ulanova, M. (2012). Expression of integrins and Toll-like receptors in cervical cancer: Effect of infectious agents. *Innate Immun.* 18, 55–69. doi: 10.1177/1753425910392934

Ye, H., Song, T., Zeng, X., Li, L., Hou, M., and Xi, M. (2018). Association between genital mycoplasmas infection and human papillomavirus infection, abnormal cervical cytopathology, and cervical cancer: A systematic review and meta-analysis. *Arch. Gynecol. Obstet.* 297, 1377–1387. doi: 10.1007/s00404-018-4733-5

Zhang, Y.-Y., Xu, X.-Q., Zhang, D., Wu, J., and Zhang, H.-X. (2020). Triage human papillomavirus testing for cytology-based cervical screening in women of different ages in primary hospitals: A retrospective clinical study. *Medicine* 99:e22320. doi: 10.1097/MD.0000000000022320

Zhang, Z., Li, T., Zhang, D., Zong, X., Bai, H., Bi, H., et al. (2021). Distinction between vaginal and cervical microbiota in high-risk human papilloma virus-infected women in China. *BMC Microbiol.* 21:90. doi: 10.1186/s12866-021-02152-y

Zhou, F.-Y., Zhou, Q., Zhu, Z.-Y., Hua, K.-Q., Chen, L.-M., and Ding, J.-X. (2020). Types and viral load of human papillomavirus, and vaginal microbiota in vaginal intraepithelial neoplasia: A cross-sectional study. *Ann. Transl Med.* 8:1408. doi: 10.21037/atm-20-622

Zhou, L., Wang, P., Huang, S., Li, Z., Gong, H., Huang, W., et al. (2021). Environmental filtering dominates bacterioplankton community assembly in a highly urbanised estuarine ecosystem. *Environ. Res.* 196:110934. doi: 10.1016/j.envres.2021.110934

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 03 November 2022
DOI 10.3389/fmicb.2022.1032623

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Guohua Huang,
Shaoyang University, China
L. V. Kebo,
Ocean University of China, China

*CORRESPONDENCE
Hua Fan
fanhua@medmail.com.cn

†These authors have contributed equally to this work

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

RECEIVED 31 August 2022
ACCEPTED 17 October 2022
PUBLISHED 03 November 2022

CITATION
Li S, Yang M, Ji L and Fan H (2022) A multi-omics machine learning framework in predicting the recurrence and metastasis of patients with pancreatic adenocarcinoma. *Front. Microbiol.* 13:1032623. doi: 10.3389/fmicb.2022.1032623

A multi-omics machine learning framework in predicting the recurrence and metastasis of patients with pancreatic adenocarcinoma

Shenming Li[1,2†], Min Yang[3,4†], Lei Ji[4] and Hua Fan[1*]

[1]Department of Hepatobiliary and Pancreaticosplenic Surgery, Beijing Chaoyang Hospital, Capital Medical University, Beijing, China, [2]Department of Nephrology, Essen University Hospital, University of Duisburg-Essen, Essen, Germany, [3]School of Electrical and Information Engineering, Anhui University of Technology, Ma'anshan, Anhui, China, [4]Genesis Beijing Co., Ltd., Beijing, China

Local recurrence and distant metastasis are the main causes of death in patients with pancreatic adenocarcinoma (PDAC). Microbial content in PDAC metastasis is still not well-characterized. Here, the tissue microbiome was comprehensively compared between metastatic and non-metastatic PDAC patients. We found that the pancreatic tissue microbiome of metastatic patients was significantly different from that of non-metastatic patients. Further, 10 potential bacterial biomarkers (*Kurthia*, *Gulbenkiania*, *Acetobacterium* and *Planctomyces* etc.) were identified by differential analysis. Meanwhile, significant differences in expression patterns across multiple omics (lncRNA, miRNA, and mRNA) of PDAC patients were found. The highest accuracy was achieved when these 10 bacterial biomarkers were used as features to predict recurrence or metastasis in PDAC patients, with an AUC of 0.815. Finally, the recurrence and metastasis in PDAC patients were associated with reduced survival and this association was potentially driven by the 10 biomarkers we identified. Our studies highlight the association between the tissue microbiome and recurrence or metastasis of pancreatic adenocarcioma patients, as well as the survival of patients.

KEYWORDS

pancreatic adenocarcinoma, multi-omics, microbial community, random forest, local recurrence, distant metastasis

Introduction

Pancreatic adenocarcinoma (PDAC) remains one of the most lethal malignancies, owing in part to its early onset of metastasis (Roe et al., 2017). Most PDAC patients have metastasized at the time of diagnosis, when there is minimal benefit from surgical or chemotherapy interventions (Ryan et al., 2014; Liu et al., 2021). Consequently,

only 5% of PDAC patients survive more than 5 years after diagnosis because of its unpredictability (Chen, 2015). Improving the dismal prognosis requires a better understanding of the mechanisms of PDAC metastasis, especially the identification of metastasis biomarkers.

The microbiota inhabiting the human body is estimated to be between 10 and 100 trillion (Costello et al., 2009). While most microorganisms reside in the gastrointestinal tracts, microbiota can be found in other organs and tissues (Li et al., 2021). They play an important role in maintaining body homeostasis, and dysbiosis of the microbiota may contribute to the pathogenesis of many diseases (Liang et al., 2018). Growing researches have suggested that microbial communities influence the occurrence, progression, and response to therapy of pancreatic adenocarcioma and other cancers (Fan et al., 2018; Riquelme et al., 2019; Yang M. et al., 2022). For example, studies have shown that cancerous pancreas has significantly richer microbiota compared to normal pancreas (Pushalkar et al., 2018). Recently, Riquelme et al. (2019) found that interaction between pancreatic adenocarcioma microbiome composition and gut microbiome affects host immune responses. Besides, studies have shown that oral antibiotic depletion of gut microbiota in mice suppresses tumor growth and metastasis while activating antitumor immunity in the tumor environment (Wei et al., 2019; Liu J. et al., 2022). However, the potential association between microbial communities of cancer tissue and pancreatic adenocarcioma metastasis remains a knowledge gap.

The occurrence and development of pancreatic adenocarcioma are affected by multiple factors. Previous studies have revealed that the development of pancreatic adenocarcioma is accompanied by changes in the expression patterns of large set of mRNAs (He et al., 2022) and non-coding RNAs, such as lncRNAs and miRNAs (Xiao et al., 2018; Wang et al., 2019; Xu et al., 2020b; Zhang et al., 2020). LncRNA PSMB8-AS1 contributes to pancreatic adenocarcioma progression *via* modulating miR-382-3p/STAT1/PD-L1 axis (Zhang et al., 2020). LncRNA DANCR promotes proliferation and metastasis in pancreatic adenocarcioma by regulating miRNA-33b (Luo et al., 2020). Wang et al. (2020) reported that the upregulation of METTL14 led to the decrease of PERP levels *via* m^6A modification, promoting the growth and metastasis of pancreatic adenocarcioma. Sohrabi et al. (2021) found that 6 out of 43 common miRNAs (hsa-miR-210, hsa-miR-375, hsa-miR-216a, hsa-miR-217, hsa-miR-216b, and hsa-miR-634) had significant differences in their expression profiles between the tumor and normal groups of pancreatic adenocarcioma. However, comparative studies on the accuracy of different omics in predicting recurrence and metastasis in pancreatic adenocarcioma patients are still vacant.

In this study, 37 samples of patients with recurrence or metastasis (RM) and 42 samples of patients without recurrence or metastasis (no-RM) were collected, and the tissue microbiome of all patients with pancreatic adenocarcioma were characterized. The main objectives of this study were: (1) to identify the bacterial biomarkers capable of discriminating between RM and non-RM, (2) to compare the differences in transcriptome levels between RM and no-RM patients, and (3) to compare the performance of microbes and mRNAs in predicting pancreatic adenocarcioma recurrence or metastasis. Our study sheds light on the ability of tissue microbial biomarkers of pancreatic adenocarcioma to predict recurrence or metastasis.

Materials and methods

Sampling populations and datasets

Microbiome data and transcriptome data were obtained from the Cancer Genome Atlas (TCGA) database.[1] The microbiome data of pancreatic adenocarcioma patient tissues were derived from the re-cleaning of the sequencing data of samples from the TCGA database by Rob Knight's team (Poore et al., 2020). The microbial RNA data of pancreatic adenocarcioma patients were selected and the clinical data of pancreatic adenocarcioma in TCGA were downloaded. The samples were divided into two groups according to whether the patients had recurrence or metastasis within 1 year after the initial diagnosis. Patients with recurrence or metastasis or both within 1 year were defined as RM, and those without recurrence or metastasis were defined as no-RM. In total, we matched 79 samples, including 37 RM and 42 no-RM. We also collected some essential clinical indicators of the patients, such as age, gender, and disease stage, etc.

Statistical analysis

Statistical analysis was performed using R language. Wilcoxon rank sum test was used to determine the relationship between different clinical features and patients' recurrence and metastasis. If the *p*-value between the two groups is less than 0.05, it is considered that there is a statistically significant difference. At the same time, by constraining the *p*-value to be less than 0.01, the microbial characteristics with significant differences were screened as potential microbial markers for downstream analysis.

1 https://portal.gdc.cancer.gov

Identification of differentially expressed genes

Differentially expressed genes (DEGs) of mRNA, lncRNA, and miRNA were identified using the "Deseq2" R package. Up-regulated genes were obtained by adjusted p-value < 0.1 and log2 Fold Change > 0. Down-regulated genes were obtained by adjusted p-value < 0.1 and log2 Fold Change < 0. Then, genes with significant differences were screened by $|\log2\ (\text{Fold Change})| \geq 1$ and adjusted p-value less than 0.05. Significantly different genes were displayed by the "pheatmap" package in R. Gene Ontology (GO) enrichment analysis was conducted by the "clusterProfiler" package in R. Enrichment pathways of DEGs were displayed by the "ggplot2" package in R.

Diversity analysis

Alpha-diversity (Richness, Chao, Shannon, and Simpson indices) were calculated using the "vegan" package in R. Principal coordinate analysis (PCoA) was conducted with the "vegan" package in R to analyze differences between microbial communities. Wilcoxon rank sum test was used for two group comparisons of microbial diversity. P-value less than 0.05 was considered statistically different.

Machine learning classification model

To evaluate the performance of different omics in predicting the recurrence and metastasis of patients with PDAC, we labeled the RM patients as "0" and the no-RM patients as "1," which turned our research into a binary classification of machine learning. Random Forest (RF) model in Python's Sklearn module was used for classification. RF randomly samples all the original data, generates n different sample datasets, builds a decision tree model for each dataset, and finally obtains the prediction result of the final model according to the voting results of each decision tree model. We estimated the performance of the classification algorithms using the fivefold cross-validation (fivefold-cv). The performance of the classification algorithm was calculated by averaging the AUC (area under curve) in the five test datasets. Finally, metrics including AUC, ACC (accuracy), precision, recall, and F1-score were used to comprehensively evaluate the performance of the model.

Survival prediction

Ten bacterial biomarkers were identified using Wilcoxon rank sun test. Then, these 10 biomarkers were used to predict the survival of patients with PDAC. The survival curve was conducted using the Kaplan–Meier (KM) method and log-rank test was used to compare the difference of survival probability. The analysis and visualization were conducted with the "survival" package in R.

TABLE 1 Clinical information.

Parameters	RM ($n = 42$)	no-RM ($n = 37$)	P-value
Gender (M/F)	22/20	23/14	NS
Age (avg years)	66.79	61.97	NS
N0/N1/unknown	9/33/0	16/18/3	*
M0/M1/MX	16/0/26	23/2/12	*
T1/T2/T3/T4/unknown	2/3/36/1/0	3/7/24/1/2	NS
Stage I and Stage II/Stage III and Stage IV/unknown	40/2/0	34/2/2	NS

Tumor Node Metastasis classification (TNM): T stage refers to the situation of the primary tumor focus. With the increase of tumor volume, the depth of invasion and the range of adjacent tissue involvement, it is expressed by T1–T4 in turn. N stage refers to the regional lymph node involvement, which is represented by N0 when the lymph node is not involved. With the increase of the degree and scope of lymph node involvement, it is indicated by N1–N2 in turn. M stage means M refers to distant metastasis, with M0 for those without distant metastasis and M1 for those with distant metastasis; RM, recurrence or metastasis; no-RM, without recurrence and metastasis; NS, no significant differences. *Indicated p-value < 0.05.

Results

Tumor node metastasis classification stages are significantly correlated with recurrence and metastasis of pancreatic adenocarcioma

The correlations between clinical phenotype and recurrence and metastasis of pancreatic adenocarcioma patients were shown in Table 1. Both N stage and M stage have significant differences between RM and no-RM (Figure 1). Specifically, PDAC patients with advanced disease (N1) had a significantly increased probability of recurrence or metastasis. That means PDAC patients with recurrence or metastasis were accompanied by increased lymph node involvement. Besides, there were no significant differences in gender and age between RM and no-RM patients. The demographics and clinical characteristics are provided in Table 1.

Bacterial profiles of pancreatic tissue differ between recurrence or metastasis and no-recurrence or metastasis patients

Previous microbial studies of pancreatic adenocarcioma have shown that bacterial composition shifted compared to non-diseased pancreatic (Pushalkar et al., 2018). Here, we intend

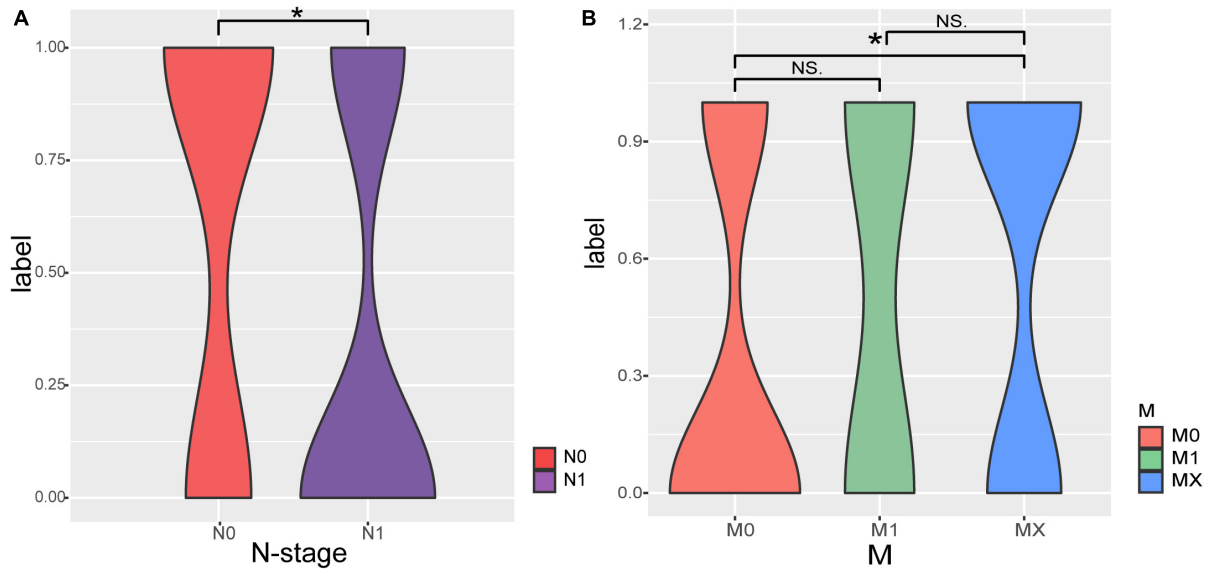


FIGURE 1
The correlations of Tumor Node Metastasis classification (TNM) stage with recurrence and metastasis. **(A)** Patients with recurrence or metastasis are accompanied by increased lymph node involvement. **(B)** Comparisons of M staging in patients with RM and no-RM; Wilcoxon test is used to compare between different groups of samples. The *X*-axis represents the different stages of patients; *Y*-axis represents the recurrence and metastasis, 0: recurrence or metastasis; 1: without recurrence and metastasis.

to examine these compositional changes in distant metastatic PDAC. As shown in **Figure 2**, *Pseudomonas* dominated the tissue microbiome of pancreatic adenocarcioma with an average relative abundance of 12.8%, followed by *Staphylococcus* (7.3%) and *Bacillus* (6.9%) (**Figure 2A**). Further, there was no difference in alpha-diversity (richness, Chao, Shannon, and Sipmson) between RM and no-RM (**Figure 2B**). PCoA plot also showed no significant differences in bacterial communities between RM and no-RM (Bray–Curtis $P = 0.172$; **Figure 2C**). These data indicated similar global community alpha-diversity and beta-diversity between the RM and no-RM patients. Riquelme et al. (2019) found higher alpha-diversity in the tumor microbiome of long-term survival (LTS) patients compared with short-term survival (STS) patients. Our results demonstrate that metastasis in PDAC patients does not alter the overall tissue microbial community structure.

Next, we identified 10 potential biomarkers capable of distinguishing between RM and no-RM (**Figure 3**). The relative abundance of *Kurthia*, *Gulbenkiania*, *Acetobacterium*, *Planctomyces*, *Xenophilus*, *Gardnerella*, *Advenella*, *Catenuloplanes*, *Leptolyngbya*, and *Proteus* was significantly different between RM and no-RM ($P < 0.01$). Among them, the relative abundance of most bacterial biomarkers decreased in patients who developed recurrence or metastasis. Only the relative abundance of *Acetobacterium*, *Catenuloplanes*, and *Leptolyngbya* increased in the RM patients. The results demonstrated that decreased relative abundance of key bacteria in PDAC patient tissues may be a contributing factor to recurrence or metastasis. Although the overall microbial communities of RM and no-RM appear to be similar, recurrence and metastasis are still accompanied by increased or decreased relative abundance of some specific bacteria. These abundance-changing bacteria may be used as important indicators for clinical prediction of recurrence and metastasis of PDAC patients, so in-depth research such as experimental verification is urgently needed to reveal the underlying functional mechanisms of these bacteria.

Transcriptome expression in pancreatic adenocarcioma patients carries information on recurrence or metastasis

Interactions and complex regulatory mechanisms among lncRNA, miRNA, and mRNA play key roles in the occurrence and development of multiple diseases (Huang, 2018; Liao et al., 2019; Cheng et al., 2020; Ma et al., 2020). In this study, instead of considering the regulation among lncRNA, miRNA, and mRNA, we analyzed the differences in these three transcriptomes between RM and no-RM separately. We performed a comprehensive analysis of the differential expression of each omics between the RM and no-RM. For lncRNA, 402 up-regulated and 288 down-regulated genes were identified. For miRNA, we identified 107 up-regulated and 44 down-regulated genes, while for mRNA, 3,074 up-regulated and 1,539 down-regulated genes were identified. After adjusting

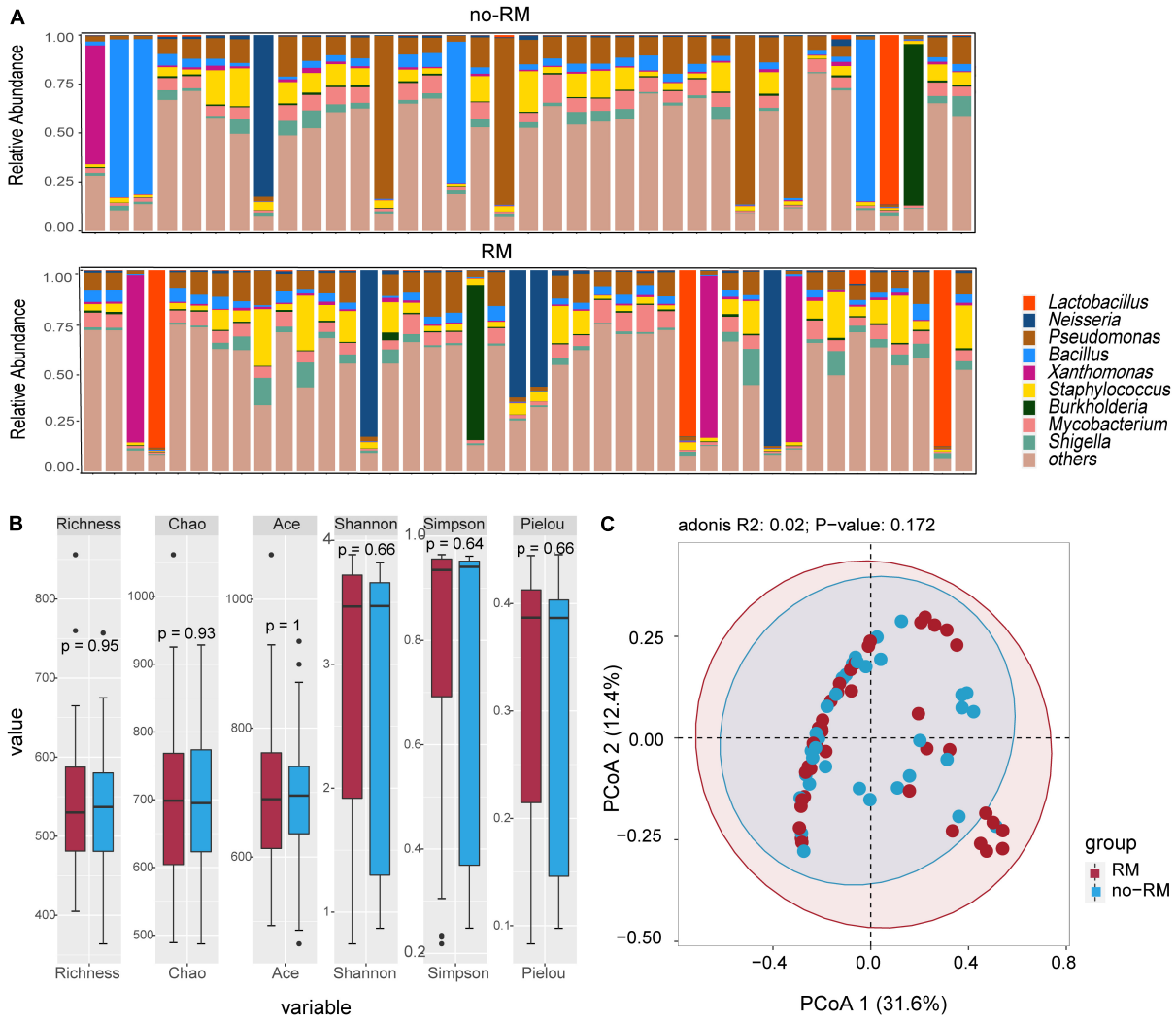


FIGURE 2
Difference in microbial composition and diversity between the two groups. **(A)** The top 10 genus levels in two groups; the stacked bar chart showed the composition of patient genus level in two groups of recurrence or metastasis. **(B)** Comparison of alpha-diversity of two groups based on different indexes. **(C)** Comparison of beta-diversity of two groups with PCoA. Wilcoxon test was used to detect variation between different groups based on the microbial composition at the genus level. Richness, Chao, and Ace index represent the richness of the microbial species; Shannon, Simpson, and Pielou index represent the diversity of the microbial species; RM, recurrence and metastasis; no-RM, no-recurrence and metastasis.

for the *P*-value, we obtained 309 significantly differentially expressed lncRNAs, 62 significantly differentially expressed miRNAs, and 1,287 significantly differentially expressed mRNAs (details in **Supplementary Tables 1–3**). Heatmap showed the differences in the expression levels of the top 40 lncRNAs, miRNAs, and mRNAs between RM and no-RM (**Figures 4A–C**).

Further, we explored the biological function of these significantly differentially genes (**Figure 4D**). For GO terms, all GO terms can be classified into three categories: (1) Biological process (BP), (2) Cellular component (CC), and (3) Molecular function (MF). First, for biological process, most of the BP terms have been confirmed to be related to the signal release and modulation of chemical synaptic transmission. Second, for cellular component, most of the CC terms can be clustered into synaptic membrane and transporter complex. Finally, as for molecular function, MF terms mostly contributed to the passive transmembrane transporter activity and channel activity. Furthermore, potential pathological pathways in PDAC metastasis were further analyzed with Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations (**Figure 4E**). The results showed that the differentially expressed genes were mainly enriched in neuroactive ligand-receptor interaction, cAMP signaling pathway, and adrenergic signaling in cardiomyocytes, and other signaling pathways.

Chen et al. (2020) reported that for lower-grade glioma (LGG) and normal tissues, neuroactive ligand-receptor

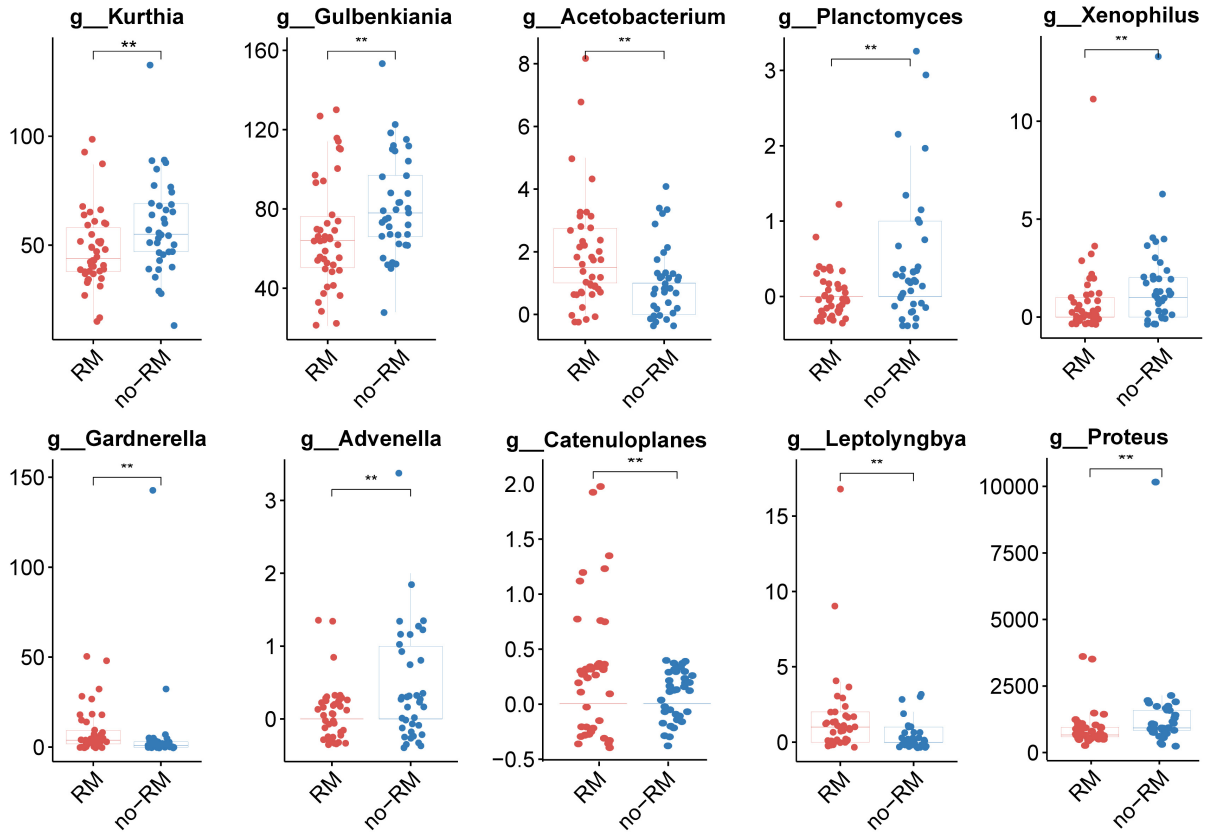


FIGURE 3
Ten potential biomarkers capable of distinguishing between RM and no-RM. Wilcoxon test was used to detect variation between different groups based on the relative abundance of tissue microbes, When the *p*-value was less than 0.01, 10 potential genus level microbial markers were identified; the boxplot was used to show the differences between the two groups; RM, recurrence or metastasis; no-RM, without recurrence and metastasis.

interaction was identified as differentially enriched pathway in KEGG. Also, in our study, a possible key pathway in RM patients with PDAC is neuroactive ligand-receptor interaction (**Figure 4E**). Different disease subjects share certain enriched pathways, which have also been reported in other studies (Priya et al., 2022). We strongly recommend further research on this topic to progressively improve the transcriptomic evidence on PDAC metastasis.

Microbes are the best predictors of the recurrence and metastasis in patients with pancreatic adenocarcioma

Predicting the recurrence or metastasis of pancreatic adenocarcioma patients plays a huge role in improving patient survival and reducing medical costs. Therefore, we further evaluated the performance of different omics in predicting the recurrence and metastasis in patients with pancreatic adenocarcioma (**Figure 5A**). Firstly, based on all the characteristics of each omics, RF fivefold cross-validation showed that lncRNA obtained the highest accuracy in predicting the recurrence and metastasis of pancreatic adenocarcioma patients (AUC = 0.791). However, when the 10 identified bacterial biomarkers were used as features, the prediction performance was the best with an AUC of 0.815. Besides AUC, other metrics (ACC, precision, recall, and F1-score) were also used to evaluate the predictive effect of each omics (**Figure 5B**). The results also showed that the 10 bacterial biomarkers performed best, which further indicate that the 10 bacteria may serve as potential biomarkers of recurrence and metastasis of PDAC.

Recurrence and metastasis in pancreatic adenocarcioma patients are associated with reduced survival

Many studies have shown that microbes are closely related to the survival of cancer patients (Chattopadhyay et al., 2019; Peters et al., 2019; Riquelme et al., 2019). Besides, in this study, we found that the tissue microbiome significantly influenced the

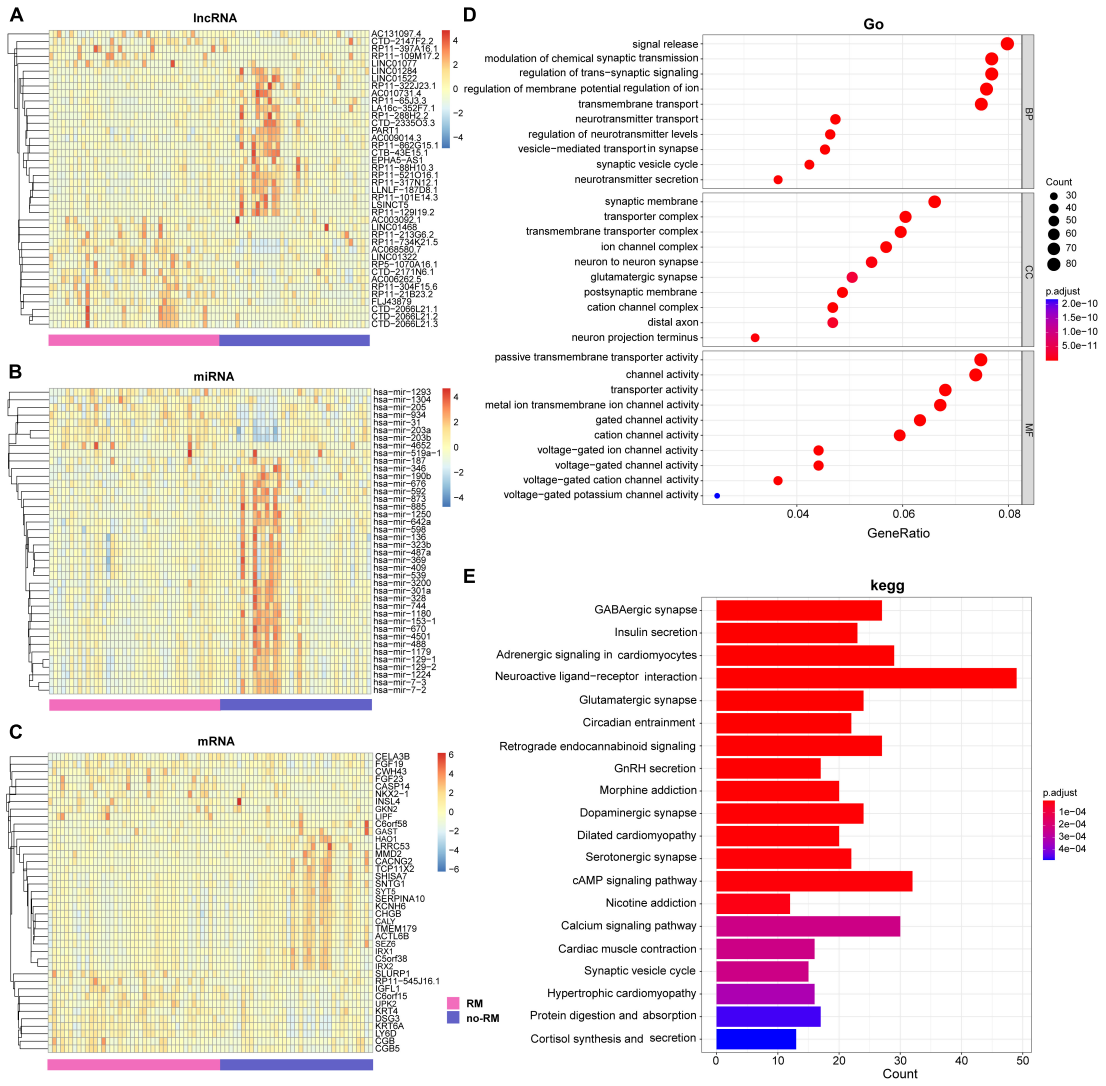


FIGURE 4
Difference analysis and enrichment analysis of different omics between two groups. The heat map of the DEGs of **(A)** lncRNA; **(B)** miRNA; **(C)** mRNA between the RM and no-RM group, the *x*-axis is the sample of two groups, and the *y*-axis is the top 40 expressions with significant differences screened by DEseq2. **(D)** GO analysis of DEGs between RM and no-RM. **(E)** KEGG analysis of the DEGs between RM and no-RM, the *X*-axis is the ratio of differentially expressed genes enriched in the corresponding pathway, and the *Y*-axis is the name of the pathway; BP, biological process category; CC, cellular component category; MF, molecular function category; RM, recurrence or metastasis; no-RM, without recurrence and metastasis.

recurrence and metastasis of patients with PDAC. Therefore, we wonder whether recurrence and metastasis in patients are associated with survival and whether this association is driven by tissue microbes.

Based on these 10 bacterial biomarkers, all patients were classified to two clusters with machine learning classification model used previously. Then, survival analysis was conducted on the predicted clusters (Figure 6). First, survival time of RM patients were significantly shorter than those of no-RM patients ($P < 0.0001$; Figure 6A). Meanwhile, similar result was found when we conducted survival analysis on the two predicted clusters, that is, there was a significant difference in survival between the two clusters ($P = 0.0059$; Figure 6B). Our results demonstrate that the recurrence and metastasis in pancreatic patients are associated with reduced survival and this association is potentially driven by key tissue microbes.

Discussion

The recurrence and metastasis have become a critical problem in cancer diagnosis, treatment, and metastasis (He et al., 2020; Shi et al., 2022). Through a comprehensive comparison of tissue microbes in non-metastatic and metastatic

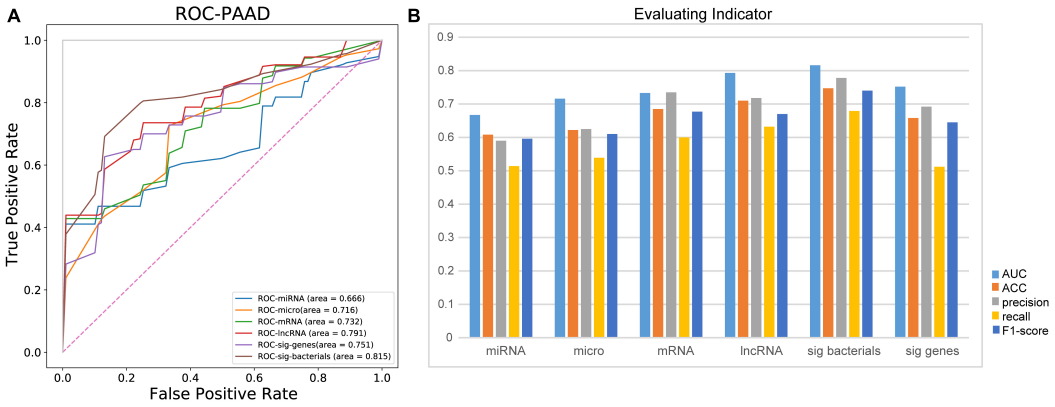


FIGURE 5
Ten identified bacterial biomarkers perform best in predicting recurrence and metastasis in patients with PDAC. **(A)** Comparison of AUC in patients with recurrence and metastasis predicted by different omics. **(B)** Evaluation of predictive ability of different evaluation indices for recurrence and metastasis of PDAC patients; micro, microbiome; sig bacteria, 10 identified bacterial biomarkers; sig genes, identified DEGs from mRNA data; AUC, area under curve; ACC, accuracy.

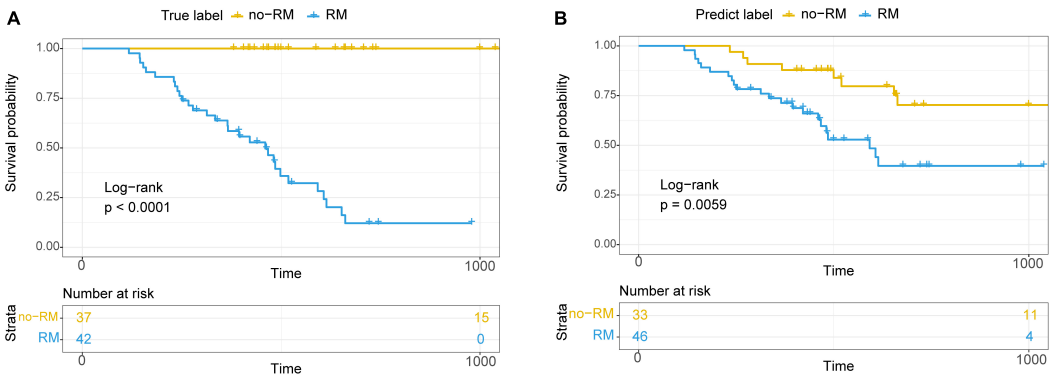


FIGURE 6
Kaplan–Meier survival curve showed significantly different overall survival between RM and no-RM. **(A)** Relationship between true recurrence and metastasis labels and overall survival of patients. **(B)** Relationship between recurrence and metastasis labels predicted by the model and the overall survival of patients; RM, recurrence or metastasis; no-RM, without recurrence and metastasis.

pancreatic adenocarcioma patients, we identified 10 bacteria that differentiate between RM and no-RM patients. Among them, the relative abundance of most bacterial biomarkers decreased in patients who developed recurrence or metastasis. Although there were significant differences in the expression patterns of multiple omics between RM and no-RM patients, the accuracy of these 10 bacteria in predicting recurrence and metastasis in pancreatic adenocarcioma patients was higher than that of other omics (lncRNA, miRNA, and mRNA). More importantly, these bacterial biomarkers potentially drive the association between metastasis and patient survival.

Beyond the simple description of tissue microbiome changes in pancreatic adenocarcioma patients, our study proposes the idea of microbe-based predictors for metastasis of PDAC. Groundbreaking, we identified 10 potential bacterial biomarkers. The microbe composition comparing normal esophagus with intestinal metaplasia, low grade dysplasia, high grade dysplasia, and adenocarcinoma showed significant decreases in the phylum Planctomycetes and the genus *Planctomyces* in diseased tissue compared with healthy controls and intrasample controls (Peter et al., 2020). We find that the relative abundance of *Planctomyces* in RM patients is significantly lower than that in no-RM patients. Bacterial dysbacteriosis, characterized by a predominance of *Gardnerella vaginalis* may accelerate the process of cervical carcinogenesis (Kovachev, 2020). Similarly, we also find an increased relative abundance of the genus *Gardnerella* in patients with recurrence or metastasis.

We find that the microbe-based predictor is more accurate compared with lncRNA, miRNA, and mRNA, possibly due to the tissue microbes play a dominant role in recurrence and metastasis of PDAC. The 10 bacterial biomarkers we identified could be used to clinically assist in the diagnosis of early stage pancreatic adenocarcioma patients for future

recurrence and metastasis. Consequently, the medical costs and patient suffering will be greatly reduced. However, the detailed link between the tissue microbes and the pathological mechanism of pancreatic metastasis remains to be further clarified. It is of note that besides microbe and molecular biomarkers, histopathological images have been adopted to evaluate recurrence and metastasis risk for many cancers (Liu X. et al., 2022; Yang J. et al., 2022; Ye et al., 2022). Feasible directions to improve prediction accuracy include exploring more advanced machine learning models used in other related biological problems (Xu et al., 2020a; Meng et al., 2022) and integrating more types of prediction data.

The strength of our study includes two accurately divided pancreatic adenocarcioma cohorts with and without recurrence or metastasis within 1 year, the microbiome data at the site of initial cancer, and detailed follow-up information for all involved patients. Several limitations to the present study exist. First, the small sample size may make the findings less generalizable. Although we comprehensively compared the tissue microbiome of RM and no-RM patients, the absence of healthy controls is not conducive to underpinning the findings. In addition, we used the public data of TCGA database, which needs to be verified by the clinical data of the Chinese population. At the same time, the image information of patients was likely to be added to the framework of predicting recurrence and metastasis, and further model fusion will help to improve the prediction accuracy. Functional experiments are needed in the future to deeply explore the physiological mechanism of tissue microbes affecting the recurrence and metastasis of PDAC. Complete and organized experiments will help unravel pancreatic adenocarcioma metastases and aid clinicians in diagnosis.

Conclusion

In conclusion, we characterize the system alterations of tissue microbiome in pancreatic adenocarcioma patients. We uncover the microbial signature associated with recurrence and metastasis of pancreatic adenocarcioma and develop a highly accurate microbe-based predictor for recurrence and metastasis diagnosis of PDAC.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov.

Author contributions

HF contributed to conception and design of the study. SL and MY organized the database. SL, HF, and MY performed the statistical analysis. SL and LJ wrote the first draft of the manuscript. HF and LJ wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Conflict of interest

Authors MY and LJ were employed by Genesis Beijing 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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.1032623/full#supplementary-material

References

Chattopadhyay, I., Verma, M., and Panda, M. (2019). Role of oral microbiome signatures in diagnosis and prognosis of oral cancer. *Technol. Cancer Res. Treat.* 18:1533033819867354. doi: 10.1177/1533033819867354

Chen, J., Wang, Z., Wang, W., Ren, S., Xue, J., Zhong, L., et al. (2020). SYT16 is a prognostic biomarker and correlated with immune infiltrates in glioma: A study based on TCGA data. *Int. Immunopharmacol.* 84:106490. doi: 10.1016/j.intimp.2020.106490

Chen, W. (2015). Cancer statistics: Updated cancer burden in China. *Chin. J. Cancer Res.* 1:27.

Cheng, Y., Su, Y., Wang, S., Liu, Y., Jin, L., Wan, Q., et al. (2020). Identification of circRNA-lncRNA-miRNA-mRNA competitive endogenous RNA network as novel prognostic markers for acute myeloid leukemia. *Genes* 11:868. doi: 10.3390/genes11080868

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

Fan, X., Alekseyenko, A. V., Wu, J., Peters, B. A., Jacobs, E. J., Gapstur, S. M., et al. (2018). Human oral microbiome and prospective risk for pancreatic cancer: A population-based nested case-control study. *Gut* 67, 120–127. doi: 10.1136/gutjnl-2016-312580

He, B., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., et al. (2020). TOOme: A novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression. *Front. Bioeng. Biotechnol.* 8:394. doi: 10.3389/fbioe.2020.00394

He, B., Wang, K., Xiang, J., Bing, P., Tang, M., Tian, G., et al. (2022). DGHNE: Network enhancement-based method in identifying disease-causing genes through a heterogeneous biomedical network. *Brief. Bioinform.* bbac405. [Epub ahead of print]. doi: 10.1093/bib/bbac405

Huang, Y. (2018). The novel regulatory role of lncRNA-miRNA-mRNA axis in cardiovascular diseases. *J. Cell. Mol. Med.* 22, 5768–5775. doi: 10.1111/jcmm.13866

Kovachev, S. M. (2020). Cervical cancer and vaginal microbiota changes. *Arch. Microbiol.* 202, 323–327. doi: 10.1007/s00203-019-01747-4

Li, J. J., Zhu, M., Kashyap, P. C., Chia, N., Tran, N. H., Mcwilliams, R. R., et al. (2021). The role of microbiome in pancreatic cancer. *Cancer Metastasis Rev.* 40, 777–789. doi: 10.1007/s10555-021-09982-2

Liang, D., Leung, R. K., Guan, W., and Au, W. W. (2018). Involvement of gut microbiome in human health and disease: Brief overview, knowledge gaps and research opportunities. *Gut Pathog.* 3:10. doi: 10.1186/s13099-018-0230-4

Liao, J., Wang, J., Liu, Y., Li, J., and Duan, L. (2019). Transcriptome sequencing of lncRNA, miRNA, mRNA and interaction network constructing in coronary heart disease. *BMC Med. Genomics* 12:124. doi: 10.1186/s12920-019-0570-z

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. *Front. Cell Dev. Biol.* 9:619330. doi: 10.3389/fcell.2021.619330

Liu, J., Lan, Y., Tian, G., and Yang, J. (2022). A systematic framework for identifying prognostic genes in the tumor microenvironment of colon cancer. *Front. Oncol.* 12:899156. doi: 10.3389/fonc.2022.899156

Liu, X., Yuan, P., Li, R., Zhang, D., An, J., Ju, J., et al. (2022). Predicting breast cancer recurrence and metastasis risk by integrating color and texture features of histopathological images and machine learning technologies. *Comput. Biol. Med.* 146:105569. doi: 10.1016/j.compbiomed.2022.105569

Luo, Y., Wang, Q., Teng, L., Zhang, J., Song, J., Bo, W., et al. (2020). LncRNA DANCR promotes proliferation and metastasis in pancreatic cancer by regulating miRNA-33b. *FEBS Open Biol.* 10, 18–27. doi: 10.1002/2211-5463.12732

Ma, N., Tie, C., Yu, B., Zhang, W., and Wan, J. (2020). Identifying lncRNA-miRNA-mRNA networks to investigate Alzheimer's disease pathogenesis and therapy strategy. *Aging* 12, 2897–2920. doi: 10.18632/aging.102785

Meng, Y., Lu, C., Jin, M., Xu, J., Zeng, X., and Yang, J. (2022). A weighted bilinear neural collaborative filtering approach for drug repositioning. *Brief Bioinform.* 23:bbab581. doi: 10.1093/bib/bbab581

Peter, S., Pendergraft, A., Vanderpol, W., Wilcox, C. M., Kyanam Kabir, Baig, K. R., et al. (2020). Mucosa-associated microbiota in barrett's esophagus, dysplasia, and esophageal adenocarcinoma differ similarly compared with healthy controls. *Clin. Transl. Gastroenterol.* 11:e00199. doi: 10.14309/ctg.0000000000000199

Peters, B. A., Hayes, R. B., Goparaju, C., Reid, C., Pass, H. I., and Ahn, J. (2019). The Microbiome in lung cancer tissue and recurrence-free survival. *Cancer Epidemiol. Biomarkers Prev.* 28, 731–740. doi: 10.1158/1055-9965.EPI-18-0966

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

Priya, S., Burns, M. B., Ward, T., Mars, R. A. T., Adamowicz, B., Lock, E. F., et al. (2022). Identification of shared and disease-specific host gene-microbiome associations across human diseases using multi-omic integration. *Nat. Microbiol.* 7, 780–795. doi: 10.1038/s41564-022-01121-z

Pushalkar, S., Hundeyin, M., Daley, D., Zambirinis, C. P., Kurz, E., Mishra, A., et al. (2018). The pancreatic cancer microbiome promotes oncogenesis by induction of innate and adaptive immune suppression. *Cancer Discov.* 8, 403–416. doi: 10.1158/2159-8290.CD-17-1134

Riquelme, E., Zhang, Y., Zhang, L., Montiel, M., Zoltan, M., Dong, W., et al. (2019). Tumor Microbiome diversity and composition influence pancreatic cancer outcomes. *Cell* 178, 795–806 e712. doi: 10.1016/j.cell.2019.07.008

Roe, J. S., Hwang, C. I., Somerville, T. D. D., Milazzo, J. P., Lee, E. J., Da Silva, B., et al. (2017). Enhancer reprogramming promotes pancreatic cancer metastasis. *Cell* 170, 875–888 e820. doi: 10.1016/j.cell.2017.07.007

Ryan, D. P., Hong, T. S., and Bardeesy, N. (2014). Pancreatic adenocarcinoma. *N. Engl. J. Med.* 371, 2140–2141. doi: 10.1056/NEJMra1404198

Shi, X., Young, S., Cai, K., Yang, J., and Morahan, G. (2022). Cancer susceptibility genes: Update and systematic perspectives. *Innovation* 3:100277. doi: 10.1016/j.xinn.2022.100277

Sohrabi, E., Rezaie, E., Heiat, M., and Sefidi-Heris, Y. (2021). An integrated data analysis of mrna, mirna and signaling pathways in pancreatic cancer. *Biochem. Genet.* 59, 1326–1358. doi: 10.1007/s10528-021-10062-x

Wang, M., Liu, J., Zhao, Y., He, R., Xu, X., Guo, X., et al. (2020). Upregulation of METTL14 mediates the elevation of PERP mRNA N(6) adenosine methylation promoting the growth and metastasis of pancreatic cancer. *Mol. Cancer* 19:130. doi: 10.1186/s12943-020-01249-8

Wang, W., Lou, W., Ding, B., Yang, B., Lu, H., Kong, Q., et al. (2019). A novel mRNA-miRNA-lncRNA competing endogenous RNA triple sub-network associated with prognosis of pancreatic cancer. *Aging* 11, 2610–2627. doi: 10.18632/aging.101933

Wei, M. Y., Shi, S., Liang, C., Meng, Q. C., Hua, J., Zhang, Y. Y., et al. (2019). The microbiota and microbiome in pancreatic cancer: More influential than expected. *Mol. Cancer* 18:97. doi: 10.1186/s12943-019-1008-0

Xiao, X., Zhu, W., Liao, B., Xu, J., Gu, C., Ji, B., et al. (2018). BPLLDA: Predicting lncRNA-disease associations based on simple paths with limited lengths in a heterogeneous network. *Front. Genet.* 9:411. doi: 10.3389/fgene.2018.00411

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020a). CMF-Impute: An accurate imputation tool for single-cell RNA-seq data. *Bioinformatics* 36, 3139–3147. doi: 10.1093/bioinformatics/btaa109

Xu, J., Zhu, W., Cai, L., Liao, B., Meng, Y., Xiang, J., et al. (2020b). LRMCMDA: Predicting miRNA-disease association by integrating low-rank matrix completion with miRNA and disease similarity information. *IEEE Access* 8, 80728–80738. doi: 10.1109/ACCESS.2020.2990533

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., et al. (2022). Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information *via* multimodal deep learning. *Comput. Struct. Biotechnol. J.* 20, 333–342. doi: 10.1016/j.csbj.2021.12.028

Yang, M., Yang, H., Ji, L., Hu, X., Tian, G., Wang, B., et al. (2022). A multi-omics machine learning framework in predicting the survival of colorectal cancer patients. *Comput. Biol. Med.* 146:105516. doi: 10.1016/j.compbiomed.2022.105516

Ye, Z., Zhang, Y., Liang, Y., Lang, J., Zhang, X., Zang, G., et al. (2022). Cervical cancer metastasis and recurrence risk prediction based on deep convolutional neural network. *Curr. Bioinform.* 17, 164–173. doi: 10.2174/1574893616666210708143556

Zhang, H., Zhu, C., He, Z., Chen, S., Li, L., and Sun, C. (2020). LncRNA PSMB8-AS1 contributes to pancreatic cancer progression *via* modulating miR-382-3p/STAT1/PD-L1 axis. *J. Exp. Clin. Cancer Res.* 39:179. doi: 10.1186/s13046-020-01687-8

frontiers | Frontiers in Microbiology

TYPE Methods
PUBLISHED 10 November 2022
DOI 10.3389/fmicb.2022.990870

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology,
China

REVIEWED BY
Minsuk Kim,
Cedars Sinai Medical Center,
United States
William Evan Johnson,
Boston University,
United States

*CORRESPONDENCE
Taesung Park
tspark@stats.snu.ac.kr

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

RECEIVED 10 July 2022
ACCEPTED 11 October 2022
PUBLISHED 10 November 2022

CITATION
Ham H and Park T (2022) Combining *p*-values from various statistical methods for microbiome data.
Front. Microbiol. 13:990870.
doi: 10.3389/fmicb.2022.990870

Combining *p*-values from various statistical methods for microbiome data

Hyeonjung Ham[1] and Taesung Park[1,2]*

[1]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, South Korea,
[2]Departement of Statistics, Seoul National University, Seoul, South Korea

Motivation: In the field of microbiome analysis, there exist various statistical methods that have been developed for identifying differentially expressed features, that account for the overdispersion and the high sparsity of microbiome data. However, due to the differences in statistical models or test formulations, it is quite often to have inconsistent significance results across statistical methods, that makes it difficult to determine the importance of microbiome taxa. Thus, it is practically important to have the integration of the result from all statistical methods to determine the importance of microbiome taxa. A standard meta-analysis is a powerful tool for integrative analysis and it provides a summary measure by combining *p*-values from various statistical methods. While there are many meta-analyses available, it is not easy to choose the best meta-analysis that is the most suitable for microbiome data.

Results: In this study, we investigated which meta-analysis method most adequately represents the importance of microbiome taxa. We considered Fisher's method, minimum value of *p* method, Simes method, Stouffer's method, Kost method, and Cauchy combination test. Through simulation studies, we showed that Cauchy combination test provides the best combined value of *p* in the sense that it performed the best among the examined methods while controlling the type 1 error rates. Furthermore, it produced high rank similarity with the true ranks. Through the real data application of colorectal cancer microbiome data, we demonstrated that the most highly ranked microbiome taxa by Cauchy combination test have been reported to be associated with colorectal cancer.

KEYWORDS

microbiome analysis, integration method, *p*-value combination, power simulation, rank simulation

Introduction

Since the roles of the microbiome in human body sites and their importance arise, there have been many studies focusing on revealing differentially expressed microbiome taxa in a variety of cancer types and diseases (Hayes et al., 2018; Osman et al., 2018; Qian et al., 2018; Dong et al., 2019; Ramsheh et al., 2021). In the meanwhile, there are certain common

characteristics among microbiome datasets that make analyses difficult: overdispersion and high sparsity (presence of zero counts; Sohn and Li, 2018; Xia et al., 2018). To account for these characteristics, many statistical methods have been developed. DESeq2 and edgeR are widely used methods to find differentially expressed features in the field of RNA-Seq data analysis, and account for overdispersion of the dataset using a negative binomial distribution modeling strategy (Robinson et al., 2010; Love et al., 2014). MetagenomeSeq was developed to account for sparsity using a distinct normalization method, known as cumulative sum scaling (CSS) and using a zero-inflated model (Paulson et al., 2013). ZIBseq and ZINB are methods that account for the sparsity through incorporating zero-inflated beta model and zero-inflated negative binomial model, respectively (Peng et al., 2016; Xia et al., 2018). There also are methods that use centered log-ratio (CLR) transformation to account for the compositional nature of relative abundance data in analysis (Gloor et al., 2017).

Microbiome analysis methods are broadly classified into two classes: taxa-level method and community-level method (Plantinga et al., 2017). Taxa-level method performs analyses in terms of each taxon, and includes aforementioned methods. The community-level method accounts for phylogenetic distances between representative sequences. MiRKAT, the microbiome regression-based kernel association test, uses kernels that incorporate microbiome-wise similarity matrix that can be calculated from various distances (Zhao et al., 2015). MiSPU, the microbiome-based sum of powered score, uses the idea of the sum of powered score (SPU) to be applied to microbiome datasets through variable weighting of representative sequences (Wu et al., 2016). OMiAT, optimal microbiome-based association test, is an approach that integrates SPU and MiRKAT by taking the minimum value of p from the two methods (Koh et al., 2017). TMAT, the phylogenetic tree-based microbiome association test, uses log-transformed read count per million (CPM) and tests whether an internal node of a phylogenetic tree is associated with the outcome, using the phylogenetic tree structure (Kim K. J. et al., 2020). All the methods introduced above are used to find the differentially expressed (DE) features. There have been studies that attempted a comprehensive review of these statistical methods (Xia and Sun, 2017; Pollock et al., 2018; Nearing et al., 2022). However, it is not easy to tell which is the best method among the individual DE method because each method is specialized for the specific characteristics of microbiome data. Furthermore, the significance results provided from different statistical methods tend to be inconsistent. In other words, a DE feature from one method does not necessarily be a DE feature from the other method (Khomich et al., 2021). Thus, several studies summarized the inconsistent results obtained from different statistical methods by using a Venn diagram that represented commonly significant features under a certain significance level (Chen et al., 2015; You et al., 2018; Nazarieh et al., 2019; Wang et al., 2019; Kim S. I. et al., 2020). In addition to the significance, the ranking of DE features is also inconsistent between the methods.

In this study, we combine the value of ps from different statistical methods to determine the importance of DE features. Rather than focusing on an individual method, our focus lies in combining different test results from a set of multiple methods. There exist many methods for combining value of ps, depending on whether value of ps are independent (Fisher, minimum value of p, Simes, Stouffer) or correlated (Kost, Cauchy). The most common method is Fisher's method that uses a chi-square distribution to calculate the combined value of p (Fisher, 1925). The method using the minimum value of p can also be taken to maximize the power (Tippett, 1931; Casella and Berger, 2017). Simes method for combining value of p is similar to the minimum value of p method, but uses ordered value of ps to determine the significance (Simes, 1986). Stouffer's method takes the inverse standard normal cumulative distribution function (CDF) of value of ps so that the statistic follows a normal distribution (Stouffer and Suchman, 1949). Kost method accounts for the correlation between p-values by modifying the chi-square distribution of the Fisher's method (Kost and McDermott, 2002). Cauchy combination test accounts for the correlation between p-values by using Cauchy distribution, which makes the distributional changes in the tail limited in the existence of p-value correlation (Liu and Xie, 2020). The combined p-values were then used to rank the importance of microbiome.

In this study, we investigate the most appropriate p-value combination method in the analysis of microbiome dataset in terms of significance testing and ranking DE features. Simulation settings were designed to assess: (i) the type 1 error and power of differentially expressed feature discovery, (ii) rank similarity between the true ranks and ranks determined by combined p-values.

In our empirical studies, we only considered the genus level. Many differential abundance analyses have been conducted only at the genus level, due to the limitation in microbiome annotation and not enough high resolution provided by 16 s rRNA sequence to classify species. Popular microbiome databases, including Silva, and Greengenes databases, recommend not to use the annotation at the species level (Ritari et al., 2015; Dueholm et al., 2020). Although databases such as NCBI and EzBioCloud EzTaxon provide more accurate annotations than Silva and Greengenes at the species level (Kim et al., 2012; Schoch et al., 2020), uncultured and unidentified species still exist and are often filtered out in the differential abundance analyses. Additionally, the microbiome resolution provided by 16 s rRNA is limited because the length of highly variable region is short for accurately classifying species except for few species. Therefore, analysis was conducted in the genus level at this study.

Materials and methods

Microbiome datasets

Baxter's colorectal cancer data

Stool samples obtained through the Great Lakes-New England Early Detection Research Network were used in this study (Baxter et al., 2016). Raw sequencing data and metadata are available at

TABLE 1 Null hypotheses of statistical methods.

Category	Method	Null hypothesis (H0)	Detail
Taxa-level	DESeq2	$\beta_i = 0$	β =LFC (Log-fold change) For i^{th} taxa
	edgeR	$\lambda_1 - \lambda_2 = 0$	For group 1, group 2
	Wilcoxon CLR	$\lambda_1^{median} - \lambda_2^{median} = 0$	For group 1, group 2
	ZIBSeq	$\beta_i = 0$	For i^{th} taxa
	MetagenomeSeq	$\beta_i = 0$	For i^{th} taxa
	ZINB	$\beta_i = 0$	For i^{th} taxa
Community-level	oMiRKAT	$\tau = 0$	Kernel regression Random effect $f(Z) \sim (0, \tau K)$ For kernel K $f(z_i) = \sum_{j=1}^{p} z_{ij}\beta_j$, if the model is linear for p OTUs
	aMiSPU	$\beta = (\beta_1, \beta_2, \ldots, \beta_p)' = 0$	For p OTUs
	aSPU	$\beta = (\beta_1, \beta_2, \ldots, \beta_p)' = 0$	For p OTUs
	TMAT	$\beta = (\beta_1, \beta_2, \ldots, \beta_{M\_1})' = 0$	For M_1 internal nodes in phylogenetic tree

NCBI Sequencing Read Archive (SRA) with the accession number SRP062005. A total of 314 samples with 187 normal and 127 colorectal cancer (CRC) were available.

Experimental procedures were previously reported as follows (Kozich et al., 2013). The V4 region of 16s rRNA gene was amplified using custom-designed primers, and sequenced using an Illumina MiSeq sequencer with paired-end sequencing. Raw FASTQ data were processed through Qiime2 pipeline from raw file processing to taxonomy assignment (https://qiime2.org/, version 2021.04). Qiime2 Cutadapt plugin was used to trim primer sequences, and representative sequences were obtained through DADA2 denoising algorithm. Taxonomies were assigned using SILVA databases (release 138) with 99% similarity. Fasttree plugin was used to generate the phylogenetic tree. After removing singletons and doublets, data comprised 4,772 representative sequences. After filtering representative sequences with <0.005% of total read count (Bokulich et al., 2013), 803 representative sequences with 80 genera were available.

Zeller's colorectal cancer data

Stool samples obtained through the European Molecular Biology Laboratory (EMBL) were used in the real analysis of this study. Raw sequencing data and metadata are available at European Nucleotide Archive (ENA) with the project number PRJEB6070. Excluding samples without the disease status information, a total of 91 samples with 50 normal and 41 CRC were available.

Experimental procedures were previously reported as follows (Zeller et al., 2014). The V4 region of 16s rRNA gene was amplified using targeted primers (F515 5'-GTGCCAGCMGCCG CGGTAA-3′, R806 5'-GGACTACHVGGGTWTCTAAT-3′), and sequenced following Illumina MiSeq platform (Illumina, San Diego, United States) at the Genomics Core Facility, EMBL, Heidelberg. Raw FASTQ data were processed through the same pipeline as the Baxter's data described above using Qiime2. After the filtering, 329 representative sequences with 81 genera were available.

Methods for identifying DE features

The methods for identifying DE features are classified into taxa-level and community level methods, as summarized in Table 1 with the corresponding null hypotheses. Taxa-level method includes DESeq2[Wald/LRT], edgeR, Wilcoxon rank sum test with CLR transformation (Wilcoxon CLR), ZIBSeq, MetagenomeSeq [Gaussian/log normal], and ZINB. Community-level method includes oMiRKAT, aMiSPU, aSPU, and TMAT. aSPU was considered instead of OMiAT, that takes the minimum value of p of SPU and MiRKAT. For this study, the value of p generated by MiRKAT was already included, so only value of p generated by SPU was considered. All analysis results were obtained at the genus level. R[1] software was used for the analyses. Unless stated, default options were used for all analysis.

Methods for integration analysis

For the value of p combination, Fisher's method, minimum value of p method (min P method), Kost method, Simes method, Stouffer's method, and Cauchy combination test were used. Details of each method are described below.

Fisher's method

It is also called Fisher's combination test. Under the null hypothesis, for independent value of ps,

$$T_{\text{Fisher}} = -\sum_{i=1}^{k} 2\log p_i \sim \chi^2_{2k}$$

for k tests to be combined, and p_i represents i^{th} value of p. Minimum value of p method (Min P method). Under the null hypothesis, for independent value of ps,

1 https://www.r-project.org/

$$T_{\min P} = \min_{i=1,2,..,k} p_i \sim \text{Beta}(1,k)$$

for k tests to be combined.

Kost method

For dependent value of *p*s, scale the chi-square distribution of Fisher's method as follows (Kost and McDermott, 2002):

$$T_{\text{Kost}} \sim c\chi^2_{2f}$$

where

$$f = \frac{E[T]^2}{\text{var}[T]}, c = \frac{\text{var}[T]}{2E[T]} = \frac{k}{f}$$

and

$$E[T] = 2k, \text{var}[T] = 4k = 2\sum_{i<j} \text{cov}(-2\log p_i, -2\log p_j)$$

Cauchy combination method

For value of *p*s under arbitrary dependency structure, defined by the weighted sum of the Cauchy transformed value of each value of *p* as follows:

$$T_{\text{Cauchy}} = \sum_{i=1}^{k} w_i \tan\{(0.5 - p_i)\pi\}$$

where w_i is nonnegative weight that satisfies $\sum_{i=1}^{k} \omega_i = 1$, and p_i is the value of p from i th test. Cauchy combination test accounts for the dependence of value of *p*s using the heaviness of the Cauchy tail (Liu and Xie, 2020). Equal weights were used in this study.

Simes method

For independent value of *p*s, let $p_1, \ldots, p_k$ be the ordered *p*-values for k tests. The null hypothesis is rejected if $p_i \le i\alpha / k$ for any $i = 1,\ldots,k$ for a significance level α. It is mainly used in multiple testing correction, but also suggested for the *p*-value combination in some studies (Cheng and Sheng, 2017; Ganju and Ma, 2017).

Stouffer method

For independent *p*-values,

$$T_{\text{stouffer}} = \frac{\sum_{i=1}^{k} \Phi^{-1}(1 - p_i)}{\sqrt{k}} \sim N(0,1)$$

where Φ represents the standard normal cumulative distribution function.

Simulation settings

Simulation setting 1

Simulation setting 1 was designed to assess type 1 error rates and power of each *p*-value combination method. The simulation datasets were generated as previously reported (Zhao et al., 2015). Microbiome datasets were simulated according to Chen and Li's approach (Chen and Li, 2013). The simulated OTU counts were generated using Dirichlet-multinomial (DM) model, that incorporates the mean OTU proportion and the overdispersion measure as the shape parameter α. The sample size was set to 300 and 20,000 total read counts were generated per sample. The OTU counts were set to have different levels of sparsity (e.g., the total proportion of zero counts) to account for the zero-inflated nature of microbiome datasets. For sparsity, sparsity parameter $\pi \in \{0.3,0.5,0.7,0.8\}$ was set. The OTU counts were simulated as follows:

$$Z_{ij} = \begin{cases} 0 & \text{with probability } \pi \\ \text{Dirichlet}-\text{multinomial}(\alpha) & \text{with probability } 1-\pi \end{cases}$$

where Z_{ij} is OTU counts for ith sample and jth OTU.

The dependent variable was generated as practiced in MiRKAT (Zhao et al., 2015). For the binary outcome variable, the outcome was simulated under the model

$$\text{logit}(E(y_i|,X_i|,Z_i)) = 0.5\text{scale}(X_{1i} + X_{2i}) + \beta\text{scale}\left(\sum_{j\in G} Z_{ij}\right)$$

where y_i represents the dependent variable of the sample i, X_i represents the covariates of sample i, scale$(\cdot)$ represents the standardization with mean 0 and standard deviation 1, β represents the degree of association and G represents the given cluster of OTUs. Here, the OTU-level datasets are simulated so that each cluster of OTUs indicates each genus. Among the statistical methods, the taxa level analysis methods used a collapsed sum of OTUs corresponding to a genus, while the community-level analysis methods used simulated OTU data as it is.

One virtual covariate X_{1i} was simulated as $\sim Bernouill(0.5)$. The other virtual covariate X_{2i} was simulated as $\sim N(0,1)$, assuming the covariate and the taxa counts Z_i were independent. β was set to have the values of $\{0,0.01,0.02,0.03,0.04,0.05,0.1,0.15,0.2\}$. Type 1 error was measured when $\beta = 0$. A total of 1,000 dependent variables were generated for each combination of β s and $\grave{A}$s to calculate the type 1 error rates and the power.

Among the DE feature analysis methods, the taxa-level analysis methods used a collapsed sum of OTUs corresponding to

a genus, while the community-level analysis methods used simulated OTU data as it is.

Simulation setting 2

Simulation setting 2 was designed to assess the rank similarity between the true rank and the rank determined using each value of p combination method. The real CRC dataset introduced in Method 2.1 was used to reflect the microbiome counts of the real data. To control the degree of association (β), the dependent variable was generated under the same model from simulation setting 1. For the same dataset, ten different dependent variables were generated by previously determined β s as in Figure 1. The larger effect size, the higher rank. For each dependent variable, 100 replications were performed.

Three scenarios were considered using different numbers of non-causal dependent variables. Non-causal dependent variables were set to have β =0, assuming microbiome features that are not related to the dependent variables. Each scenario was designed to, respectively, have 6, 4, and 8 non-causal dependent variables, and the causal dependent variables were generated to have different degrees of association (with different βs).

The rank difference was presented with two measurements: rank squared difference and Spearman correlation coefficient. The rank squared difference was measured using

$$\sum_{g=1}^{N} d_g^2$$

where $d_g = \left(\text{real } rank - \text{computed } rank\right)$ of g th genus.

Similarly, the Spearman rank correlation coefficient was used as:

$$\rho = 1 - \frac{6\sum d_g^2}{N\left(N^2 - 1\right)}$$

With each value of p combination method, both measures were applied and the results of 100 replications were compared.

Results

Simulation result

Result of simulation setting 1

In this section, the type 1 error rates and the power of individual statistical method are first shown, then that of value of p combination methods are subsequently shown.

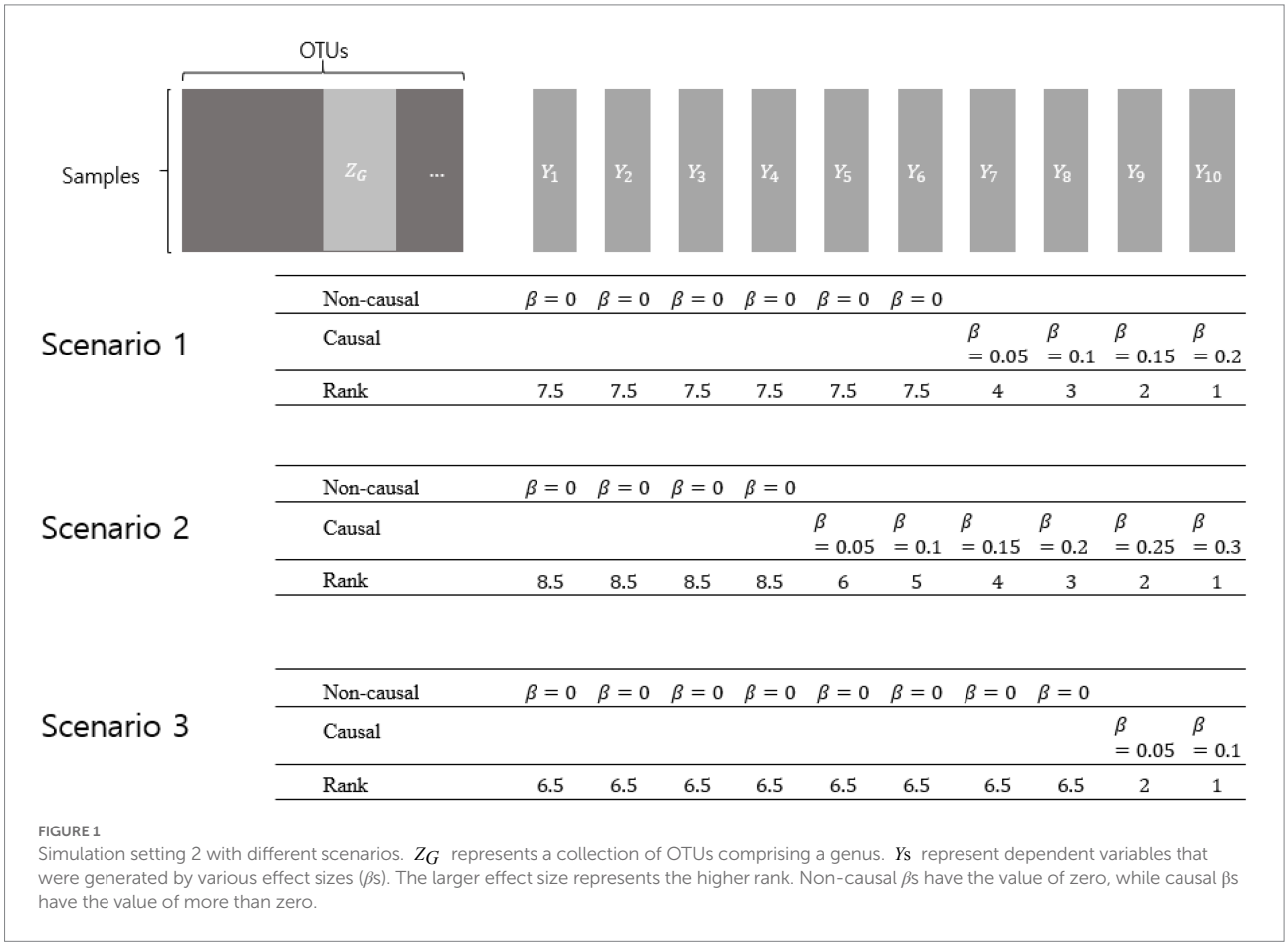


FIGURE 1
Simulation setting 2 with different scenarios. Z_G represents a collection of OTUs comprising a genus. Ys represent dependent variables that were generated by various effect sizes (βs). The larger effect size represents the higher rank. Non-causal βs have the value of zero, while causal βs have the value of more than zero.

TABLE 2 Type 1 error rates of individual statistical methods.

Sparsity	DESeq2 LRT	DESeq2 Wald	edgeR	Wilcoxon	ZIBSeq	ZIG Gaussian	ZIG Log Normal	ZINB	aSPU	oMiRKAT	aMiSPU	TMAT
0.3	0.000	0.000	0.000	0.017	0.002	0.000	0.000	0.000	0.014	0.014	0.014	0.014
0.5	0.037	0.037	0.026	0.006	0.031	**0.094**	0.000	0.043	0.042	0.042	0.042	0.042
0.7	0.000	0.000	0.000	0.013	0.000	0.009	0.000	0.000	0.000	0.000	0.000	0.000
0.8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Bold value indicates to inflated type I errors.

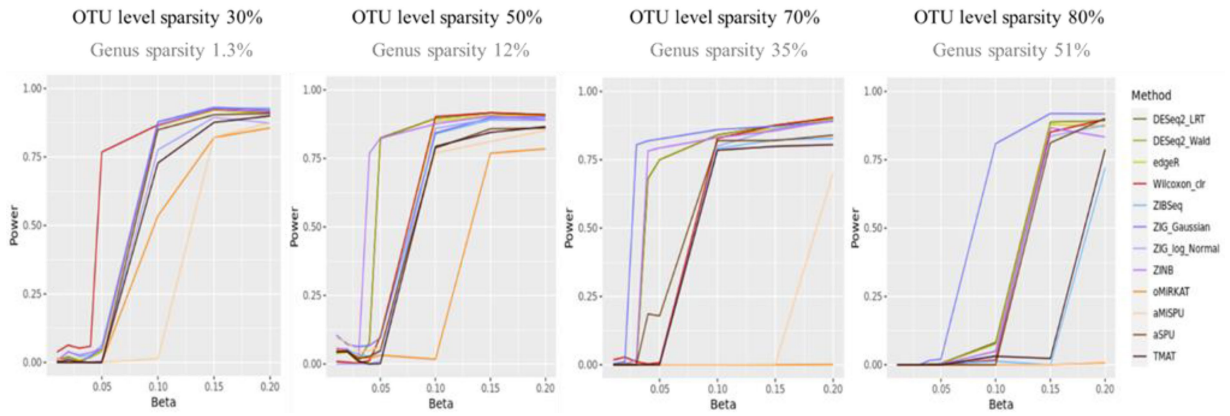


FIGURE 2
The statistical power of the individual methods. The *x*-axis represents the degree of association (β). The value of βs were given as $\{0.01,0.02,0.03,0.04,0.05,0.1,0.15,0.2\}$. The y-axis represents the power. The blueish colors represent methods that consider zero-inflation.

The type 1 error rates of individual methods are given in Table 2. Under the significance level of 0.05, the type 1 error rates of most statistical methods were well-controlled below 0.05. The type 1 error rates of ZIG Gaussian was uncontrolled in some cases, but not in ZIG log Normal. It was previously reported that the type 1 error rate of ZIG Gaussian was off the nominal range, compared to other statistical methods (Calgaro et al., 2020).

Figure 2 shows the statistical power of the individual methods in terms of the degree of association. The power tended to decrease as the level of sparsity increased, and the power of community-level analysis methods tended to be lower than the taxa-level analysis methods. The methods used in RNA-Seq data analysis showed higher performances in terms of power (ZIG, DESeq2). The Wilcoxon rank sum method showed a higher performance when the sparsity level was low (Genus sparsity 1.3%).

Table 3 represents the type 1 error rates of value of *p* combination methods. The type 1 error rates were not controlled in Fisher's method and Stouffer's method. The type 1 error rates were considered to be not controlled if the confidence interval for proportion test did not include 0.05 (i.e., for Stouffer's method with sparsity 0.3, the 95% confidence interval of [0.0694, 0.1051] did not include 0.05, for Cauchy combination test with sparsity 0.5, the 95% confidence interval of [0.0435, 0.0732] include 0.05.). The type 1 error rates of other value of p combination methods did not exceed the given significance level of 0.05 considering the confidence interval. Since the type 1 error rates of Fisher's combination method and Stouffer's methods were not controlled, we focused only on the other methods for value of *p* combination. The results for Fisher's and Stouffer's methods can be found in the Supplementary Figure 1.

TABLE 3 The type 1 error rates of p-value combination methods.

Sparsity	Fisher	MinP	Kost	Cauchy	Simes	Stouffer
0.3	0.032	0	0.005	0	0	**0.086**
0.5	**0.09**	0.029	0.045	0.057	0.032	**0.096**
0.7	0.01	0	0	0	0	0.016
0.8	0	0	0	0	0	0.005

Bold value indicates to inflated type I errors.

Figure 3 shows the statistical power of the value of *p* combination methods as the degree of association increases. Although the performances of value of *p* combination methods were similar, the power of Cauchy combination test was observed to be the best for all levels of sparsity. The performance of min *P* method was the worst. The differences in power between the methods tended to be smaller as the sparsity level becomes higher.

Result of simulation setting 2

Three scenarios were considered to evaluate the rank difference. In scenario 1, the rank squared difference was the lowest when combined with Cauchy combination test, Min P and Simes methods being next (Figure 4). Similarly, the Spearman

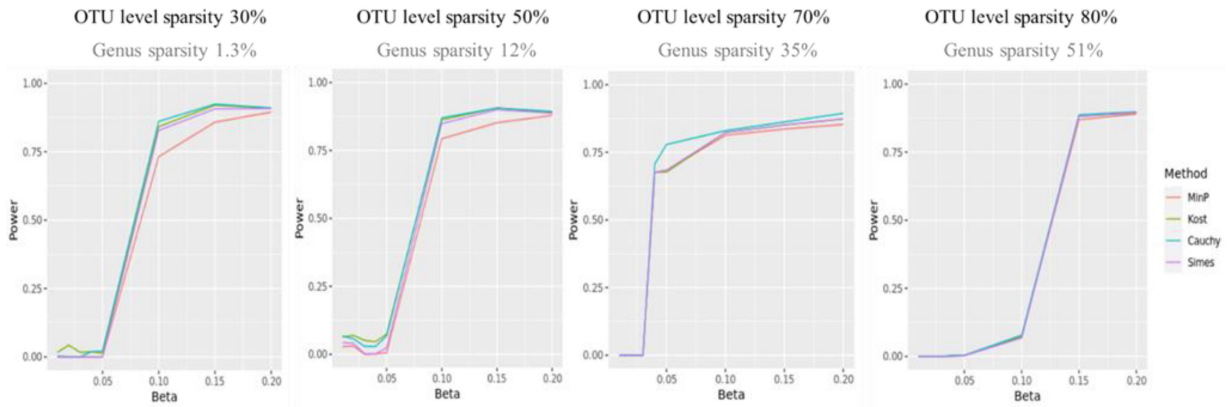


FIGURE 3
The statistical power of value of p combination methods. The x-axis represents the degree of association (β). The value of βs were given as $\{0.01,0.02,0.03,0.04,0.05,0.1,0.15,0.2\}$. The y-axis represents the power.

rank correlation was the highest for Cauchy combination test. In both measures, the paired Wilcoxon test value of p between Cauchy combination test results and others were significant (value of $p < 0.001$). Similarly, Cauchy combination test showed the lowest rank squared difference and the highest correlation coefficient in scenarios 2 and 3 (Figure 4).

Real microbiome data analysis

Baxter's colorectal cancer data analysis

The differentially expressed microbiome feature analysis was conducted for every genus in the Baxter's CRC dataset, and the importance was determined by the magnitude of value of ps generated for each genus. DE feature analyses were used as described in the Method section. The Spearman rank correlation between each pair of statistical methods was compared as in Figure 5. A Spearman rank correlation coefficient of 0.46 was observed between DESeq2 and edgeR, which are both used in RNA-Seq analysis and based on the negative binomial distribution in common. A lower spearman rank correlation coefficient was observed between edgeR and Wilcoxon rank sum test results, between ZIBSeq and others, ZIG and others, ZINB and others except for RNA-Seq analysis methods, and the community-level analysis methods (oMiRKAT, aSPU, aMISPU, and TMAT) and others. The correlation tests were significant between some pairs of methods, that means there was a linear trend between value of ps ranks generated for those methods. However, the linear trend does not assure that the pairwise p-values have the same ranks. For example, although the correlation test between edgeR and ZINB is significant with the coefficient of 0.79, and thus they have a linear trend of p-value ranks, the pairwise p-values are not aligned as DESeq2_LRT and DESeq2_Wald. Furthermore, except for the DESeq2_LRT and DESeq2_Wald, which are both derived from DESeq2, no pair of methods produced similar rank list of microbiome genera (Supplementary Figure 2).

CRC stool samples were analyzed with different statistical methods and the resulting p-values were combined using Cauchy combination test. These p-values were further adjusted for controlling the false discovery rate (FDR) as practiced (Yoon et al., 2021). Table 4 shows the top microbiome genera in the order of adjusted p-values (q-values).

The first taxon was the most significant. Although it was uncultured in both genus and family levels, *Rhodospirillales* in order level was previously identified in the dextran sulfate sodium-induced colitis group but not in the control group (Yang et al., 2017). Also, the microbiome family *Rhodospirillaceae* was increased in colitic mice and IBD patients (Burrello et al., 2018). The bacterial genus *Megasphaera* was found to be a butyrate-producer, that induces epigenetic modifications in CRC development (Tarashi et al., 2019). *Gastranaerophilales* was previously reported as correlated with the late phase of aging through gene expression profiles of C57BL/6J mice (van der Lugt et al., 2018). The genus *Cloacibacillus* was observed to be enriched in CRC patients with stage IV (Sheng et al., 2019). The bacterial species *Porphyromonas asaccharolytica* and *Porphyromonas gingivalis*, both rarely detectable in healthy individuals, were shown to be enriched in CRC patients in previous studies (Sinha et al., 2016; Okumura et al., 2021; Wang et al., 2021). Clostridia vadinBB60 group was observed to be enriched in low-graded; right-sided/transverse tumors (Zwinsová et al., 2021). The genus *Sutterella* was reported to be the most representative in the colorectal adenocarcinoma groups (Mori et al., 2018). The bacterial species *Odoribacter splanchnicus* was previously reported as a potential inducer of TH17cells and might protect against colitis and CRC in wild type mice (Xing et al., 2021; York, 2021). The abundance of *Turicibacter* was observed to be higher in the colitis or CRC group than in the groups with treatments, but the causative role of *Turicibacter* is to be further studied (Wu M. et al., 2019). The genus *Slackia* was studied to be overrepresented in CRC (Coleman and Nunes, 2016).

Most microbiome genera in Table 4 that had high ranks from Cauchy combination test had been previously reported as associated

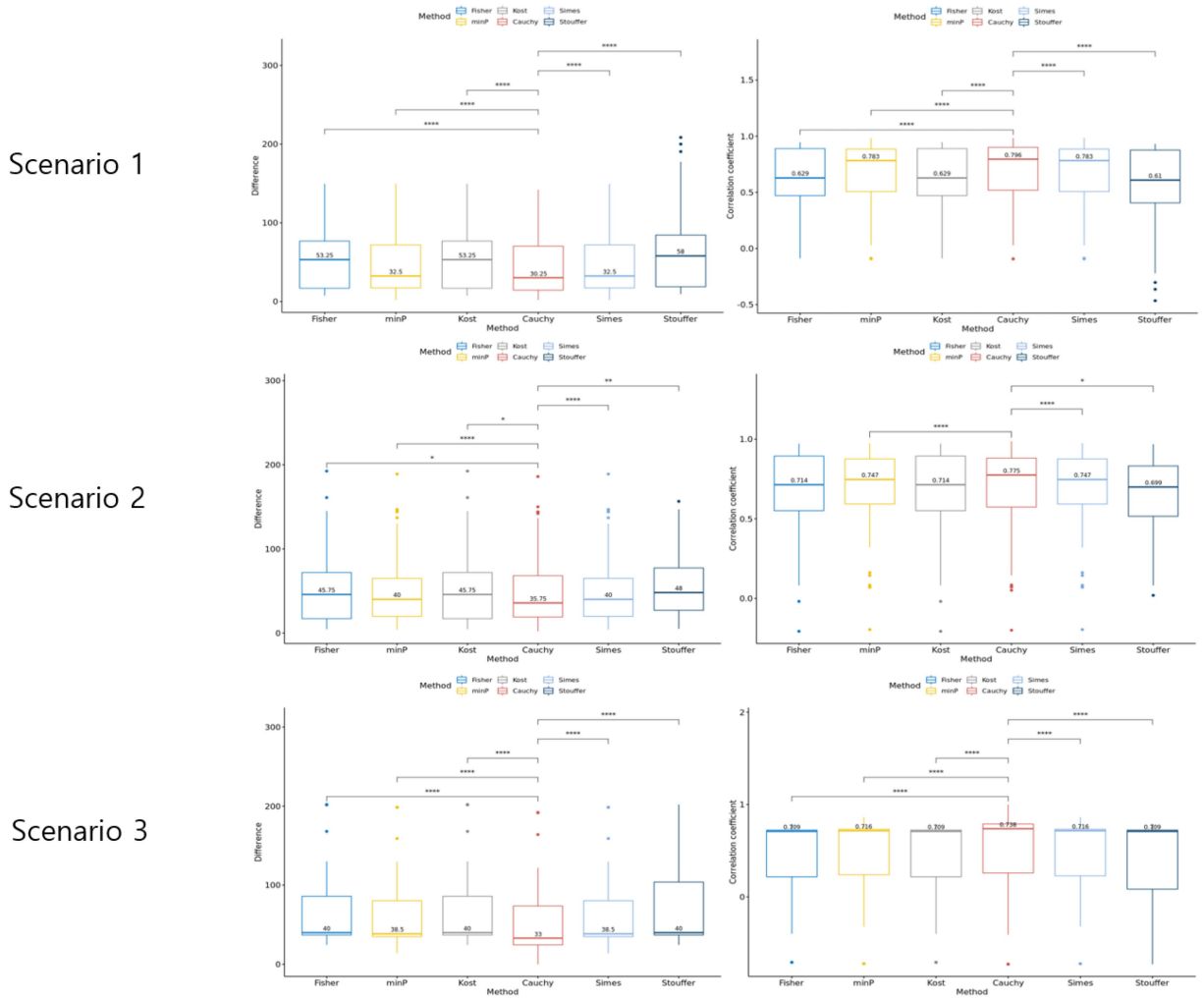


FIGURE 4
Results of the simulation setting 2. The graphs in the left column represent rank difference of each value of *p* combination method. The graphs in the right column represent the Spearman rank correlation of each *p*-value combination method.

with CRC or related symptoms. The ranks generated by min P and Simes method were similar to the Cauchy combination test, which corresponds to the results from the simulation setting 2. On the other hand, the other methods did not include some highly ranked taxa discovered from Cauchy combination test in the lists of their top 10 taxa (Supplementary Table 1). For example, Cauchy combination test ranked the genera *Sutterella* and *Odoribacter* at 7th and 8th, while Stouffer's method ranked them at 18th and 13th, respectively, despite their reported associations with CRC.

Zeller's colorectal cancer data analysis

A different CRC stool samples were analyzed with statistical methods and the resulting *p*-values were combined using Cauchy combination test. Table 5 shows the top microbiome genera in the order of q-values.

The most significant microbiome, *Porphyromonas* has been reported to be enriched in gut microbiota profiling of CRC patients in several studies (Yang et al., 2019). *Hungatella* was found to be a CRC-enriched marker, and was found to be depleted after the removal of CRC compared with newly diagnosed CRC patients (Cronin et al., 2022). Also, the species *Hungatella hathewayi WAL-18680* is a common cancer-associated biomarker (Wu et al., 2021). *Fusobacterium nucleatum* is commonly associated with CRC, and found to promote tumor development by inducing several immune responses including inflammation (Wu J. et al., 2019; Queen et al., 2022). *Rikenellaceae RC9 gut group* was suggested as a potential biomarker of CRC from gut microbiota profiles in mice (Shao et al., 2022). *Cloacibacillus* was reported to show statistical differences in the gut microbiota between CRC patients with stage III and IV (Sheng et al., 2019). *Veillonella* and a strain of *Streptococcus* together were reported to modulate inflammation, and were increased in fibrosis and cirrhosis compared to samples without cirrhosis (Jia et al., 2021). The relative abundance of *Catenibacterium* was found to be significantly different between CRC and normal patients (Yang et al., 2019). A low abundance of *Mitsuokella* in CRC patients

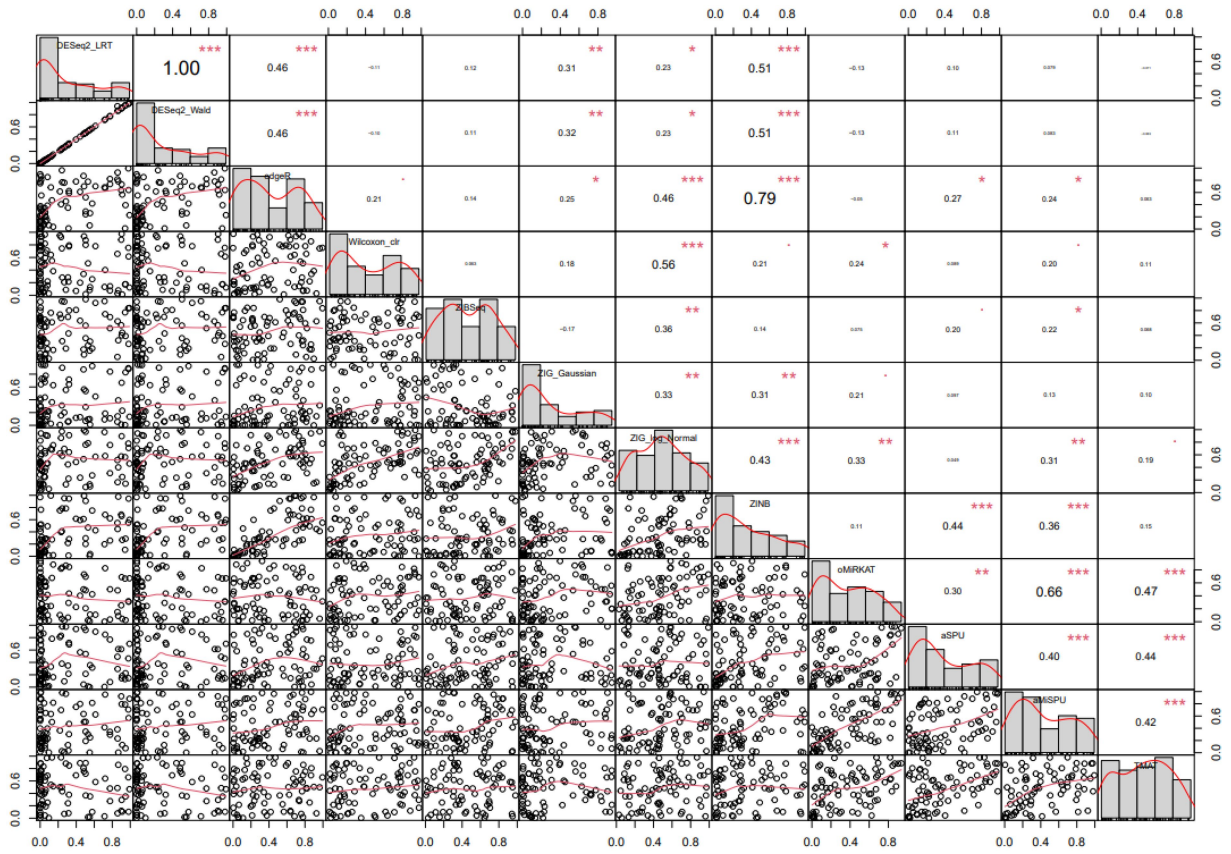


FIGURE 5
Pairwise Spearman correlation coefficients computed from various statistical methods. Spearman correlation of each pair of methods are represented in the upper diagonal graphs. The bigger the numbers the stronger the correlation. Lower diagonal scatterplots represent *p*-values. Diagonal graphs with the method name have histogram of p-values. Methods are in the order of DESeq2[LRT/Wald], edgeR, Wilcoxon CLR, ZIBSeq, ZIG[Gaussian/log normal], ZINB, oMiRKAT, aSPU, aMiSPU, TMAT.

compared to healthy controls was reported (Sobhani et al., 2019). *Bilophila wadsworthia* was reported to produce genotoxic hydrogen sulfide in the gut, enhancing carcinogenesis (Coker et al., 2022). The relative abundance of *Anaerostipes* were reported to be reduced in CRC patients compared to healthy controls (Chen et al., 2012).

Similar to the previous results with Baxter's data, Fisher's method, Kost's method, and Stouffer's method ranked CRC-related important genera lower than Cauchy combination test. For example, *Fusobacterium*, which was ranked 3rd by Cauchy combination test, was ranked 12th, 12th, and 36th, respectively. Similarly, *Cloacibacillus*, which was ranked 5th by Cauchy combination test, was ranked 15.5th, 15th, and 18th, respectively.

We also compared the results obtained from two CRC datasets (Baxter's data and Zeller's data). A total of 64 common genera were found. *Fusobacterium* was found to be the rank of 27.5 out of 80 genera in Baxter's data, but the rank of 3 out of 81 genera in Zeller's data. The value of *p* trend of the two datasets, and there were four commonly significant genera (*q*-value <0.05). *Fusobacterium* was found to be significant in Zeller's data, but not in Baxter's data with *q*-value of 0.133.

The commonly significant genera from the real datasets were investigated. There were 22 significant microbiome genera (*q*-value <0.05) from Zeller's data, and 9 significant microbiome genera from Baxter's data (*q*-value <0.05). Among them, there were four commonly significant microbiome genera from the two datasets. *Cloacibacillus* was previously found to be related to late-stage CRC patients (Sheng et al., 2019). *Porphyromonas* has been reported to be enriched in gut microbiota profiling of CRC patients in several studies (Yang et al., 2019). *Clostridia vadinBB60 group* was previously found to be enriched in low-graded; right-sided/transverse tumors (Zwinsová et al., 2021). *Streptococcus* was reported to have increased relative abundance in CRA compared to healthy controls (Sun et al., 2020). Furthermore, *Streptococcus gallolyticus* is known as opportunistic pathogen causing infections associated with colon neoplasia in the elderly (Périchon et al., 2022).

Discussion

In this study, we conducted empirical studies to determine the most appropriate value of *p* combination method for microbiome

TABLE 4 Top 10 microbiome genera ranked by Cauchy combination test.

Taxa (o:order, f:family, g:genus)	*q*-value
o__Rhodospirillales; f__uncultured; g__uncultured	5.64E-20
o__Veillonellales-Selenomonadales; f__Veillonellaceae; g__Megasphaera	1.22E-16
o__Gastranaerophilales; f__Gastranaerophilales; g__Gastranaerophilales	3.72E-15
*o__Synergistales; f__Synergistaceae; g__Cloacibacillus**	7.42E-13
*o__Bacteroidales; f__Porphyromonadaceae; g__Porphyromonas**	4.23E-09
*o__Clostridia_vadinBB60_group; f__Clostridia_vadinBB60_group; g__Clostridia_vadinBB60_group**	1.12E-07
o__Burkholderiales; f__Sutterellaceae; g__Sutterella	2.13E-05
o__Bacteroidales; f__Marinifilaceae; g__Odoribacter	1.09E-05
o__Erysipelotrichales; f__Erysipelotrichaceae; g__Turicibacter	1.51E-04
o__Coriobacteriales; f__Eggerthellaceae; g__Slackia	1.87E-04

*Commonly significant microbiome genera with Zeller's data.

TABLE 5 Top 10 microbiome genera ranked by Cauchy combination test.

Taxa (o:order, f:family, g:genus)	*q*-value
*o__Bacteroidales; f__Porphyromonadaceae; g__Porphyromonas**	1.80E-14
o__Lachnospirales; f__Lachnospiraceae; g__Hungatella	3.26E-13
o__Fusobacteriales; f__Fusobacteriaceae; g__Fusobacterium	2.78E-09
o__Bacteroidales; f__Rikenellaceae; g__Rikenellaceae_RC9_gut_group	1.11E-06
*o__Synergistales; f__Synergistaceae; g__Cloacibacillus**	1.35E-06
o__Veillonellales-Selenomonadales; f__Veillonellaceae; g__Veillonella	1.47E-06
o__Erysipelotrichales; f__Erysipelatoclostridiaceae; g__Catenibacterium	1.11E-05
o__Veillonellales-Selenomonadales; f__Selenomonadaceae; g__Mitsuokella	1.36E-05
o__Desulfovibrionales; f__Desulfovibrionaceae; g__Bilophila	5.75E-05
o__Lachnospirales; f__Lachnospiraceae; g__Anaerostipes	1.05E-04

*Commonly significant microbiome genera with Baxter's data.

data. Cauchy combination test was determined to be the most appropriate in terms of type 1 error rates, power, and showed the highest consistency with the true rank than other methods.

The power and type 1 error rates were assessed because it was important to know whether the combined value of *p*s controlled type 1 error rates. For Fisher's method and Stouffer's method, the uncontrolled type 1 error rates were observed. Since it was shown that the value of *p*s produced from various methods had significant correlations, Fisher's method and Stouffer's method that combine value of *p*s based on the independent assumption of *p*-values tended to show uncontrolled type 1 error rates in some conditions. On the other hand, Kost method incorporating the correlation between the combined *p*-values yielded well-controlled type 1 error rates. Cauchy combination test is a powerful *p*-value combination method robust to arbitrary dependency structures, effectively accounting for the dependency structure of the microbiome dataset.

In our analysis, we considered 12 DE analyses and proposed combining all 12 value of *p*s. Our method can be applicable to any number of DE analyses. For illustrative purposes, we performed the similar analyses using only a fewer DE methods. We considered combining the following methods: (1) taxa-level methods, (2) community-level methods, (3) three randomly chosen methods, (4) five randomly chosen methods, (5) seven randomly chosen methods, (6) a correlated set of methods, (7) another correlated set of methods, and (8) less correlated set of methods. For the randomly chosen three/five/seven methods, we simply applied on a single random set of methods each. Each case resulted similar power trend with that of using all 12 methods (Supplementary Figures 3–10).

In this study, we formulated the difficulty of analyzing microbiome datasets in the sense of overdispersion and high sparsity, by using different analysis methods accounting for these traits. However, one may want to focus on other traits, such as different normalization strategies. We leave it as a future study.

From the rank simulation, Cauchy combination test showed the best performance with significant differences from other value of *p* combination methods for scenarios 1 and 3, while it showed similar performance in scenario 2. Note that scenarios 1 and 3 had six and eight non-causal dependent variables, respectively, while scenario 2 had four non-causal dependent variables and six different causal dependent variables. This implies that Cauchy combination test has the better performance when several non-causal microbiome genera exist. This corresponds to the real microbiome dataset that has several non-causal microbiome taxa and few causal taxa.

The microbiome ranks generated by Cauchy combination test and min P or Simes method did not differ much for the top ranks

in the real data analysis. Rather, similar trends of value of *p*s and high correlation coefficients between those methods were observed (Supplementary Figure 11). The difference of microbiome ranks was most obvious with Stouffer's method, and it was shown that the top ranks generated by Cauchy combination test and Stouffer's method were quite different. The top ranks generated using Fisher's method and Kost method did not differ much from those generated using Cauchy combination test. The ranks generated using Fisher's method and Kost method were the same because they both follow chi-square distributions with different degrees of freedom. Kost method follows a scaled chi-square distribution, but scaling did not alter the resulting ranks.

Most microbiome features have very high sparsity and low abundance, making the statistical analysis difficult. In this study, we considered those characteristics in assessing the different value of *p* combination methods by simulating different levels of sparsity and setting a microbiome feature with high sparsity and low abundance as causal.

The value of *p* combination approach used to determine microbiome importance considering microbiome-specific characteristics can be easily extended to other omics data analyses. For example, our approach can be applied to analysis in RNA-seq or copy number variation data considering data-specific characteristics. There also are several methods to analyze each type of dataset. Note that there is "no one real winner that performs the best." Thus, combining the results from various methods can have the advantage of using all methods available and being robust to the method-specific assumptions. Cauchy combination test can effectively combine different statistical methods, and produces a representative result of all methods, instead of using a single method that could possibly have a good performance in one dataset, but not in others. Our empirical study showed that the performance of Cauchy combination method provided robust and reasonable result compared to the best performing individual DE method, and performed the best among the value of *p* combination methods in terms of power and rank similarity, and controlling type 1 error rates (supplementary Figure 12). Furthermore, we made a python script with the module "mpmath" that enables floating point arithmetic in case the resulting value of *p*s from individual analysis methods are minute for the combined value of *p* of Cauchy combination test (Cauchy_pval.py). All combination methods used in this script are provided as a R script in https://github.com/HyeonJungHam/P_value_combination, that also includes automatic execution of python script for calculating Cauchy combination test *p*-value.

While Cauchy combination test was introduced with equal weights for each method, it can be easily extended to handle unequal weights. By the authors, Cauchy combination test still accounts for the arbitrary dependency structure when the weights are random variables and independent of test statistics (Liu and Xie, 2020). Thus, it is reasonable to assign a larger weight to the method providing more reliable and accurate result. We expect that the optimal weights would result in an increased performance of Cauchy combination test. However, the choice of optimal weights can change across dataset. Thus, given a dataset, it would not be straightforward to choose the optimal weights. We will leave the choice of optimal weights as a future research topic.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: The raw sequencing data and metadata of Baxter's data are available from Sequence Read Archive (SRA) publicly under the accession number of SRP062005. The raw sequencing data and metadata of Zeller's data are available at European Nucleotide Archive (ENA) with the project number PRJEB6070.

Ethics statement

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

Author contributions

TP contributed to conception and design of the study, and manuscript revision, read, and approved the submitted version. HH performed the statistical analysis and wrote the first draft of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the National Research Foundation, the Korea Health Technology R8D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare (HI16C2037).

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.2022.990870/full#supplementary-material

References

Baxter, N. T., Koumpouras, C. C., Rogers, M. A., Ruffin, M. T. 4th, and Schloss, P. D. (2016). DNA from fecal immunochemical test can replace stool for detection of colonic lesions using a microbiota-based model. *Microbiome* 4:59. doi: 10.1186/s40168-016-0205-y

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., et al. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* 10, 57–59. doi: 10.1038/nmeth.2276

Burrello, C., Garavaglia, F., Cribiù, F. M., Ercoli, G., Lopez, G., Troisi, J., et al. (2018). Therapeutic faecal microbiota transplantation controls intestinal inflammation through IL10 secretion by immune cells. *Nat. Commun.* 9:5184. doi: 10.1038/s41467-018-07359-8

Calgaro, M., Romualdi, C., Waldron, L., Risso, D., and Vitulo, N. (2020). Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biol.* 21:191. doi: 10.1186/s13059-020-02104-1

Casella, G, and Berger, R. (2017). *Statistical Inference*. *2nd Edn*. Belmont:Cengage Learning. p. 229.

Chen, J., and Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* 7, 418–442. doi: 10.1214/12-AOAS592

Chen, W., Liu, F., Ling, Z., Tong, X., and Xiang, C. (2012). Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PLoS One* 7:e39743. doi: 10.1371/journal.pone.0039743

Chen, H. I. H., Liu, Y., Zou, Y., Lai, Z., Sarkar, D., Huang, Y., et al. (2015). Differential expression analysis of RNA sequencing data by incorporating non-exonic mapped reads. *BMC Genomics* 16:S14. doi: 10.1186/1471-2164-16-S7-S14

Cheng, L., and Sheng, X. S. (2017). Combination of combinations of p values. *Empir. Econ.* 53, 329–350. doi: 10.1007/s00181-017-1230-9

Coker, O. O., Liu, C., Wu, W. K. K., Wong, S. H., Jia, W., Sung, J. J. Y., et al. (2022). Altered gut metabolites and microbiota interactions are implicated in colorectal carcinogenesis and can be non-invasive diagnostic biomarkers. *Microbiome* 10:35. doi: 10.1186/s40168-021-01208-5

Coleman, O. I., and Nunes, T. (2016). Role of the microbiota in colorectal cancer: updates on microbial associations and therapeutic implications. *Biores. Open Access.* 5, 279–288. doi: 10.1089/biores.2016.0028

Cronin, P., Murphy, C. L., Barrett, M., Ghosh, T. S., Pellanda, P., O'Connor, E. M., et al. (2022). Colorectal microbiota after removal of colorectal cancer, NAR. *Cancer* 4:zcac011. doi: 10.1093/narcan/zcac011

Dong, Z., Chen, B., Pan, H., Wang, D., Liu, M., Yang, Y., et al. (2019). Detection of microbial 16S rRNA gene in the serum of patients with gastric cancer. *Front. Oncol.* 9:608. doi: 10.3389/fonc.2019.00608

Dueholm, M. S., Andersen, K. S., Mcllroy, S. J., Kristensen, J. M., Yashiro, E., Karst, S. M., et al. (2020). Generation of comprehensive ecosystem-specific reference databases with species-level resolution by high-throughput full-length 16S rRNA gene sequencing and automated taxonomy assignment AutoTax. *MBio* 11, e01557–e01520. doi: 10.1128/mBio.01557-20

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Ganju, J., and Ma, G. J. (2017). The potential for increased power from combining P-values testing the same hypothesis. *Stat. Methods Med. Res.* 26, 64–74. doi: 10.1177/0962280214538016

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Hayes, R. B., Ahn, J., Fan, X., Peters, B. A., Ma, Y., Yang, L., et al. (2018). Association of oral microbiome with risk for incident head and neck squamous cell cancer. *JAMA Oncol.* 4, 358–365. doi: 10.1001/jamaoncol.2017.4777

Jia, W., Rajani, C., Xu, H., and Zheng, X. (2021). Gut microbiota alterations are distinct for primary colorectal cancer and hepatocellular carcinoma. *Protein Cell* 12, 374–393. doi: 10.1007/s13238-020-00748-0

Khomich, M., Måge, I., Rud, I., and Berget, I. (2021). Analysing microbiome intervention design studies: comparison of alternative multivariate statistical methods. *PLoS One* 16:e0259973. doi: 10.1371/journal.pone.0259973

Kim, O. S., Cho, Y. J., Lee, K., Yoon, S. H., Kim, M., Na, H., et al. (2012). Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int. J. Syst. Evol. Microbiol.* 62, 716–721. doi: 10.1099/ijs.0.038075-0

Kim, S. I., Kang, N., Leem, S., Yang, J., Jo, H., Lee, M., et al. (2020). Metagenomic analysis of serum microbe-derived extracellular vesicles and diagnostic models to differentiate ovarian cancer and benign ovarian tumor. *Cancers* 12:1309. doi: 10.3390/cancers12051309

Kim, K. J., Park, J., Park, S. C., and Won, S. (2020). Phylogenetic tree-based microbiome association test. *Bioinformatics* 36, 1000–1006. doi: 10.1093/bioinformatics/btz686

Koh, H., Blaser, M. J., and Li, H. (2017). A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. *Microbiome* 5:45. doi: 10.1186/s40168-017-0262-x

Kost, J., and McDermott, M. (2002). Combining dependent *P*-values. *Stat. Probab. Lett.* 60, 183–190. doi: 10.1016/S0167-7152(02)00310-3

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., and Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* 79, 5112–5120. doi: 10.1128/AEM.01043-13

Liu, Y., and Xie, J. (2020). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *JASA* 115, 393–402. doi: 10.1080/01621459.2018.1554485

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Mori, G., Rampelli, S., Orena, B. S., Rengucci, C., Maio, G. D., Barbieri, G., et al. (2018). Shifts of faecal microbiota during sporadic colorectal carcinogenesis. *Sci. Rep.* 8:10329. doi: 10.1038/s41598-018-28671-9

Nazarieh, M., Rajula, H., and Helms, V. (2019). Topology consistency of disease-specific differential co-regulatory networks. *BMC Bioinformat.* 20:550. doi: 10.1186/s12859-019-3107-8

Nearing, J. T., Douglas, G. M., Hayes, M. G., MacDonald, J., Desai, D. K., Allward, N., et al. (2022). Microbiome differential abundance methods produce different results across 38 datasets. *Nat. Commun.* 13:342. doi: 10.1038/s41467-022-28034-z

Okumura, S., Konishi, Y., Narukawa, M., Sugiura, Y., Yoshimoto, S., Arai, Y., et al. (2021). Gut bacteria identified in colorectal cancer patients promote tumourigenesis via butyrate secretion. *Nat. Commun.* 12:5674. doi: 10.1038/s41467-021-25965-x

Osman, M.-A., Neoh, H., Ab Mutalib, N.-S., Chin, S.-F., and Jamal, R. (2018). 16S rRNA gene sequencing for deciphering the colorectal cancer gut microbiome: current protocols and workflows. *Front. Microbiol.* 9:767. doi: 10.3389/fmicb.2018.00767

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

Peng, X., Li, G., and Liu, Z. (2016). Zero-inflated beta regression for differential abundance analysis with metagenomics data. *J. Comput. Biol.* 23, 102–110. doi: 10.1089/cmb.2015.0157

Périchon, B., Lichtl-Häfele, J., Bergsten, E., Delage, V., Trieu-Cuot, P., Sansonetti, P., et al. (2022). Detection of streptococcus gallolyticus and four other CRC-associated bacteria in patient stools reveals a potential driver role for enterotoxigenic *Bacteroides fragilis*. *Front. Cell. Infect. Microbiol.* 11:794391. doi: 10.3389/fcimb.2022.794391

Plantinga, A., Zhan, X., Zhao, N., Chen, J., Jenq, R. R., and Wu, M. C. (2017). MiRKAT-S: a community-level test of association between the microbiota and survival times. *Microbiome* 5:17. doi: 10.1186/s40168-017-0239-9

Pollock, J., Glendinning, L., Wisedchanwet, T., and Watson, M. (2018). The madness of microbiome: attempting to find consensus best practice for 16S microbiome studies. *Appl. Environ. Microbiol.* 84, e02627–e02617. doi: 10.1128/AEM.02627-17

Qian, Y., Yang, X., Xu, S., Wu, C., Qin, N., Chen, S. D., et al. (2018). Detection of microbial 16S rRNA gene in the blood of patients with Parkinson's disease. *Front. Aging Neurosci.* 10:156. doi: 10.3389/fnagi.2018.00156

Queen, J., Domingue, J. C., White, J. R., Stevens, C., Udayasuryan, B., Nguyen, T. T. D., et al. (2022). Comparative analysis of colon cancer-derived fusobacterium nucleatum subspecies: inflammation and colon tumorigenesis in murine models. *Bacteriology* 8:e0299121. doi: 10.1128/mbio.02991-21

Ramsheh, M. Y., Haldar, K., Esteve-Codina, A., Purser, L. F., Richardson, M., Müller-Quernheim, J., et al. (2021). Lung microbiome composition and bronchial epithelial gene expression in patients with COPD versus healthy individuals: a bacterial 16S rRNA gene sequencing and host transcriptomic analysis. *Lancet Microb.* 2, E300–E310. doi: 10.1016/S2666-5247(21)00035-5

Ritari, J., Salojärvi, J., Lahti, L., and de Vos, W. M. (2015). Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics* 16:1056. doi: 10.1186/s12864-015-2265-y

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., et al. (2020). NCBI taxonomy: A comprehensive update on curation, resources and tools. *Database* 2020:baaa062. doi: 10.1093/database/baaa062

Shao, L., Guo, Y., Wang, L., Chen, M., Zhang, W., Deng, S., et al. (2022). Effects of ginsenoside compound K on colitis-associated colorectal cancer and gut microbiota profiles in mice. *Ann. Transl. Med* 10:408. doi: 10.21037/atm-22-793

Sheng, Q., Du, H., Cheng, X., Cheng, X., Tang, Y., Pan, L., et al. (2019). Characteristics of fecal gut microbiota in patients with colorectal cancer at different stages and different sites. *Oncol. Lett.* 18, 4834–4844. doi: 10.3892/ol.2019.10841

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73, 751–754.

Sinha, R., Ahn, J., Sampson, J. N., Shi, J., Yu, G., Xiong, X., et al. (2016). Fecal microbiota, fecal metabolome, and colorectal cancer interrelations. *PLoS One* 11:e0152126. doi: 10.1371/journal.pone.0152126

Sobhani, I., Bergsten, E., Couffin, S., Amiot, A., Nebbad, B., Barau, C., et al. (2019). Colorectal cancer-associated microbiota contributes to oncogenic epigenetic signatures. *Proc. Natl. Acad. Sci. U. S. A.* 116, 24285–24295. doi: 10.1073/pnas.1912129116

Sohn, M. B., and Li, H. (2018). A GLM-based latent variable ordination method for microbiome samples. *Biometrics* 74, 448–457. doi: 10.1111/biom.12775

Stouffer, S. A., and Suchman, E. A. (1949). The American soldier, vol. 1. Adjustment during army life. *J. Consult. Psychol.* 13:310.

Sun, W., Wang, L., Zhang, Q., and Dong, Q. (2020). Microbial biomarkers for colorectal cancer identified with random Forest model. *ERHM* 5, 19–26. doi: 10.14218/ERHM.2019.00026

Tarashi, S., Siadat, S. D., Badi, S. A., Zali, M., Biassoni, R., Ponzoni, M., et al. (2019). Which one is the defendant for colorectal cancer? *Microorganisms* 7:561. doi: 10.3390/microorganisms7110561

Tippett, L.H.C. (1931). *The Methods of Statistics*. London: Williams Norgate Ltd.

van der Lugt, B., Rusli, F., Lute, C., Lamprakis, A., Salazar, E., Boekschoten, M. V., et al. (2018). Integrative analysis of gut microbiota composition, host colonic gene expression and intraluminal metabolites in aging C57BL/6J mice. *Aging* 10, 930–950. doi: 10.18632/aging.101439

Wang, X., Jia, Y., Wen, L., Mu, W., Wu, X., Liu, T., et al. (2021). Porphyromonas gingivalis promotes colorectal carcinoma by activating the hematopoietic NLRP3 inflammasome. *Cancer Res.* 81, 2745–2759. doi: 10.1158/0008-5472.CAN-20-3827

Wang, Z., Jin, S., and Zhang, C. (2019). A method based on differential entropy-like function for detecting differentially expressed genes across multiple conditions in RNA-Seq studies. *Entropy* 21:242. doi: 10.3390/e21030242

Wu, C., Chen, J., Kim, J., and Pan, W. (2016). An adaptive association test for microbiome data. *Genome Med.* 8:56. doi: 10.1186/s13073-016-0302-3

Wu, Y., Jiao, N., Zhu, R., Zhang, Y., Wu, D., Wang, A., et al. (2021). Identification of microbial markers across populations in early detection of colorectal cancer. *Nat. Commun.* 12:3063. doi: 10.1038/s41467-021-23265-y

Wu, M., Li, J., An, Y., Li, P., Xiong, W., Li, J., et al. (2019). Chitooligosaccharides prevents the development of colitis-associated colorectal cancer by modulating the intestinal microbiota and mycobiota. *Front. Microbiol.* 10:2101. doi: 10.3389/fmicb.2019.02101

Wu, J., Li, Q., and Fu, X. (2019). Fusobacterium nucleatum contributes to the carcinogenesis of colorectal cancer by inducing inflammation and suppressing host immunity. *Transl. Oncol.* 12, 846–851. doi: 10.1016/j.tranon.2019.03.003

Xia, Y., and Sun, J. (2017). Hypothesis testing and statistical analysis of microbiome. *Genes. Dis.* 4, 138–148. doi: 10.1016/j.gendis.2017.06.001

Xia, Y., Sun, J., and Chen, D. G. (2018). "Modeling zero-inflated microbiome data," in *Statistical Analysis of Microbiome Data with R. ICSA Book Series in Statistics.* Singapore: Springer.

Xing, C., Wang, M., Adebusola, A. A., Tan, P., Fu, C., Chen, L., et al. (2021). Microbiota regulate innate immune signaling and protective immunity against cancer. *Cell Host Microbe* 29, 959–974.e7. doi: 10.1016/j.chom.2021.03.016

Yang, Y., Chen, G., Yang, Q., Ye, J., Cai, X., Tsering, P., et al. (2017). Gut microbiota drives the attenuation of dextran sulphate sodium-induced colitis by Huangqin decoction. *Oncotarget* 8, 48863–48874. doi: 10.18632/oncotarget.16458

Yang, J., McDowell, A., Kim, E. K., Seo, H., Lee, W. H., Moon, C. M., et al. (2019). Development of a colorectal cancer diagnostic model and dietary risk assessment through gut microbiome analysis. *Exp. Mol. Med.* 51, 1–15. doi: 10.1038/s12276-019-0313-4

Yoon, S., Baik, B., Park, T., and Nam, D. (2021). Powerful p-value combination methods to detect incomplete association. *Sci. Rep.* 11:6980. doi: 10.1038/s41598-021-86465-y

York, A. (2021). Guarding against colorectal cancer. *Nat. Rev. Microbiol.* 19:405. doi: 10.1038/s41579-021-00572-1

You, J., Corley, S. M., Wen, L., Hodge, C., Höllhumer, R., Madigan, M. C., et al. (2018). RNA-Seq analysis and comparison of corneal epithelium in keratoconus and myopia patients. *Sci. Rep.* 8:389. doi: 10.1038/s41598-017-18480-x

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10:766. doi: 10.15252/msb.20145645

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., et al. (2015). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.* 96, 797–807. doi: 10.1016/j.ajhg.2015.04.003

Zwinsová, B. A., Petrov, V., Hrivňáková, M., Smatana, S., Micenková, L., Kazdová, N., et al. (2021). Colorectal tumour mucosa microbiome is enriched in Oral pathogens and defines three subtypes that correlate with markers of tumour progression. *Cancers* 13:4799. doi: 10.3390/cancers13194799

frontiers | Frontiers in Microbiology

TYPE Methods
PUBLISHED 05 December 2022
DOI 10.3389/fmicb.2022.1062281

OPEN ACCESS

EDITED BY
Fei Ma,
Chinese Academy of Medical Sciences and Peking Union Medical College, China

REVIEWED BY
Chirasmita Nayak,
Alagappa University, India
Guohua Huang,
Shaoyang University, China

*CORRESPONDENCE
Ju Xiang
xiang.ju@foxmail.com
Jianjun He
hejianjun@csmu.edu.cn
Binsheng He
hbscsmu@163.com

†These authors have contributed equally to this work

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

RECEIVED 05 October 2022
ACCEPTED 21 November 2022
PUBLISHED 05 December 2022

CITATION
Wang Y, Xiang J, Liu C, Tang M, Hou R, Bao M, Tian G, He J and He B (2022) Drug repositioning for SARS-CoV-2 by Gaussian kernel similarity bilinear matrix factorization.
Front. Microbiol. 13:1062281.
doi: 10.3389/fmicb.2022.1062281

Drug repositioning for SARS-CoV-2 by Gaussian kernel similarity bilinear matrix factorization

Yibai Wang[1†], Ju Xiang[1,2*†], Cuicui Liu[1], Min Tang[3], Rui Hou[4,5], Meihua Bao[6,7], Geng Tian[4,5], Jianjun He[2,6,7*] and Binsheng He[2,6,7*]

[1]School of Information Engineering, Changsha Medical University, Changsha, China, [2]Academician Workstation, Changsha Medical University, Changsha, China, [3]School of Life Sciences, Jiangsu University, Zhenjiang, Jiangsu, China, [4]Geneis (Beijing) Co., Ltd., Beijing, China, [5]Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China, [6]School of Pharmacy, Changsha Medical University, Changsha, China, [7]Key Laboratory Breeding Base of Hunan Oriented Fundamental and Applied Research of Innovative Pharmaceutics, Changsha Medical University, Changsha, China

Coronavirus disease 2019 (COVID-19), a disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is currently spreading rapidly around the world. Since SARS-CoV-2 seriously threatens human life and health as well as the development of the world economy, it is very urgent to identify effective drugs against this virus. However, traditional methods to develop new drugs are costly and time-consuming, which makes drug repositioning a promising exploration direction for this purpose. In this study, we collected known antiviral drugs to form five virus-drug association datasets, and then explored drug repositioning for SARS-CoV-2 by Gaussian kernel similarity bilinear matrix factorization (VDA-GKSBMF). By the 5-fold cross-validation, we found that VDA-GKSBMF has an area under curve (AUC) value of 0.8851, 0.8594, 0.8807, 0.8824, and 0.8804, respectively, on the five datasets, which are higher than those of other state-of-art algorithms in four datasets. Based on known virus-drug association data, we used VDA-GKSBMF to prioritize the top-k candidate antiviral drugs that are most likely to be effective against SARS-CoV-2. We confirmed that the top-10 drugs can be molecularly docked with virus spikes protein/human ACE2 by AutoDock on five datasets. Among them, four antiviral drugs ribavirin, remdesivir, oseltamivir, and zidovudine have been under clinical trials or supported in recent literatures. The results suggest that VDA-GKSBMF is an effective algorithm for identifying potential antiviral drugs against SARS-CoV-2.

KEYWORDS

SARS-CoV-2, drug repositioning, bilinear matrix factorization, molecular docking, machine learning

Introduction

Caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a new infectious disease called coronavirus disease 2019 (COVID-19) has caused a big pandemic worldwide since 2019 (Eurosurveillance editorial team, 2020; Cheng et al., 2021a; Zhang et al., 2021). SARS-CoV-2 can transmit by human-to-human contacts, and is currently spreading rapidly to more than 400 countries around the world, causing millions of deaths (Coronaviridae Study Group of the International Committee on Taxonomy of V, 2020; Li et al., 2020; Cohain et al., 2021). Thus, SARS-CoV-2 seriously threatens human life and health as well as the development of world economy (Wu et al., 2020; Zhou P. et al., 2020; Zhu et al., 2020; Cheng et al., 2021b), and it is critical to find effective measures to prevent the transmission and fight against this virus.

One effective way to prevent the transmission of a virus is through vaccination. However, viruses like SARS-CoV-2 and influenzas are under rapid genetic and antigenic evolution, especially in their spike proteins (Yao et al., 2017; Zhang et al., 2017), which will make the vaccine less effective. Another method is to develop specific drug against the viruses. However, traditional methods to develop new drugs usually take years and cost tens of millions of dollars (Novac, 2013). With the development of various computational algorithms for mining intrinsic associations in biomedical data (Zhang et al., 2019; Xu et al., 2020a; Liu et al., 2021; Xiang et al., 2021a, 2022b; He et al., 2022; Yang et al., 2022), drug repositioning has become an effective way of exploring new uses for approved drugs, since it can significantly reduce the time and cost in the development of drugs (Liu et al., 2016, 2020; Yang J. et al., 2020; Zhu et al., 2021).

There are a few studies to prioritize approved drugs against SARS-CoV-2. For example, Zhou et al. proposed a KATZ method to probe antiviral drugs against SARS-CoV-2 through virus-drug association prediction (Zhou L. et al., 2020). More recently, Tang et al. prioritized drugs for COVID-19 through an indicator regularized non-negative matrix factorization method (Tang et al., 2020). Peng et al. collected an antivirial drug database and minied it to repurpose drugs aginst SARS-CoV-2 (Peng et al., 2020; Zhou L. et al., 2020). Wang et al. predicted anti-SARS-COV-2 drugs by bound nuclear norm regularization (Wang et al., 2021). Meng et al. builded the human drug virus database and identified anti-SARS-COV-2 drugs by similarity constrained probabilistic matrix factorization (Lu et al., 2021; Meng et al., 2021; Parsza et al., 2021). Shen et al. prioritized anti-SARS-CoV-2 drugs by combining an unbalanced bi-random walk and Laplacian regularized least squares (Shen et al., 2022). Though these methods achieved relatively good prediction performance in cross-validation and literature mining, the accuracy of prediction is yet to be improved and a more robust validation method is needed for further wet-lab experiments. Therefore, in this study, we collected the data of well-studied viruses that are similar to SARS-CoV-2 and their known antiviral drugs, forming a virus-drug association matrix (VDA). Then, we proposed a novel method for exploring potential virus-drug associations of SARS-CoV-2 by using Gaussian kernel similarity bilinear matrix factorization (VDA-GKSBMF).

The rest of the work is organized as follows. First, we collect five datasets and propose the details of the VDA-GKSBMF method for predicting potential virus-drug associations of SARS-CoV-2. Then, we study the effectiveness of the method by the 5-fold cross-validation experiments and compare VDA-GKSBMF with other state-of-art algorithms. Based on known virus-drug association data, we use VDA-GKSBMF to prioritize top-10 candidate antiviral drugs that are most likely to fight against SARS-CoV-2, and then evaluate the molecular binding activity between predicted antiviral drugs and SARS-CoV-2 spike protein (Gralinski, 2020) or human ACE2 (Zhao et al., 2020), to confirm whether the top-10 drugs are to be molecularly docked with the virus spikes protein or human ACE2. We also explore literatures to check if the top predicted drugs are under clinical trials or experiments against SARS-CoV-2.

Materials and methods

The overall workflow of the method is illustrated in Figure 1. We first introduce the datasets in this study, and then describe the details of the VDA-GKSBMF method for drug repositioning of SARS-CoV-2, including the construction of virus–drug heterogeneous network and the VDA-GKSBMF model, along with the alternating direction method of multipliers (ADMM) for solving the model to fill out unknown associations in virus–drug matrix.

Materials

To identify potential VDAs involving SARS-COV-2, we collect five datasets. There is Virus similarity matrix, drug similarity matrix, and VDA matrix in each dataset. Viruses are similar to SARS-CoV-2, small-molecule drugs and VDAs between them from the DrugBank (Wishart et al., 2018), PubChem (Kim et al., 2016), and NCBI (Wheeler et al., 2004) databases (see Table 1 for details).

These VDAs are represented by a VDA matrix $B_{m\times n}$, where $B_{dv}=1$ if the d-th drug is associated with the v-th virus, otherwise, $B_{dv}=0$. This forms a virus-drug association network, which can be denoted as a bipartite graph $G(V,D,E)$, where $E(G)=\{e_{ij}\}\subseteq V\times D$ contains edges representing known associations between viruses and drugs.

For viruses, we obtain the sequence-based similarities between viruses that are calculated by MAFFT (Katoh and Toh, 2008). For drugs, we obtain the chemical structure-based similarity scores between drugs by RDKit (Landrum, 2014), where chemical structures of drugs are obtained from the DrugBank database (Wishart et al., 2018). The details are shown in Table 1.

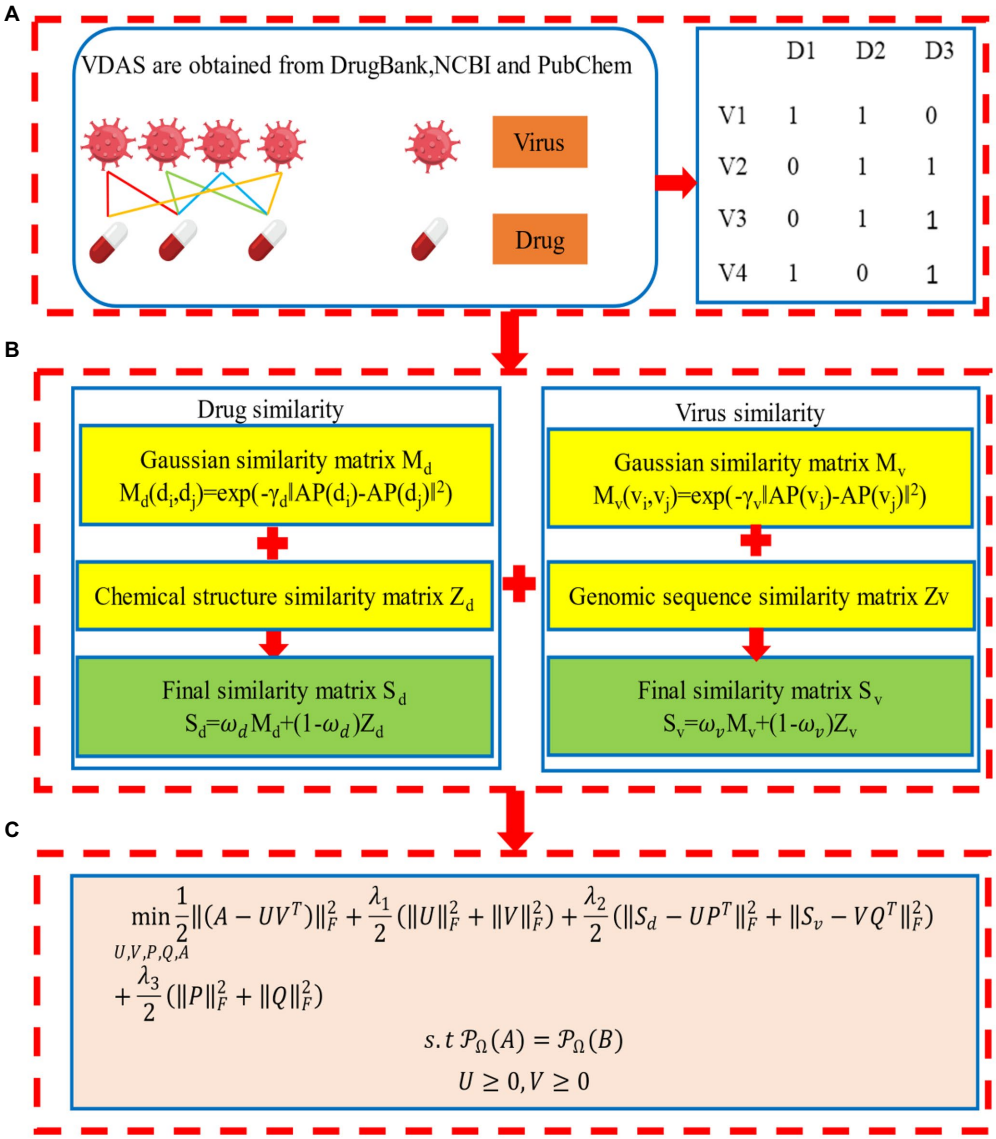


FIGURE 1
Workflow of Gaussian kernel similarity bilinear matrix factorization (VDA-GKSBMF). **(A)** Virus–drug association network and its association matrix. **(B)** Drug–drug similarity matrix and Virus–virus similarity matrix. **(C)** The model of VDA-GKSBMF.

TABLE 1 The statistics of datasets.

Datasets	No. of viruses	No. of drug	No. of VDAS	Sparsity
Dataset1	12	78	96	89.7%
Dataset2	69	128	770	91.3%
Dataset3	34	203	407	95.0%
Dataset4	34	210	437	93.9%
Dataset5	34	219	455	93.9%

Methods

Drug similarity matrix

Considering that drugs with common associated viruses may be similar, we denote the Gaussian association profile (AP) of drug d_i by $\mathrm{AP}(d_i)$, i.e., the i-th row of the VDA matrix B, which is a binary vector encoding the associations between this drug and viruses in the VDA matrix. Then, we calculate the similarity $M_d(d_i,d_j)$ between two drugs d_i and d_j based on association profiles of drugs by,

$$M_d(d_i,d_j)=\exp\left(-\gamma_d\|\mathrm{AP}(d_i)-\mathrm{AP}(d_j)\|^2\right)$$

where $\gamma_d=\gamma'_d/(\frac{1}{m}\sum_{k=1}^{m}\|AP(d_k)\|^2)$ is the normalized core band-width based on bandwidth parameter γ'_d, and m denotes the number of drugs.

Then, we obtain the chemical structure (CS)-based similarity between drugs calculated by RDKit (Landrum, 2014), which is

denoted as Z_d. Finally, we generate the drug–drug similarity matrix (DDS) by,

$$S_d = \omega_d M_d + (1-\omega_d) Z_d,$$

where $\omega_d \in [0,1]$ balances the contribution of the CS-based and AP-based drug similarity matrices. This forms a drug–drug network with edges weighted by the pairwise drug similarity scores.

Virus similarity matrix

Considering that viruses with common associated drugs may be similar, in the same way, we denote the Gaussian association profile (AP) of virus v_a by $\mathrm{AP}(v_a)$, i.e., the a -th column of the VDA matrix B, which is a binary vector encoding the associations between this virus and drugs in the VDA matrix. We calculate the AP-based similarity $M_v(v_a,v_b)$ between two viruses by,

$$M_v(v_a,v_b) = \exp\left(-\gamma_v \| \mathrm{AP}(v_a) - \mathrm{AP}(v_b) \|^2\right),$$

where $\gamma_v = \gamma'_v / (\frac{1}{n}\sum_{k=1}^{n} \| AP(v_k) \|^2)$, and n denotes the number of viruses.

Then, we obtain the sequence (SQ)-based similarity matrix calculated by MAFFT (Katoh and Toh, 2008), which is denoted as Z_v. Finally, the virus-virus similarity matrix (VVS) is calculated by,

$$S_v = \omega_v M_v + (1-\omega_v) Z_v,$$

where $\omega_v \in [0,1]$ balances the contribution of the SQ-based and AP-based virus similarity matrices. This forms a virus-virus network with edges weighted by the pairwise virus similarity scores.

Constructing heterogeneous network

To make use of information in the above DDS, VVS, and VDA matrices, we integrate them to construct a heterogeneous virus–drug network, by connecting the virus–virus network and drug–drug network through virus–drug associations. In the heterogeneous network, there are a set of m viruses $V = \{v_1, v_2, v_3, \ldots, v_m\}$ and a set of n drugs $D = \{d_1, d_2, d_3, \ldots, d_n\}$; the edge between drugs (d_i, d_j) is weighted by the score $S_d(d_i, d_j)$ in the DDS matrix, the edge between viruses (v_a, v_b) is weighted by the score $S_v(v_a, v_b)$ in the VVS matrix, and the edge between drug d_i and virus v_a denotes the existence of association between them.

The VDA matrix B is extremely sparse due to the rarity of known virus–drug associations, where 1/0 denotes known/unknown virus–drug associations, respectively. We would like to fill out the missing values in the matrix as scores to predict unknown VDAs. The integration of information of DDSs, VVSs, and known VDAs into the heterogeneous network will benefit the discovery of unknown VDAs due to the intrinsic correlation among drugs and viruses.

VDA-GKSBMF model to predict virus–drug associations

To predict potential virus-drug associations of COVID-19, we define the VDA prediction as a problem of completing virus-drug matrix in a heterogeneous virus-drug network, and explore potential VDAs of COVID-19 by Gaussian kernel similarity bilinear matrix factorization (Yang M. et al., 2020; called as VDA-GKSBMF).

Matrix factorization is an effective method, which intends to calculate an optimal approximation to the target matrix by decomposing it into two low-rank matrices. In a word, the mathematical model of matrix factorization is formulated as

$$\min_{U,V} B - UV^T{}_F^2, \quad (1)$$

where $B \in \mathbb{R}^{n\times m}$ is the given incomplete matrix with n drugs and m viruses, $\mathrm{U} \in \mathbb{R}^{n\times k}$ and $\mathrm{V} \in \mathbb{R}^{m\times k}$ are the indicator feature matrices of B and k is the subspace dimensionality [$\mathrm{k} \ll \min(\mathrm{n},\mathrm{m})$], $\|\cdot\|_F$ denotes the Frobenius norm. Many algorithms have been designed to provide numerical solutions for the above model or alternative forms. However, compared with other algorithms, the classic ADMM algorithm is superior to solving our proposed matrix factorization model.

The elements in the association matrix B are either 0 or 1. Thus, the predicted values in the un-known entries are expected to be in the interval of [0, 1], where a predicted value closer to 1 indicates that this is likely to be an indication and vice versa. Nevertheless, in the above matrix completion model, the entries in the completed matrix can be any real value in $(-\infty, +\infty)$.

Moreover, based on the assumption that similar drugs share similar molecular pathways to treat similar viruses, the underlying factors that determine drug-virus associations are highly correlated. Since B is extremely rare and low rank, usually less than 1% of known associations are present, while the rest of the elements are unknown. Therefore, the error term is only computed on items with known associations. At the same time, Tikhonov regularization terms are often used to avoid overfitting. To achieve this, the matrix factorization model can be expressed as,

$$\min_{U,V} \frac{1}{2}\mathcal{P}\Omega \| \left(B - UV^T\right) \|_F^2 + \frac{\lambda_1}{2}\left(\| U \|_F^2 + \| V \|_F^2\right), \quad (2)$$

where Ω is a set containing index pairs (i,j) of all known entries in B and $\mathcal{P}_\Omega$ is the projection operator onto Ω, λ_1 is regularization parameter. However, the above objective function does not involve a large amount of prior information about viruses and drugs, such as disease similarity and drug similarity. Since U and V are matrices containing potential eigenvectors of drugs and viruses, given a drug similarity matrix Z_d and a virus similarity matrix Z_v, UU^T and VV^T are expected to match S_d and S_v, respectively. Therefore, model (2) is described as follows:

$$\min_{U,V} \frac{1}{2}\| \mathcal{P}_{\Omega}\left(B - UV^T\right)\|_F^2 + \frac{\lambda_1}{2}\left(\| U \|_F^2 + \| V \|_F^2\right) + \frac{\lambda_2}{2}\left(\| Z_d - UU^T \|_F^2 + \| Z_v - VV^T \|_F^2\right) \tag{3}$$

Model (3) deals with a single drug and virus similarity measure. Here, in order to integrate the Gaussian kernel similarity measure, we propose the VDA-GKSBMF model, which is expressed as follows:

$$\min_{U,V,P,Q,A} \frac{1}{2}\| \left(A - UV^T\right)\|_F^2 + \frac{\lambda_1}{2}\left(\| U \|_F^2 + \| V \|_F^2\right) + \frac{\lambda_2}{2}\left(\| S_d - UP^T \|_F^2 + \| S_v - VQ^T \|_F^2\right) + \frac{\lambda_3}{2}\left(\| P \|_F^2 + \| Q \|_F^2\right) \tag{4}$$

$$s.t\ \mathcal{P}_{\Omega}(A) = \mathcal{P}_{\Omega}(B)$$

$$U \geq 0, V \geq 0,$$

where S_d and S_v are matrices concatenating Gaussian kernel similarity measure of drug and virus, and λ_1, λ_2, and λ_3 are balancing parameters. A is an auxiliary matrix for facilitating optimization. The approximation of similarity matrix S_d and S_v are constructed based on characteristic matrices U and V, where P and Q are potential characteristic matrices representing drug similarity and virus similarity, respectively. We solve model (4) by ADMM framework. Introducing two riving matrices X and Y, model (4) is transformed into

$$\min_{U,V,P,Q,X,Y,A} \frac{1}{2}\| \left(A - UV^T\right)\|_F^2 + \frac{\lambda_1}{2}\left(\| U \|_F^2 + \| V \|_F^2\right) + \frac{\lambda_2}{2}\left(\| S_d - UP^T \|_F^2 + \| S_v - VQ^T \|_F^2\right) + \frac{\lambda_3}{2}\left(\| P \|_F^2 + \| Q \|_F^2\right) \tag{5}$$

$$s.t\ \mathcal{P}_{\odot}(A) = \mathcal{P}_{\odot}(B)$$

$$U = X, V = Y$$

$$X \geq 0, Y \geq 0.$$

The augmented Lagrangian function becomes

$$\mathrm{L} = \| \left(A - UV^T\right)\|_F^2 + \frac{\lambda_1}{2}\left(\| U \|_F^2 + \| V \|_F^2\right) + \frac{\lambda_2}{2}\left(\| S_d - UP^T \|_F^2 + \| S_v - VQ^T \|_F^2\right) + \frac{\lambda_3}{2}\left(\| P \|_F^2 + \| Q \|_F^2\right) + Tr\left(W^T(U - X)\right) + Tr\left(R^T(U - X)\right) + \frac{\rho}{2}\left(\| U - X \|_F^2 + \| V - Y \|_F^2\right) \tag{6}$$

where W and R are the Lagrange multiplier and ρ >0 is the penalty parameter. At the i-th iteration, it requires alternatively computing $U_{i+1}, V_{i+1}, P_{i+1}, Q_{i+1}, X_{i+1}, Y_{i+1}, A_{i+1}$.

Molecular docking method

Molecular docking method can be used to study the behavior of small molecules at the binding sites of target proteins. It has been widely used in drug design, since structures of more and more target proteins have been confirmed by experiments. AutoDock (Goodsell, 1996) is an open source molecular simulation software available to identify the conformation of a small molecule binding to a large molecule target. AutoDock has an affinity scoring function, which can sort candidate poses according to the sum of van der Waals and electrostatic energy. We used AutoDock to evaluate the molecular binding activity between predicted antiviral drugs and biomolecules.

Evaluation metrics

In this work, we evaluate the predictive performance of our method by 5-fold cross-validation. Popular evaluation metrics: AUC and AUPR are used to quantify the predictive performance of methods. Given a threshold of predictive scores, the candidate associations above this threshold are regarded as positives, and others are negatives. Then, true positive rate (TPR), false positive rate (FPR) and Precision can be calculated by,

$$\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{7}$$

$$\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP}+\mathrm{TN}) \tag{8}$$

$$\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{9}$$

where TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively. TPR is also called as Recall, which measures the ratio of correctly predicted positive samples to all positive samples. Precision measures the ratio of correctly predicted positive samples to all predicted positive samples.

With the increases of the threshold, TPR/Recall, FPR, and Precision will vary. TPR and FPR can form a TPR- FPR curve, called as the receiver-operating characteristic (ROC) curve. The area under the ROC curve is generally denoted as AUC. Precision and Recall (equivalent to TPR) can form a Precision–Recall (PR) curve. The area under the PR curve is generally denoted as AUPR. AUC and AUPR are scalar with the evaluation criterion: the larger AUC/AUPR is, the better the predictive performance is. AUC and AUPR can evaluate the overall performance of prediction algorithms.

Results

Parameter setting

In VDA-GKSBMF algorithm, there are tunable parameters $\gamma', \omega, \lambda_1, \lambda_2$ and λ_3. In order to prevent

multi-parameter overfitting, we set λ_1, λ_2 and λ_3 to the same value and remove two parameters. Because they are used to punish the related terms of U and V, P and Q in model (3) and model (4). VDA-GKSBMF has three parameters ($\gamma', \omega, \lambda_l$) needed to be determined. We first set γ' to 0.5, and then ω, λ_l are set in range of {0, 0.1, 0.2,…, 1}, {0.001, 0.01, 0.1, 1} by using the fivefold cross-validation on the training dataset. Table 2 displays the top 3 AUCS values as a function of $\gamma', \omega, \lambda_1, \lambda_2$ and λ_3 in five datasets.

Comparison with other methods

By 5-fold cross-validation experiment, we evaluate the performance of VDA-GKSBMF. We plot its ROC curve in Figure 2, and we find that it has a high AUC value in five datasets.

Further, we compare the VDA-GKSBMF method with other methods for drug repositioning: VDA-KATZ (Yang et al., 2019), IRNMF (Tang et al., 2020), VDA-GBNNR (Wang et al., 2021), and SCPMF (Meng et al., 2021). VDA-KATZ (Yang et al., 2019) used a KATZ algorithm to infer drug-virus association. The Indicator Regularized non-negative Matrix Factorization (IRNMF) method (Tang et al., 2020) introduced the indicator matrix and Karush-Kuhn-Tucker condition into the non-negative matrix factorization algorithm. VDA-GBNNR based on kernel similarity to predict anti-SARS-COV-2 drug. SCPMF used similarity constrained probabilistic matrix to infer drug-virus association. The experiment was carried out 50 times, with average performance as the final result. Table 3 shows sensitivities, specificities, accuracies, and AUCs of the five models on the five datasets. From Table 3, VDA-GBNNR obtains the best performance for other methods in dataset 1. However, VDA-GKSBMF achieves the best sensitivity, accuracy, specificity, and AUC on dataset 2, dataset 3, dataset 4, and dataset 5. Figure 2 displays the results of the methods in five datasets. The results show that the VDA-GKSBMF method outperforms the baseline methods in terms of the ROC curves and the corresponding AUC values, meaning that it can better discover antiviral drugs.

TABLE 2 The top three AUCs using different $\gamma', \omega, \lambda_1, \lambda_2$, and λ_3 values in 5-fold cross-validation.

Dataset	γ'	ω	λ_1	λ_2	λ_3	AUC
Dataset1	0.5	0.3	1	1	1	**0.8851**
	0.5	0.4	1	1	1	0.8825
	0.5	0.5	1	1	1	0.8663
Dataset2	0.5	0.1	0.1	0.1	0.1	**0.8594**
	0.5	0.2	0.1	0.1	0.1	0.8590
	0.5	0.3	0.1	0.1	0.1	0.8583
Dataset3	0.5	0.4	1	1	1	**0.8807**
	0.5	0.3	1	1	1	0.8793
	0.5	0.2	1	1	1	0.8756
Dataset4	0.5	0.2	0.1	0.1	0.1	**0.8824**
	0.5	0.3	0.1	0.1	0.1	0.8809
	0.5	0.4	0.1	0.1	0.1	0.8766
Dataset5	0.5	0.4	1	1	1	**0.8804**
	0.5	0.3	1	1	1	0.8789
	0.5	0.5	1	1	1	0.8787

Bold represented the best AUC values of different parameters in the same datasets.

Case study

After verifying the good performance of VDA-GKSBMF, to discover unknown antiviral drugs against SARS-CoV-2, we predict potential associations between SARS-CoV-2 and small molecule drugs based on known drug-virus association data, and we obtain the top-10 drugs with the highest score (see Table 4) in five datasets. Among the top-10 predicted drugs, there are 10 drugs that have been reported in the relevant literature, but the small molecule drugs were never confirmed to be anti-SARS-CoV-2 antiviral drugs. Ribavirin, Remdesivir, Oseltamivir, and Zidovudine were existed in at least four datasets.

Ribavirin is a road-spectrum antiviral drug that can inhibit the replication of respiratory syncytial virus (van Laarhoven and Marchiori, 2013). It can prevent respiratory syncytial virus infection in lung transplant recipients, and has been used to treat SARS-CoV and MERS-CoV. Similar to SARS-CoV and MERS-CoV, SARS-CoV-2 are a respiratory syndrome beta coronavirus that may cause severe respiratory diseases, and a few studies have reported that ribavirin may take an inhibitory effect on SARS-CoV-2 (Peng et al., 2020).

Remdesivir is a nucleoside analog with antiviral activity. Remdesivir has broad-spectrum activities against RNA viruses, such as SARS and MERS, and has been studied in a clinical trial for Ebola.

Oseltamivir is an antiviral neuraminidase inhibitor (Oseltamivir, n.d.) and has been used to prevent the infection of influenza A virus (for example, A-H1N1; Meijer et al., 2009, A-H5N1; De Jong et al., 2005, and influenza B virus). Oseltamivir can prevent the germination, replication, and infectivity of the virus in the host cell. More importantly, Oseltamivir combined with other drugs has been reported to inhibit the infection of SARS-CoV-2 (Huang et al., 2020).

Molecular docking

To further study the effectiveness of predicted drugs against SARS-CoV-2, the top 10 predicted small molecules are molecularly docked with SARS-CoV-2 spike protein/ACE2. From the DrugBank database, the chemical structures of these small molecule drugs have been obtained. The structure of spinous process protein of SARS-CoV-2 is calculated based on the homology model of Zhang lab (Wang et al., 2020). We used AutoDock, a bioinformatics tool, to conduct molecular docking between the predicted antiviral drug and SARS-CoV-2 spike

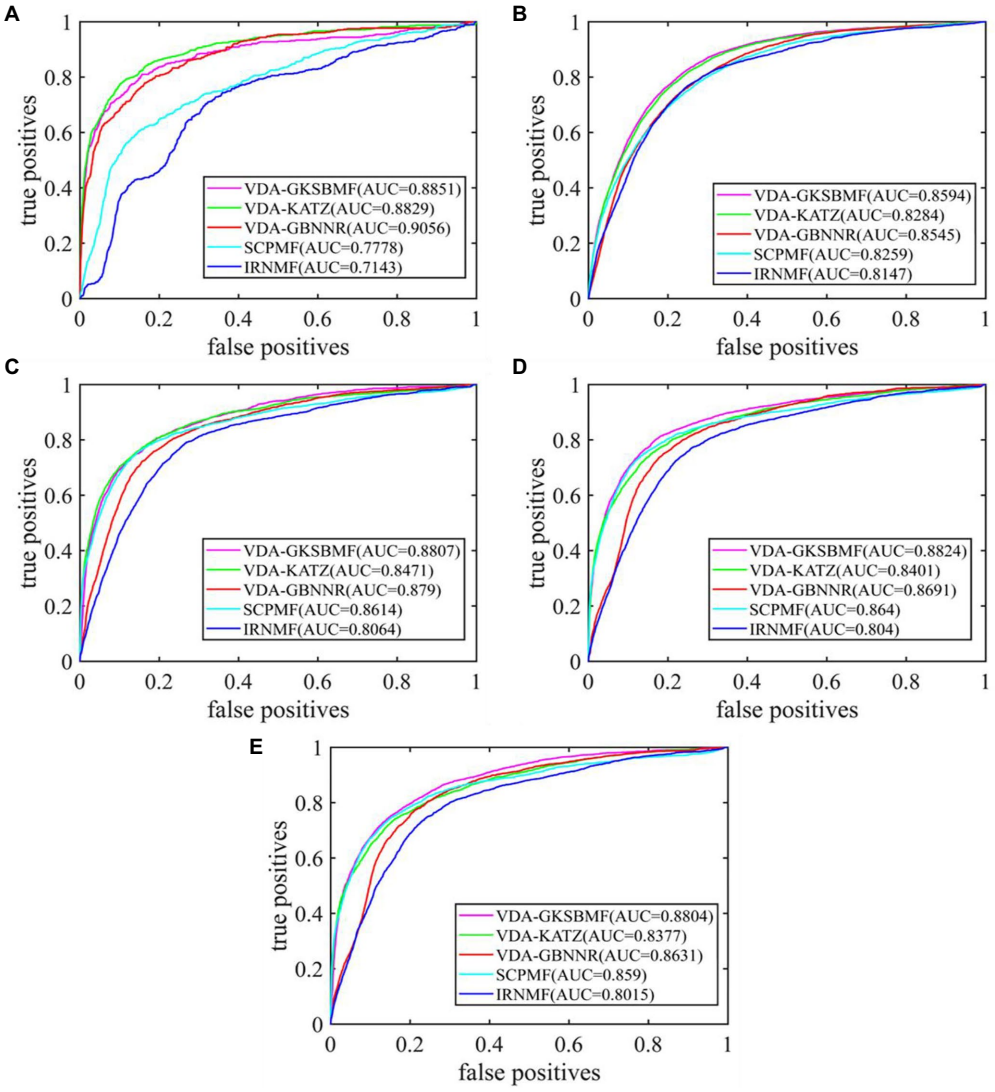


FIGURE 2
The performance of all methods in predicting virus–drug associations on five datasets: **(A)** Dataset1, **(B)** Dataset2, **(C)** Dataset3, **(D)** Dataset4, and **(E)** Dataset5.

protein/ACE2. The search algorithm scans the entire protein in AutoDock by genetic algorithm and grid box.

We calculate the predicted molecular binding energies of ribavirin, remdesivir, oseltamivir, and zidovudine small molecules with the spinous process protein and ACE2 of SARS-CoV-2 in Table 5. The results show that the binding activities of ribavirin with these two proteins are −5.29 and −6.39 kcal/mol, followed by remdesivir with −5.22 and −7.4 kcal/mol, and oseltamivir with −4.04 and − 4.73 kcal/mol. More importantly, ribavirin and remdesivir have been used to treat SARS, and their sequence homology with SARS-CoV-2 is about 79%.

Zidovudine has molecular binding energies of −6.54 and − 7.93 kcal/mol. Zidovudine is the drug which is an effective HIV replication inhibitor, which can improve immune function and partially reverse the neurological dysfunction caused by HIV. zidovudine, as an HIV nucleoside/nucleotide analogues reverse transcriptase inhibitor, has the potential to be a clue for SARS-COV-2 treatment.

Figures 3, 4 represent the docking results of four small molecules including ribavirin, remdesivir, oseltamivir, and zidovudine with two target proteins. The circles in each subgraph indicate the binding sites of the drug to the target protein. For example, the amino acids L387, L368, P565, and V209 are inferred to be the key residues for ribavirin binding to the SARS-CoV-2 spike protein/ACE2, while L849, T827, W1212, L144, and P504 are predicted as the key residues for remdesivir binding to these two target proteins.

TABLE 3 Performance indicators for different models.

Datasets	Methods	Accuracy	Sensitivity	Specificity	AUC
Dataset1	VDA-GKSBMF	0.5172	0.8757	0.5091	0.8851
	VDA-GBNNR	**0.5181**	**0.8957**	**0.5095**	**0.9056**
	VDA-KATZ	0.5171	0.8735	0.5090	0.8829
	SCPMF	0.5126	0.7708	0.5067	0.7778
	IRNMF	0.5098	0.7088	0.5052	0.7142
Dataset2	VDA-GKSBMF	**0.5136**	**0.8515**	**0.5072**	**0.8594**
	VDA-GBNNR	0.5134	0.8466	0.5071	0.8544
	VDA-KATZ	0.5125	0.8211	0.5066	0.8284
	SCPMF	0.5124	0.8187	0.5065	0.8259
	IRNMF	0.5120	0.8077	0.5063	0.8146
Dataset3	VDA-GKSBMF	**0.5097**	**0.8748**	**0.5052**	**0.8807**
	VDA-GBNNR	0.5097	0.8731	0.5051	0.8790
	VDA-KATZ	0.5089	0.8416	0.5047	0.8471
	SCPMF	0.5093	0.8557	0.5049	0.8613
	IRNMF	0.5079	0.8015	0.5042	0.8063
Dataset4	VDA-GKSBMF	**0.5102**	**0.8763**	**0.5054**	**0.8824**
	VDA-GBNNR	0.5098	0.8631	0.5052	0.8691
	VDA-KATZ	0.5091	0.8345	0.5048	0.8400
	SCPMF	0.5097	0.8581	0.5051	0.8639
	IRNMF	0.5081	0.7990	0.5044	0.8040
Dataset5	VDA-GKSBMF	**0.5101**	**0.8743**	**0.5054**	**0.8804**
	VDA-GBNNR	0.5096	0.8572	0.5051	0.8630
	VDA-KATZ	0.5090	0.8322	0.5048	0.8376
	SCPMF	0.5095	0.8532	0.5051	0.8590
	IRNMF	0.5081	0.7966	0.5043	0.8015

Bold represented the best value of different methods under the same evaluation condition.

TABLE 4 The predicted top-10 antiviral drugs against SARS-CoV-2 in five datasets.

Dataset1-drug	Dataset2-drug	Dataset3-drug	Dataset4-drug	Dataset5-drug
Remdesivir	Favipiravir	**Ribavirin**	Nitazoxanide	**Ribavirin**
Oseltamivir	**Remdesivir**	Nitazoxanide	**Ribavirin**	Chloroquine
Zanamivir	Cidofovir	Chloroquine	**Oseltamivir**	**Zidovudine**
ribavirin	**ribavirin**	Camostat	Camostat	Camostat
Laninamivir	Mycophenolic acid	Umifenovir	**Zidovudine**	Umifenovir
Peramivir	Navitoclax	**Remdesivir**	Favipiravir	Favipiravir
Presatovir	Itraconazole	**Zidovudine**	Hexachlorophene	Rifamycin
zidovudine	BCX4430 (Galidesivir)	Berberine	**Remdesivir**	**Oseltamivir**
Mycophenolic acid	Pleconaril	Amantadine	Sirolimus	Berberine
Mizoribine	Cyclosporine	**Oseltamivir**	Suramin	Niclosamide

Bold indicated that the drug existed in at least four datasets.

TABLE 5 The molecular binding energies between the predicted 4 antiviral drugs and two target proteins at least four datasets.

Drugs	Binding energies of target proteins	
	Spike protein	ACE2
Ribavirin	−5.29	−6.39
Remdesivir	−5.22	−7.40
Oseltamivir	−4.04	−4.73
Zidovudine	−6.54	−7.93

Discussion

Severe acute respiratory syndrome coronavirus 2 is quickly diffusing throughout the world, and it is urgent to find effective treatments against this virus. Drug repositioning, seeking to find new uses, offers a new strategy for the treatment of SARS-COV-2. However, to date, only a few databases have collated relevant drugs that may be used to treat SARS-COV-2. Thus, we developed a drug-virus as well

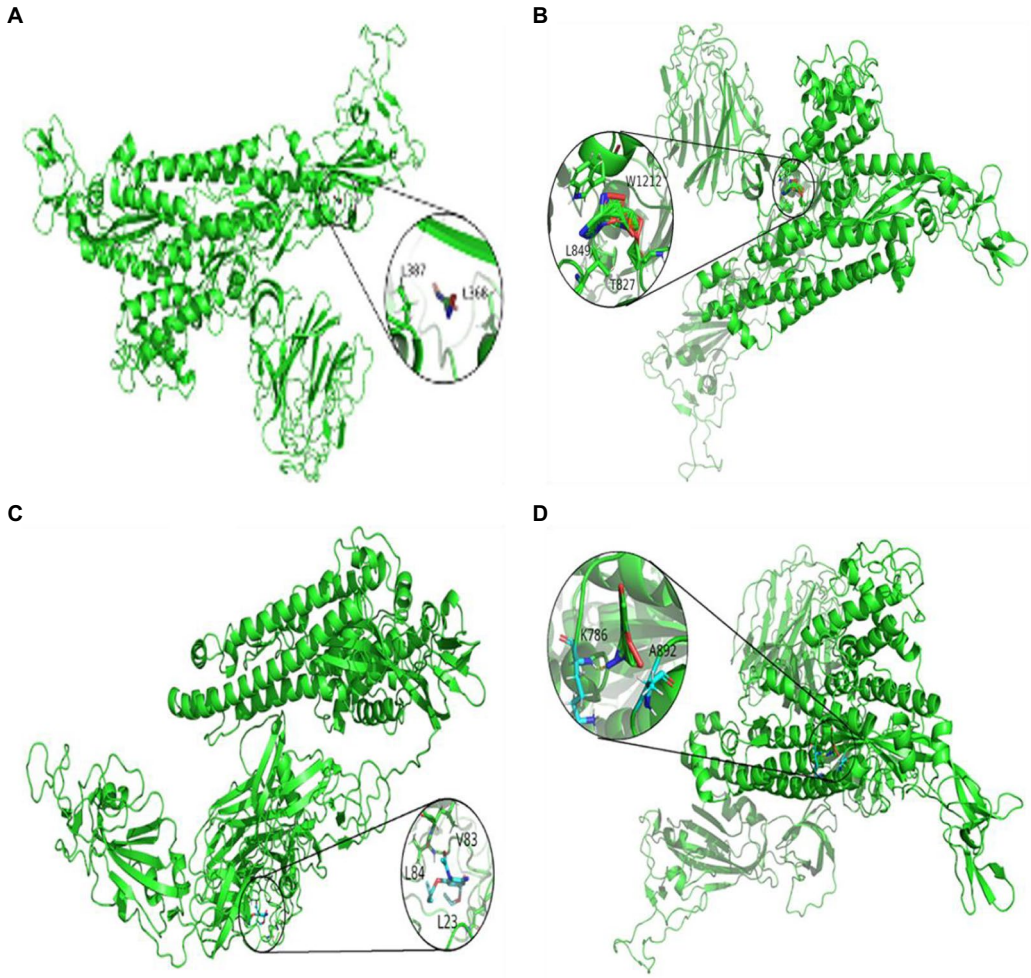


FIGURE 3
Molecular docking between the spike protein and four drugs: **(A)** ribavirin, **(B)** remdesivir, **(C)** oseltamivir, and **(D)** zidovudine.

as a method VDA-GKSBMF to prioritize drugs against SARS-COV-2.

Specifically, VDA-GKSBMF has a high AUC in cross-validation, which is better than other state-of-art methods in four datasets. We measured the molecular binding activity between predicted antiviral drugs and SARS-CoV-2 spike protein/human ACE2 (Zhao et al., 2020). Among them, the molecular binding energies between ACE2 and the four drugs were: Ribavirin (−6.39 kcal/mol), Remdesivir (−7.4 kcal/mol), Oseltamivir (−4.73 kcal/mol), zidovudine (−7.93 kcal/mol), and the four drugs have been in clinical trials or supported in recent publications. The results suggest that the VDA-GKSBMF algorithm can effectively infer unknown drugs of SARS-COV-2.

However, there a few limitations of this study. First, due to the limited size of the current virus-drug dataset and the complexity of intrinsic relationship in biomedical data, VDA-GKSBMF still has room for further improvement. On the one hand, we would like to expand the virus-drug dataset by including more virus-related and drug-related information, so as to further improve the predictive power of mining hidden virus-drug associations. On the other hand, it is also possible to enhance the ability of discovering potential drugs against SARS-COV-2 by more advanced and methods in related fields (Xu et al., 2020b; Xiang et al., 2021b, 2022a; Meng et al., 2022). Second, though we performed literature mining and molecular docking to validate our results, they are all in-silico methods. The prioritized drugs should be validated using wet-lab experiments. However, it is out of the scope of this study.

Conclusion

In this study, we collected five virus-drug datasets including VDAs matrix, virus genomic sequence similarity matrix, and drug chemical structure similarity matrix and explored drug repositioning of SARS-COV-2 by a novel method called VDA-GKSBMF.VDA-GKSBMF combined Gaussian similarity and extracted useful features to deduce potential virus-drug

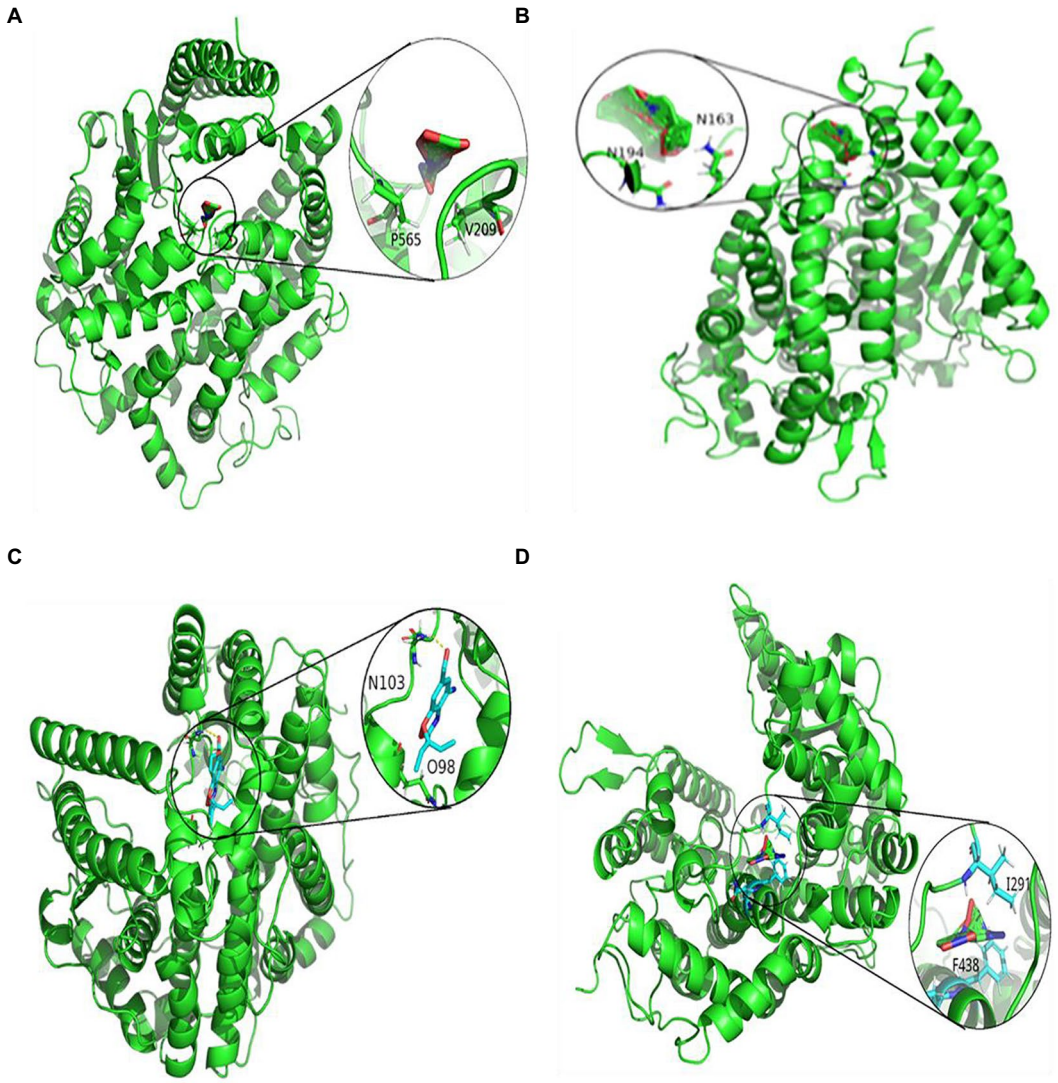


FIGURE 4
Molecular docking between ACE2 and four drugs: **(A)** ribavirin, **(B)** remdesivir, **(C)** oseltamivir, and **(D)** zidovudine.

associations. It combined Gaussian similarity and virus-drug association into the target function. The non-negative constraint was used in VDA-GKSBMF, ensuring that the predicted scores of association matrix were non-negative for the biological interpretability. Our results showed that VDA-GKSBMF is an effective approach for discovering new drugs of SARS-COV-2. In the future, we will combine different data resources to create larger dataset and design integrated algorithm, integrating multiple heterogeneous network and multiple similarities for predicting potential virus-drug associations.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/xiangju0208/VDA_GMSBMF.

Author contributions

BH and JH contributed to conception and design of the study. YW and JX organized the data and the prediction model. MT, RH, CL, and GT performed the statistical analysis. YW, JX, MB, JH, and BH wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Training Program for Excellent Young Innovators of Changsha (Grant Nos. kq1802024, kq1905045, kq2009093, and kq2106075), Hunan key laboratory cultivation base of the research and development of novel pharmaceutical preparations (No. 2016TP1029), Hunan Provincial Innovation Platform and Talents Program (No. 2018RS3105), the Foundation of Hunan Educational Committee (Grant No. 19A060), and the

Provincial key R & D projects of Hunan Provincial Science and Technology Department (No. 2022SK2074). This research was funded by the Natural Science Foundation of Hunan province (No. 2018JJ2461), the Project to Introduce Intelligence from Oversea Experts to Changsha City (Grant No. 2089901), and General project of Education Department of Hunan Province (Grant No. 19C0190), and supported by the special fund of "Young and Middle-aged Key Teachers Training Program" of Changsha Medical College, the National Natural Science Foundation of China (32002235).

Conflict of interest

RH and GT are employed by Genesis (Beijing) 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.

References

Cheng, L., Han, X., Zhu, Z., Qi, C., Wang, P., and Zhang, X. (2021a). Functional alterations caused by mutations reflect evolutionary trends of SARS-CoV-2. *Brief. Bioinform.* 22, 1442–1450. doi: 10.1093/bib/bbab042

Cheng, L., Zhu, Z., Wang, C., Wang, P., He, Y. O., and Zhang, X. (2021b). COVID-19 induces lower levels of IL-8, IL-10, and MCP-1 than other acute CRS-inducing diseases. *Proc. Natl. Acad. Sci. U. S. A.* 118:e2102960118. doi: 10.1073/pnas.2102960118

Cohain, A. T., Barrington, W. T., Jordan, D. M., Beckmann, N. D., Argmann, C. A., Houten, S. M., et al. (2021). An integrative multiomic network model links lipid metabolism to glucose regulation in coronary artery disease. *Nat. Commun.* 12:547. doi: 10.1038/s41467-020-20750-8

Coronaviridae Study Group of the International Committee on Taxonomy of V (2020). The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* 5, 536–544. doi: 10.1038/s41564-020-0695-z

De Jong, M. D., Tran, T. T., Truong, H. K., Vo, M. H., Smith, G. J., Nguyen, V. C., et al. (2005). Oseltamivir resistance during treatment of influenza a (H5N1) infection. *N. Engl. J. Med.* 353, 2667–2672. doi: 10.1056/NEJMoa054512

Eurosurveillance editorial team (2020). Note from the editors: World Health Organization declares novel coronavirus (2019-nCoV) sixth public health emergency of international concern. *Eur. Secur.* 25:200131e. doi: 10.2807/1560-7917.ES.2020.25.5.200131e

Goodsell, D. S. (1996). Automated docking of flexible ligands: Applications of autodock molecular recognition.

Gralinski, L. E. (2020). Menachery VD: return of the coronavirus: 2019-nCoV. *Viruses* 12:135. doi: 10.3390/v12020135

He, B., Wang, K., Xiang, J., Bing, P., Tang, M., Tian, G., et al. (2022). DGHNE: network enhancement-based method in identifying disease-causing genes through a heterogeneous biomedical network. *Brief. Bioinform.* 23:bbac405. doi: 10.1093/bib/bbac405

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395, 497–506. doi: 10.1016/S0140-6736(20)30183-5

Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* 9, 286–298. doi: 10.1093/bib/bbn013

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., et al. (2016). PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–D1213. doi: 10.1093/nar/gkv951

Landrum, G. (2014). RDKit:open-source cheminformatics. Release 2014.03.1.

Li, J., Wang, X., Li, N., Jiang, Y., Huang, H., Wang, T., et al. (2020). Feasibility of mesenchymal stem cell therapy for COVID-19: a mini review. *Curr. Gene Ther.* 20, 285–288. doi: 10.2174/1566523220999200820172829

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. *Front. Cell Dev. Biol.* 9:619330. doi: 10.3389/fcell.2021.772380

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression. *Mol. Ther. Nucleic Acids* 21, 676–686. doi: 10.1016/j.omtn.2020.07.003

Liu, X., Yang, J., Zhang, Y., Fang, Y., Wang, F., Wang, J., et al. (2016). A systematic study on drug-response associated genes using baseline gene expressions of the cancer cell line encyclopedia. *Sci. Rep.* 6:22811. doi: 10.1038/srep22811

Lu, K., Wang, F., Ma, B., Cao, W., Guo, Q., Wang, H., et al. (2021). Teratogenic toxicity evaluation of bladder cancer-specific oncolytic adenovirus on mice. *Curr. Gene Ther.* 21, 160–166. doi: 10.2174/1566523220999201217161258

Meijer, A., Lackenby, A., Hungnes, O., Lina, B., Van-Der-Werf, S., Schweiger, B., et al. (2009). On behalf of the European influenza surveillance scheme: oseltamivir-resistant influenza virus a (H1N1), Europe, 2007-08 seasona. *Emerg. Infect. Dis.* 15, 552–560. doi: 10.3201/eid1504.181280

Meng, Y., Jin, M., Tang, X., and Xu, J. (2021). Drug repositioning based on similarity constrained probabilistic matrix factorization: COVID-19 as a case study. *Appl. Soft Comput.* 103:107135. doi: 10.1016/j.asoc.2021.107135

Meng, Y., Lu, C., Jin, M., Xu, J., Zeng, X., and Yang, J. (2022). A weighted bilinear neural collaborative filtering approach for drug repositioning. *Brief. Bioinform.* 23:bbab581. doi: 10.1093/bib/bbab581

Novac, N. (2013). Challenges and opportunities of drug repositioning. *Trends Pharmacol. Sci.* 34, 267–272. doi: 10.1016/j.tips.2013.03.004

Oseltamivir (n. d.). Oseltamivir: Description Available at: https://www.drugbank.ca/drugs/DB00198

Parsza, C. N., Gomez, D. L. M., Simonin, J. A., Nicolas Belaich, M., and Ghiringhelli, P. D. (2021). Evaluation of the Nucleopolyhedrovirus of Anticarsia gemmatalis as a vector for gene therapy in mammals. *Curr. Gene Ther.* 21, 177–189. doi: 10.2174/1566523220999201217155945

Peng, L., Tian, X., Shen, L., Kuang, M., Li, T., Tian, G., et al. (2020). Identifying effective antiviral drugs against SARS-CoV-2 by drug repositioning through virus-drug association prediction. *Front. Genet.* 11:577387. doi: 10.3389/fgene.2020.577387

Shen, L., Liu, F., Huang, L., Liu, G., Zhou, L., and Peng, L. (2022). VDA-RWLRLS: an anti-SARS-CoV-2 drug prioritizing framework combining an unbalanced bi-random walk and Laplacian regularized least squares. *Comput. Biol. Med.* 140:105119. doi: 10.1016/j.compbiomed.2021.105119

Tang, X., Cai, L., Meng, Y., Xu, J., Lu, C., and Yang, J. (2020). Indicator regularized non-negative matrix factorization method-based drug repurposing for COVID-19. *Front. Immunol.* 11:603615. doi: 10.3389/fimmu.2020.603615

van Laarhoven, T., and Marchiori, E. (2013). Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PLoS One* 8:e66952. doi: 10.1371/journal.pone.0066952

Wang, J., Wang, C., Shen, L., Zhou, L., and Peng, L. (2021). Screening potential drugs for COVID-19 based on bound nuclear norm regularization. *Front. Genet.* 12:817672. doi: 10.3389/fgene.2021.817672

Wang, F., Yang, J., Lin, H., Li, Q., Ye, Z., Lu, Q., et al. (2020). Improved human age prediction by using gene expression profiles from multiple tissues. *Front. Genet.* 11:1025. doi: 10.3389/fgene.2020.01025

Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L., et al. (2004). Database resources of the National Center for biotechnology information: update. *Nucleic Acids Res.* 32, 35D–340D. doi: 10.1093/nar/gkh073

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

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

Xiang, J, Meng, X, Zhao, Y, Wu, F-X, and Li, M. (2022a). HyMM: hybrid method for disease-gene prediction by integrating multiscale module structure. *Brief. Bioinform.* doi: 10.1093/bib/bbac072 [Epub ahead of print].

Xiang, J., Zhang, N.-R., Zhang, J.-S., Lv, X.-Y., and Li, M. (2021a). PrGeFNE: predicting disease-related genes by fast network embedding. *Methods* 192, 3–12. doi: 10.1016/j.ymeth.2020.06.015

Xiang, J, Zhang, J, Zhao, Y, Wu, F-X, and Li, M. (2022b). Biomedical data, computational methods and tools for evaluating disease–disease associations. *Brief. Bioinform.* doi: 10.1093/bib/bbac006 [Epub ahead of print].

Xiang, J., Zhang, J., Zheng, R., Li, X., and Li, M. (2021b). NIDM: network impulsive dynamics on multiplex biological network for disease-gene prediction. *Brief. Bioinform.* 22:bbab080. doi: 10.1093/bib/bbab080

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020a). CMF-impute: an accurate imputation tool for single-cell RNA-seq data. *Bioinformatics* 36, 3139–3147. doi: 10.1093/bioinformatics/btaa109

Xu, J., Zhu, W., Cai, L., Liao, B., Meng, Y., Xiang, J., et al. (2020b). LRMCMDA: predicting miRNA-disease association by integrating low-rank matrix completion with miRNA and disease similarity information. *IEEE Access* 8, 80728–80738. doi: 10.1109/ACCESS.2020.2990533

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., et al. (2022). Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput. Struct. Biotechnol. J.* 20, 333–342. doi: 10.1016/j.csbj.2021.12.028

Yang, M., Luo, H., Li, Y., and Wang, J. (2019). Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* 35, i455–i463. doi: 10.1093/bioinformatics/btz331

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020). Human geroprotector discovery by targeting the converging subnetworks of aging and age-related diseases. *Geroscience* 42, 353–372. doi: 10.1007/s11357-019-00106-x

Yang, M., Wu, G., Zhao, Q., Li, Y., and Wang, J. (2020). Computational drug repositioning based on multi-similarities bilinear matrix factorization. *Brief. Bioinform.* 22:bbaa267. doi: 10.1093/bib/bbaa267

Yao, Y., Li, X., Liao, B., Huang, L., He, P., Wang, F., et al. (2017). Predicting influenza antigenicity from Hemagglutintin sequence data based on a joint random forest method. *Sci. Rep.* 7:1545. doi: 10.1038/s41598-017-01699-z

Zhang, Y., Huang, H., Zhang, D., Qiu, J., Yang, J., Wang, K., et al. (2017). A review on recent computational methods for predicting noncoding RNAs. *Biomed. Res. Int.* 2017:9139504. doi: 10.1155/2017/9139504

Zhang, Y., Xiang, J., Tang, L., Li, J., Lu, Q., Tian, G., et al. (2021). Identifying breast cancer-related genes based on a novel computational framework involving KEGG pathways and PPI network modularity. *Front. Genet.* 12:596794. doi: 10.3389/fgene.2021.809608

Zhang, W., Zhang, H., Yang, H., Li, M., Xie, Z., and Li, W. (2019). Computational resources associating diseases with genotypes, phenotypes and exposures. *Brief. Bioinform.* 20, 2098–2115. doi: 10.1093/bib/bby071

Zhao, Y, Zhao, Z, Wang, Y, Zhou, Y, Ma, Y, and Zuo, W. (2020). Single-cell RNA expression profiling of ACE2, the receptor of SARS-COV-2. *Am. J. Respir. Crit. Care. Med.* 202, 756–759. doi: 10.1164/rccm.202001-0179LE

Zhou, L., Wang, J., Liu, G., Lu, Q., Dong, R., Tian, G., et al. (2020). Probing antiviral drugs against SARS-CoV-2 through virus-drug association prediction based on the KATZ method. *Genomics* 112, 4427–4434. doi: 10.1016/j.ygeno.2020.07.044

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

Zhu, Z., Zhang, S., Wang, P., Chen, X., Bi, J., Cheng, L., et al. (2021). A comprehensive review of the analysis and integration of omics data for SARS-CoV-2 and COVID-19. *Brief. Bioinform.* 23:bbab446. doi: 10.1093/bib/bbab302

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

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 19 December 2022
DOI 10.3389/fmicb.2022.1085086

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Zheng Yan,
Henan Provincial Cancer Hospital, China
Tao Tan,
Kunming University of Science and Technology, China

*CORRESPONDENCE
Yuanyuan Han
✉ hyy@imbcams.com.cn
Yan Zhang
✉ 2817621@qq.com
Libin Shi
✉ 1992246775@qq.com

†These authors have contributed equally to this work

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

RECEIVED 31 October 2022
ACCEPTED 02 December 2022
PUBLISHED 19 December 2022

CITATION
Huang Z, Yi L, Jin L, Chen J, Han Y, Zhang Y and Shi L (2022) Systematic analysis of virus nucleic acid sensor DDX58 in malignant tumor. *Front. Microbiol.* 13:1085086. doi: 10.3389/fmicb.2022.1085086

Systematic analysis of virus nucleic acid sensor DDX58 in malignant tumor

Zhijian Huang[1†], Limu Yi[2†], Liangzi Jin[3], Jian Chen[1], Yuanyuan Han[3*], Yan Zhang[2,4*] and Libin Shi[5*]

[1]Department of Breast Surgical Oncology, Clinical Oncology School of Fujian Medical University, Fujian Cancer Hospital, Fuzhou, China, [2]Department of Pathology, The First Affiliated Hospital of Guangdong University of Pharmacy, Guangzhou, China, [3]Institute of Medical Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Kunming, China, [4]Department of Pathology, Maternity and Child Healthcare Hospital of Longhua District, Shenzhen, China, [5]Department of Nuclear Medicine, Clinical Oncology School of Fujian Medical University, Fujian Cancer Hospital, Fuzhou, China

Introduction: In December 2019, a novel epidemic of coronavirus pneumonia (COVID-19) was reported，and population-based studies had shown that cancer was a risk factor for death from COVID-19 infection. However, the molecular mechanism between COVID-19 and cancer remains indistinct. In this paper, we analyzed the nucleic acid sensor (*DDX58*) of SARS-CoV-2 virus, which is a significant gene related to virus infection. For purpose of clarifying the characteristics of *DDX58* expression in malignant tumors, this study began to systematically analyze the *DDX58* expression profile in the entire cancer type spectrum.

Methods: Using TCGA pan-cancer database and related data resources, we analyzed the expression, survival analysis, methylation expression, mutation status, microsatellite instability (MSI), immune related microenvironment, gene related network, function and drug sensitivity of *DDX58*.

Results: The expression level of *DDX58* mRNA in most cancers was higher than the expression level in normal tissues. Through TIMER algorithm mining, we found that *DDX58* expression was closely related to various levels of immune infiltration in pan-cancer. The promoter methylation level of *DDX58* was significantly increased in multiple cancers. In addition, abnormal expression of *DDX58* was related to MSI and TMB in multiple cancers, and the most common type of genomic mutation was "mutation." In the protein–protein interaction (PPI) network, we found that type I interferon, phagocytosis, ubiquitinase, and tumor pathways were significantly enriched. Finally, according to the expression of *DDX58* indicated potential sensitive drugs such as Cediranib, VE–821, Itraconazole, JNJ–42756493, IWR–1, and Linsitinib.

Discussion: In conclusion, we had gained new insights into how *DDX58* might contribute to tumor development, and *DDX58* could be used as an immune-related biomarker and as a potential immunotherapeutic target for COVID-19 infected cancer patients.

KEYWORDS

***DDX58*, pan-cancer, biomaker, immune infiltration, SARS-CoV-2**

1. Introduction

Globally, 590 million cases of COVID-19 infection and 6.4 million deaths have been reported as of August 15, 2022. A recent study found that about 66 cancer patients were immunosuppressed with increasing susceptibility to infection and risk of serious complications (Al-Quteimat and Amer, 2020). Compared with other diseases, the genomes of 68 cancers have been fully studied. However, the gene information associated with COVID-19 remains largely unknown.

Genome-wide association study on COVID-19 patients with severe and critical illness showed that *DDX58* gene was closely associated to severe COVID-19. It is urgent to study the role of this gene in different cancers. RIG-I or DExD/H-box helicase 58 (*DDX58*) is a protein that recognizes viral double-stranded RNA and produces type I interferon, an antiviral and innate immune response medium as previous described (Morelli et al., 2021). At the same time, *DDX58* is considered as a potential novel target for COVID-19 treatment and a key component of COVID-19 infection and progress (Yamada et al., 2021). In cancer patients exposed to viruses, their condition worsened and their mortality increased (Han et al., 2021). Therefore, we aimed to find the role of *DDX58* in cancer immunotherapy, in order to provide a more suitable treatment idea for cancer patients infected with COVID-19.

Here, we showed the landscape analysis of *DDX58* expression level in healthy tissues and pan-cancer tissues using GTEx and TCGA, and then studied the relationship between *DDX58* and various tumor prognoses. We explored the relationship between *DDX58* and immune cell infiltration in specific tumor patients and studied the potential role of *DDX58* in tumor patients, we also analyzed the methylation profile of *DDX58* promoter and the mutation of *DDX58* in the UALCAN database. These findings might have important significance in preventing SARS CoV-2 infection and mitigate cytokine storm in patients infected with cancer. This study might also point out the therapeutic potential of *DDX58* inhibitors in preventing or mitigating SARS CoV-2 infection in specific cancer patients.

2. Materials and methods

2.1. Transcriptome data analysis

TCGA database and genotypic tissue expression (GTEx) database were used to obtain gene expression profiles. An analysis of 31 normal tissues was performed using mRNA data obtained from the GTEx project. Cancer cell lines were analyzed in 31 tissues according to their expression levels, and then the Kruskal Wallis test was performed on the mRNA data between adjacent tissues and tumor tissues, as well as healthy tissues and tumor tissues, to determine the difference of *DDX58* expression. *DDX58* expression levels were compared between healthy tissues and tumor tissues, as well as between adjacent tumor tissues and tumor tissues.

HPA[1] contains normal tissue and tumor tissue protein levels of human gene expression profile information. In this study, we compared the expression of *DDX58* protein in normal tissues and cancer tissues of four different organs by HPA. The significance of the difference was calculated using the Wilcoxon test. $p < 0.05$ suggests that the expression of tumor tissue is different from that of normal tissue.

2.2. Clinical relevance analysis

The expression level of *DDX58* was examined using univariate COX regression analysis to determine whether it was associated with tumor prognosis in various cancers. According to the median of *DDX58*, samples were divided into two groups based on their expression levels: high-and low-expression groups of *DDX58*. In order to determine the importance of survival differences, a log rank test was used, with a threshold of $p = 0.05$. What's more, we used the limma package to learn the relationship between *DDX58* and T stage in pan-cancer.

2.3. Construction and enrichment analysis of gene-gene, protein-protein and gene-disease networks

We constructed gene–gene interaction network through GeneMANIA[2] and built PPI network through STRING database.[3] We had further constructed a gene disease network on the OPENTARGET platform. Gene ontology (GO) terminology, Kyoto Encyclopedia of Genes and Genomes (KEGG) and GSEA were used to gene enrichment analysis. The term "GO" refers to molecular function (MF), cellular components (CC), and biological processes (BP). Use the "ClusterProfiler" package to perform GO, KEGG analysis, and GSEA. The TIMER[4] was a comprehensive online database, analysis of a wide variety of cancer types related to immune infifiltrating. In this study, we used TIMER to determine the relationship between *DDX58* expression and ACE2.

2.4. Epigenetic methylation analysis and association analysis of methyltransferase

As a form of DNA chemical modification, DNA methylation controls gene expression by changing epigenetics without changing DNA sequence. To analyze the methylation level of tumor and normal tissues, we obtained them from the methylation

1 https://proteinatlas.org/
2 https://genemania.org/
3 https://string-db.org/
4 https://cistrome.shinyapps.io/timer/

module of UALCAN database. Later, from UCSC[5] database, we have downloaded a standardized pan-cancer dataset: TCGA Pan Cancer (PANCAN, $N = 10{,}535$, $G = 60{,}499$), from which we further extracted the expression data of *DDX58* gene and 44 marker genes of three kinds of RNA modified m6A genes in each sample. We filtered the samples from: Primary Blood Derived Cancer-Peripheral Blood, Primary Tumor. Further, log2 $(x + 1)$ transformation has been performed for each expression value. Next, we had calculated the spearman correlation between *DDX58* and marker genes of five different immune pathways.

2.5. Analysis of tumor mutation load and genome changes in pan-cancer

The total number of substitutions, insertions and deletions per megabase in the coding region of tumor gene exons was used to calculate the tumor mutation load (TMB). We got the expression data of *DDX58* gene in every sample from the previously downloaded datasets, combined with the previously screened samples. In addition, we also had download the Simple Nucleotide Variation dataset of level4 of all TCGA samples processed through MuTect2 software from GDC (https://portal.gdc.cancer.gov/; Beroukhim et al., 2010). To calculate the tumor mutation burden (TMB), we used the TMB function of the R software package maftools (version 2.8.01). Then we integrated the TMB and gene expression data of the samples. Finally, we obtained the expression data of 37 cancer species after removing those with fewer than three samples in a single cancer species. Through cBioPortal resources,[6] we had analyzed the genetic changes of *DDX58* in the TCGA dataset (Reimer et al., 2021). The gene changes and mutation sites of *DDX58* were obtained in the "Oncoprint," "Cancer Type Summary," and "Mutations" sub modules.

2.6. Analysis of immune checkpoint genes and new immune antigens

Biological phenomena such as gene fusion, deletion mutation and point mutation are called new antigens encoded by mutated genes in tumor cells. We had calculated the binding affinity score of epitopes with 8–11 amino acids of a certain length and the epitopes with a score less than 500 nm were defined as new antigens. Then, we ranked the predicted new antigens according to antigenicity index value, affinity and mutation allele frequency. In each tumor sample, scannedo was used to count the new antigens and analyze the relationship between *DDX58* expression and new antigens. The immune checkpoint genes had been extracted and analyzed along with the *DDX58* expression to further investigate their relationship.

2.7. *DDX58* expression and microsatellite instability analysis

From UCSC (see Footnote 5) database we had downloaded a standardized pan-cancer dataset: TCGA Pan Cancer (PANCAN, $N = 10{,}535$, $G = 60{,}499$). Based on the previously extracted expression data and screened samples, we obtained MSI (Microsatellite instance) scores of each tumor from the previous study (Gounder et al., 2022). Next, the MSI and gene expression data of the samples were integrated.

2.8. Immune infiltration analysis

We screened the metastatic samples from the following sources: Primary Blood Derived Cancer - Peripheral Blood (TCGA-LAML), Primary Tumor, and TCGA-SKCM. The gene expression profiles of each tumor were extracted, mapped to GeneSymbol, and further analyzed using the Timer method of the R software package IOBR (version 0.99.9, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8283787/; Li et al., 2017). The B cell, T cell CD4, T cell CD8, Neutrophil, Macrophage and DC infiltration scores of each patient in each tumor were reevaluated according to gene expression.

2.9. Drug sensitivity of *DDX58* in pan-cancer

To study the drug sensitivity of pan-cancer patients to *DDX58*, the CallMinerTM database was used[7] to get activity data and RNA seq expression profile of NCI-60 compounds. In order to analyze and select drugs approved by FDA or clinical trials, R packages "impute," "limma," "ggplot2," and "ggpubr" were used for analysis.

3. Results

3.1. Differential *DDX58* expression analysis in pan-cancer tissues and normal tissues

The analysis of gene disease network interaction showed that *DDX58* was mainly related to genetic, familial or genetic disease, immune system disease, infectious disease, benign tumor, etc. In particular, *DDX58* had a certain relationship with benign tumor (Figure 1A). Subsequently, we investigated the role of human *DDX58* expression in pan-cancer. A comparison of the expression levels of *DDX58* in tumors and normal tissues was performed using the TCGA database. As compared to normal tissues, *DDX58* was found to be highly expressed in BRCA, ESCA, STES, KIPAN,

5 https://xenabrowser.net/
6 http://www.cbioportal.org/
7 https://discover.nci.nih.gov/cellminer

STAD, HNSC, KIRC, LIHC, CHOL, while it was low expressed in LUAD, COAD, READ, KIRP, LUSC, KICH (Figures 1B,C). At the same time, *DDX58* protein levels in four different organs and tissues also showed significant differences (Figure 1D).

3.2. Pan-cancer analysis of prognostic value of *DDX58* expression in different stages of cancers

Next, an analysis of *DDX58* expression and cancer prognosis was conducted using univariate Cox regression. According to the forest map of pan-cancer, the expression of *DDX58* had a significant impact on the OS of LGG, KIRC, SKCM, MESO, TGCT, PAAD, LUAD patients (Figure 2). In addition, we also analyzed the expression of *DDX58* in different cancer T stages, and the results showed that there were significant differences in the expression of *DDX58* in different stages of 10 cancers (Supplementary Figure S1).

3.3. Construction of *DDX58* gene, protein, disease network correlation with SARS CoV-2 receptor–ACE2

In order to understand the *DDX58* related network, the STRING and GeneMANIA protein-protein and gene-gene interaction networks that interact with *DDX58* were used (20 potential related genes were selected respectively; Figures 3B,C). We obtained 8 genes from the intersection of two data sets (Figure 3A) and carried out GO and KEGG analysis on 9 genes including *DDX58* (Figure 3D). We found that BP was enriched in negative regulation of type I interchange production, regulation of type I interchange production, type I interchange production. CC was mainly enriched in phagophore assembly site membrane, phagophore assembly site, phagocytic vascular membrane. MF was significantly enriched in protein tag, Lys63 specific dehydrogenase activity, and Lys48 specific dehydrogenase activity. KEGG analysis showed that many related pathways were significantly enriched, including RIG-I-like receiver signaling pathway, NF kappa B signaling pathway, Influenza A. In addition, it could be seen from the correlation analysis with AEC2 (SARS CoV-2 receptor) that there was a positive correlation between the expression of *DDX58* and ACE2 in many cancers (Supplementary Figure S2).

3.4. Epigenetic modification of *DDX58*

According to promoter methylation analysis, *DDX58* is hypermethylated in a variety of cancer types (Figure 4A). *DDX58* methylation seems to be correlated with the level of DNA methyltransferase mRNA expression in various cancers (all $p < 0.05$; Figure 4B). As we all know, DNA methylation is the result of DNA methyltransferase, which plays a role by covalently binding to the methyl at the 5′ carbon position of cytosine, a CpG dinucleotide in the genome. A correlation was found between methyl related genes and various cancers. There was a positive correlation between the expression of *DDX58* in pan-cancer and methyl-related genes, which meant that *DDX58* may mediate tumor genesis and progression by regulating epigenetic status. Moreover, it was worth noting that the correlation coefficient was higher in DLBC and UVM.

3.5. Genetic variation analysis of *DDX58* in pan carcinoma

Based on the cBioPortal database, we found that there were higher *DDX58* gene changes in LUSC, UCEC, STAD, and SKCM, and mutation was the main type (Figure 5A). It further proved the type, location and quantity of *DDX58* gene modification. R244K/I changes were detected in 4 patients with *DDX58* (Figure 5B). Then the 3D structure of *DDX58* protein at this mutation site was mapped (Figure 5C). The most common type of mutation found in pan-cancer analysis were gain and diploid (Figure 5D). In addition, TRAJ6, TMEM158, YY1P2, TTN, TAF1L, TP53, TOPORS, MUC16, ACO1, RYR2 gene changes were more common in the altered group than in the unchanged group (Figure 5E).

3.6. *DDX58* is associated with TMB and MSI in some cancers

TMB and MSI are effective prognostic biomarkers and indicators of immunotherapeutic response in many tumors. From these two analyses, we can conclude the relationship between *DDX58* and immunotherapy prognosis of specific cancer types.

The tumor cell genome's TMB is usually expressed as the total number of non-synonymous mutations within an average 1 M base region. In some cases, it is also expressed directly as the number of somatic mutations. Base substitution, frameshift mutation, deletion mutation, insertion mutation and other mutation types are the most common mutation type. In tumor cells, TMB is a quantifiable indicator of mutation frequency. The correlation between *DDX58* and TMB was calculated for each tumor. Ten tumors showed a significant correlation, including a significant positive correlation in 6 tumors, such as GBMLGG ($N = 650$; $R = 0.1430$, $p = 0.0002$), COAD ($N = 282$; $R = 0.1328$, $p = 0.0257$), COADREAD ($N = 372$; $R = 0.1103$, $p = 0.0334$), KIPAN ($N = 679$; $R = 0.194$, $p = 3.4926e\text{-}7$), UCS ($N = 57$; $R = 0.3099$, $p = 0.01895$), BLCA ($N = 407$; $R = 0.0996$, $p = 0.0444$), significantly negative correlation in 4 tumors, for example: BRCA ($N = 981$; $R = -0.0669$, $p = 0.0361$), HNSC ($N = 498$; $R = -0.1226$, $p = 0.0061$), THCA ($N = 489$; $R = -0.2149$, $p = 0.00001$), UVM ($N = 79$; $R = -0.3177$, $p = 0.0043$; Figure 6A).

DDX58 expression correlated with MSI in different types of cancer and we had calculated their spearman correlation in each

10.3389/fmicb.2022.1085086

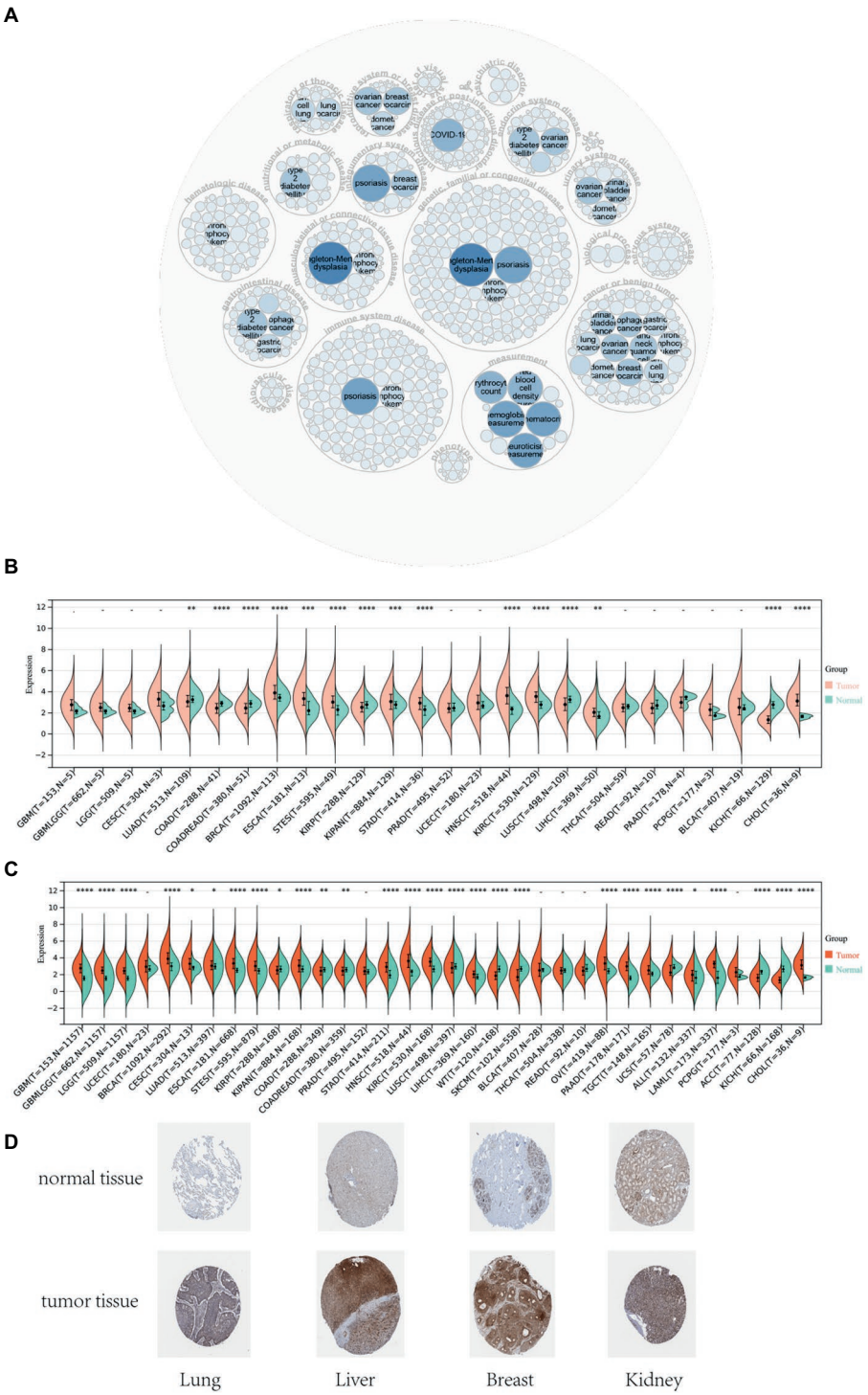


FIGURE 1
Differential expression analysis of *DDX58* in pan-cancer tissues and normal tissues. **(A)** *DDX58* related disease prediction **(B)** cancer and normal tissues in TCGA database **(C)** cancer and normal tissues in GTEx database **(D)** the protein expression level of *DDX58* in normal and tumor tissues of four different organs *$p<0.05$, **$p<0.01$, ***$p<0.001$. BLCA, Bladder Urothelial Carcinoma; BRCA, Bladder Urothelial Carcinoma; CHOL, Cholangiocarcinoma, COAD, Colon adenocarcinoma; ESCA, Esophageal carcinoma, GBM, Glioblastoma multiforme; HNSC, Head and Neck squamous cell carcinoma; KICH, Kidney Chromophobe; KIRC, Kidney renal clear cell carcinoma; KIRP, Kidney renal papillary cell carcinoma; LIHC, Liver hepatocellular carcinoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; PRAD, Prostate adenocarcinoma; READ, Rectum adenocarcinoma; STAD, Stomach adenocarcinoma; THCA, Thyroid carcinoma; UCEC, Uterine Corpus Endometrial Carcinoma.

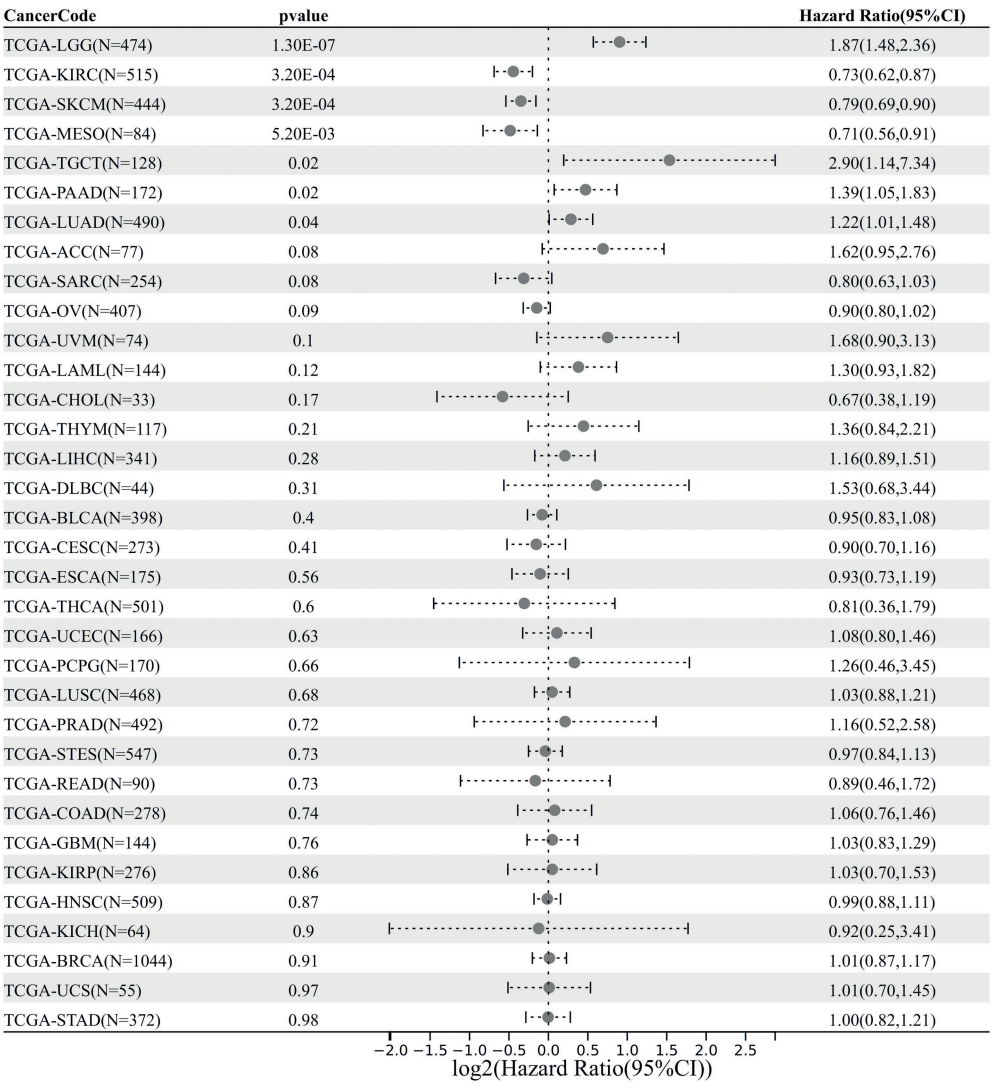


FIGURE 2
Pancancerous analysis of the diagnostic and prognostic value of *DDX58* expression. The forest map shows the HR and 95% CI of *DDX58* expression related to cancer OS.

tumor. A significant correlation was observed in 10 tumors, including significant positive correlation in 3 tumors, such as COAD ($N = 285$; $R = 0.1526$, $p = 0.0098$), COADREAD ($N = 374$; $R = 0.1068$, $p = 0.0389$), THYM ($N = 118$; $R = 0.1868$, $p = 0.0428$), and the significant negative correlation in 7 tumors, such as GBMLGG ($N = 657$; $R = -0.1286$, $p = 0.0009$) LGG ($N = 506$; $R = -0.0878$, $p = 0.0483$), KIPAN ($N = 688$; $R = -0.3528$,$p = 1.351e\text{-}21$), PRAD ($N = 495$; $R = -0.1286$, $p = 0.0041$), THCA ($N = 493$; $R = -0.0967$, $p = 0.0317$), PAAD ($N = 176$; $R = -0.1677$, $p = 0.0260$), DLBC ($N = 47$; $R = -0.4937$, $p = 0.0004$; Figure 6B).

It was worth noting that the absolute coefficients associated with TMB or MSI in the COAD cohort were relatively high compared with other cancer types, suggesting that the it may be sensitive to immunotherapy.

3.7. *DDX58* might regulate tumor immune microenvironment by influencing immune invasion of various cancer types and expression of immune checkpoints

To determine whether this pathway affects the tumor immune microenvironment, we studied the expression of *DDX58* with the degree of immune cell infiltration in each cancer type. Using the data collected from TCGA and the six types of immune cells available in TIMER database (B cells, CD4 + T cells, CD8 + T cells, neutrophils, macrophages and dendritic cells) for analysis, the results indicated that there was significant correlation in multiple tumors (Figure 7). It was worth noting that CD8 + T cells

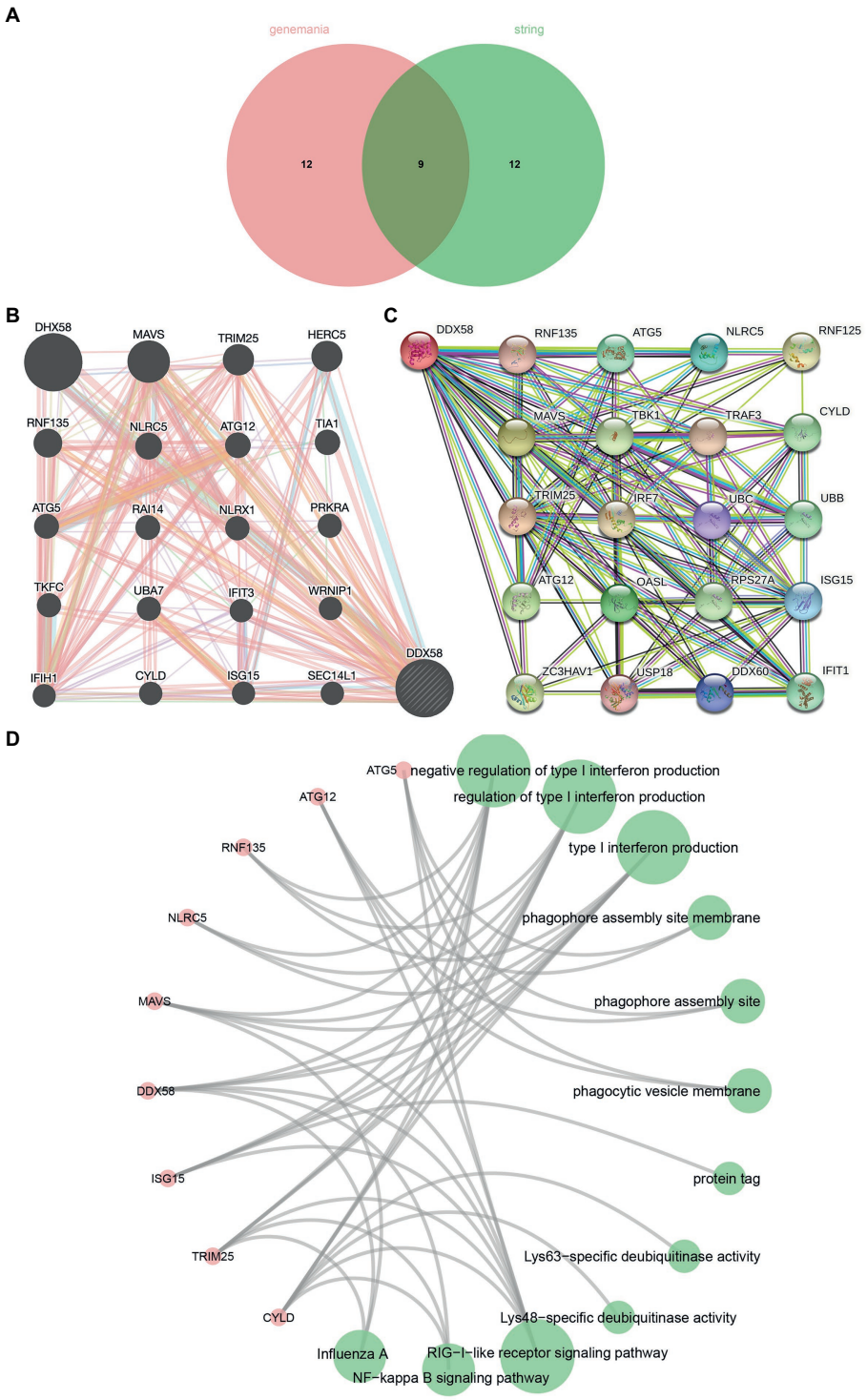


FIGURE 3
Gene, protein and disease networks use the *DDX58* related gene network mapped by GeneMANIA. **(A)** The Venn diagram where STRING and GeneMANIA intersect. **(B)** *DDX58* related gene network mapped by GeneMANIA **(C)** *DDX58* related protein network mapped by STRING. **(D)** Enrichment analysis of cross genes.

had the highest DLBC correlation coefficient. Their corresponding linear regression diagram showed that the high expression of *DDX58* may be related to the increased level of immune cell infiltration. Similarly, *DDX58* also affected the expression of immune checkpoints in different cancers (Supplementary Figure S3).

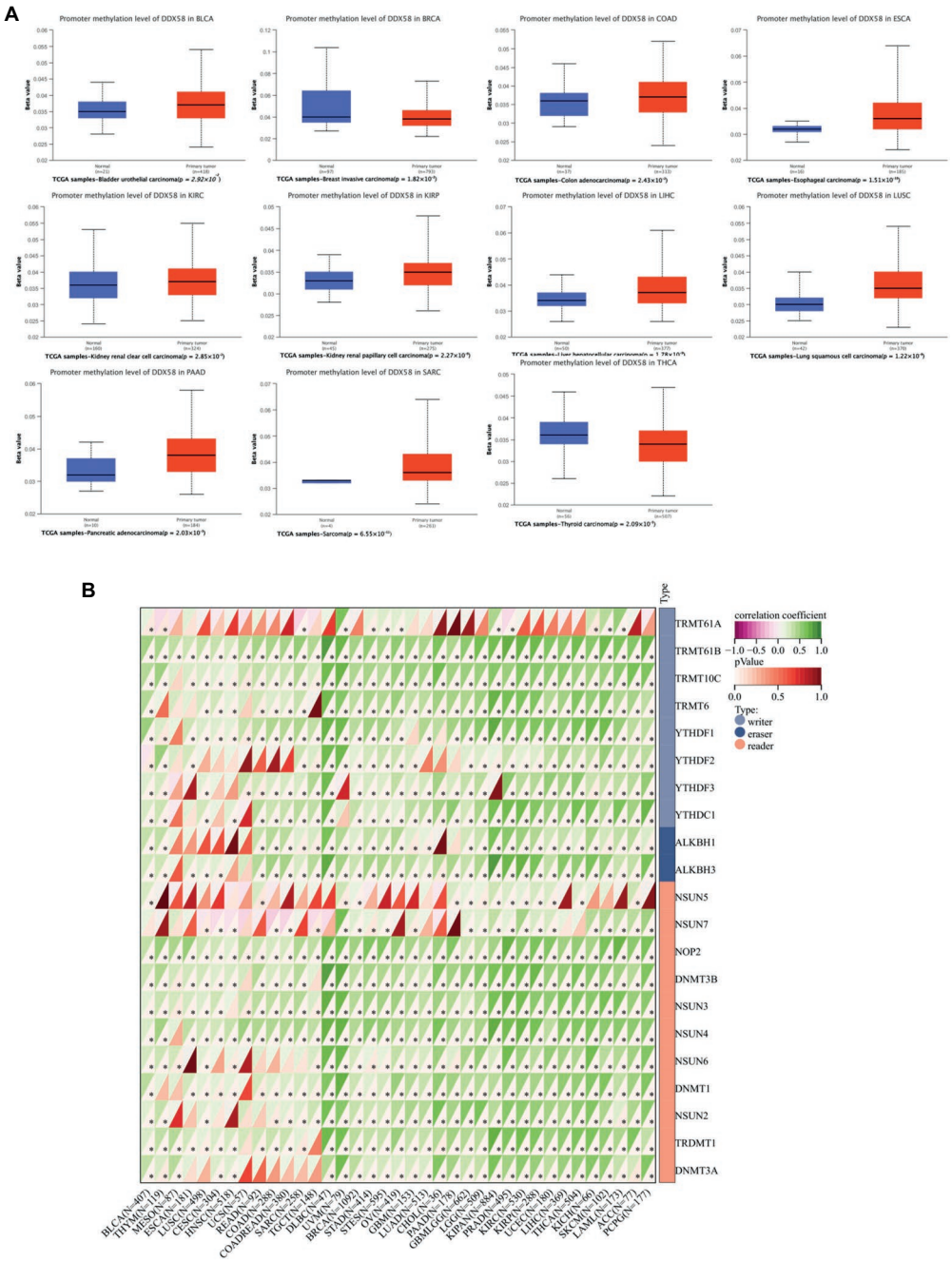


FIGURE 4
Correlation analysis between *DDX58* methylation level and methyltransferase expression level in pan-cancer tissues. **(A)** Display the difference of *DDX58* methylation level between tumor and adjacent normal tissues in TCGA database (β Value). **(B)** Correlation between *DDX58* expression and methylation related gene expression.

3.8. *DDX58* drug sensitivity analysis

The database was further analyzed to determine whether *DDX58* expression was correlated with drugs using CellMiner™ (Figure 8). Our results indicated that the expression of *DDX58* was positively correlated with the sensitivity to Cediranib, VE−821, Itraconazole, JNJ−42756493, IWR−1, Linsitinib. And the expression of *DDX58* was negatively correlated with the drug sensitivity of geldanamycin analysis, Tanespimecin, TYROTHRICIN, Panobinostat, Alvespimycin, Quisinostat, XR−5944, Lapiphone, Paclitaxel,Tamoxifen.

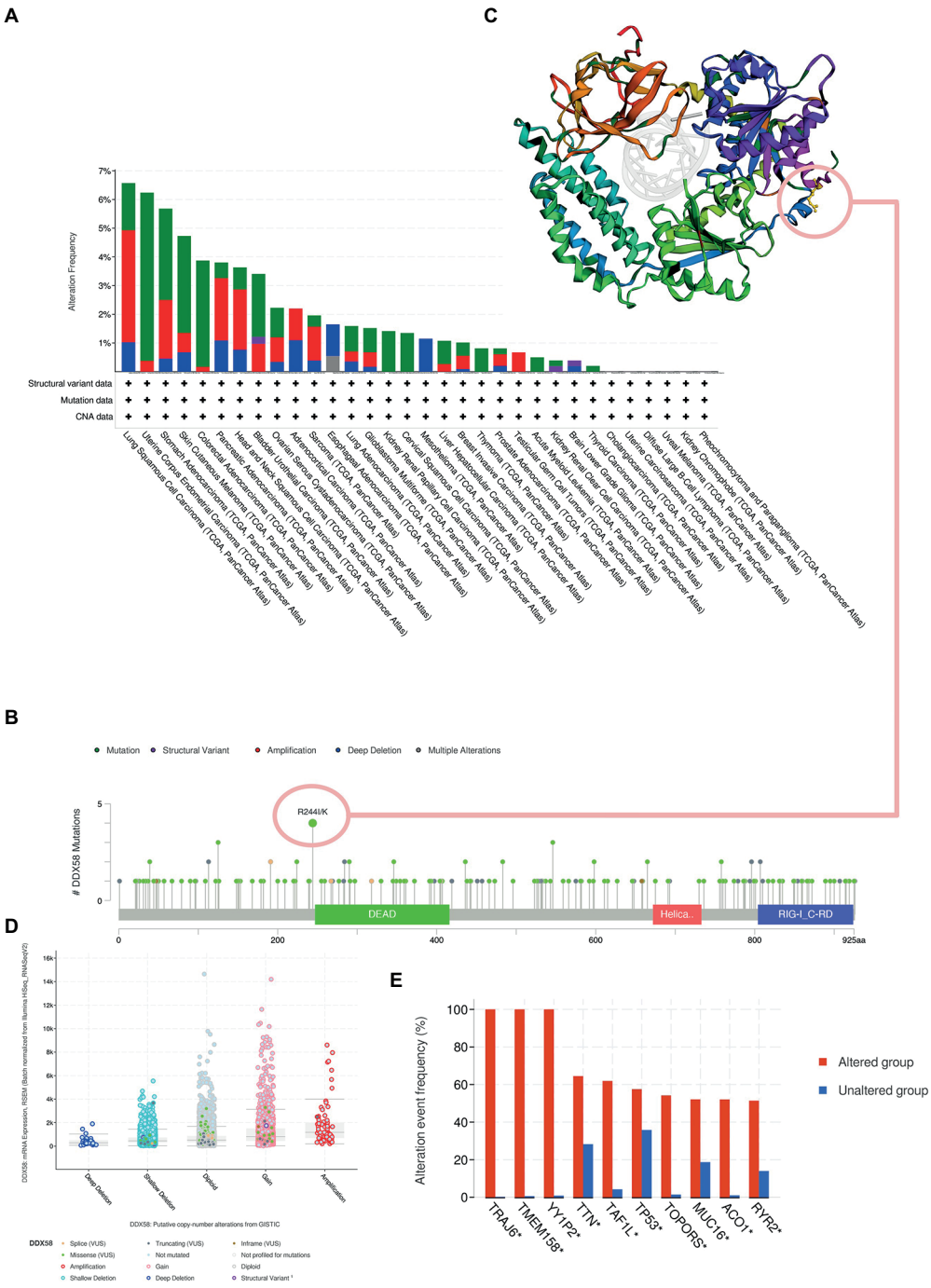


FIGURE 5
Genetic changes of *DDX58*. **(A)** Summary of *DDX58* changes in TCGA pan-cancer dataset. **(B)** The type, number and location of mutations in *DDX58* gene changes. **(C)** The 3D structure of *DDX58* at 232 mutation site. **(D)** The type of *DDX58* change in pan carcinoma. **(E)** Change frequency of related genes in *DDX58* changed and unchanged groups.

4. Discussion

As of December 2019, COVID-19 had caused a worldwide pandemic and posed a serious threat to global public health (Talic et al., 2021). As a result of the COVID-19 pandemic, cancer patients were more likely to be infected with SARS CoV-2. According to these findings, COVID-19 might have an impact on cancer patients' survival. RNA sensor RIG-I (*DDX58*) was a protein coding gene. The diseases related to RIG-I included Singleton Merten syndrome2 and Singleton Merten syndrome2. Signaling pathways that leaded to the production of type I interferon and proinflammatory cytokines in response to

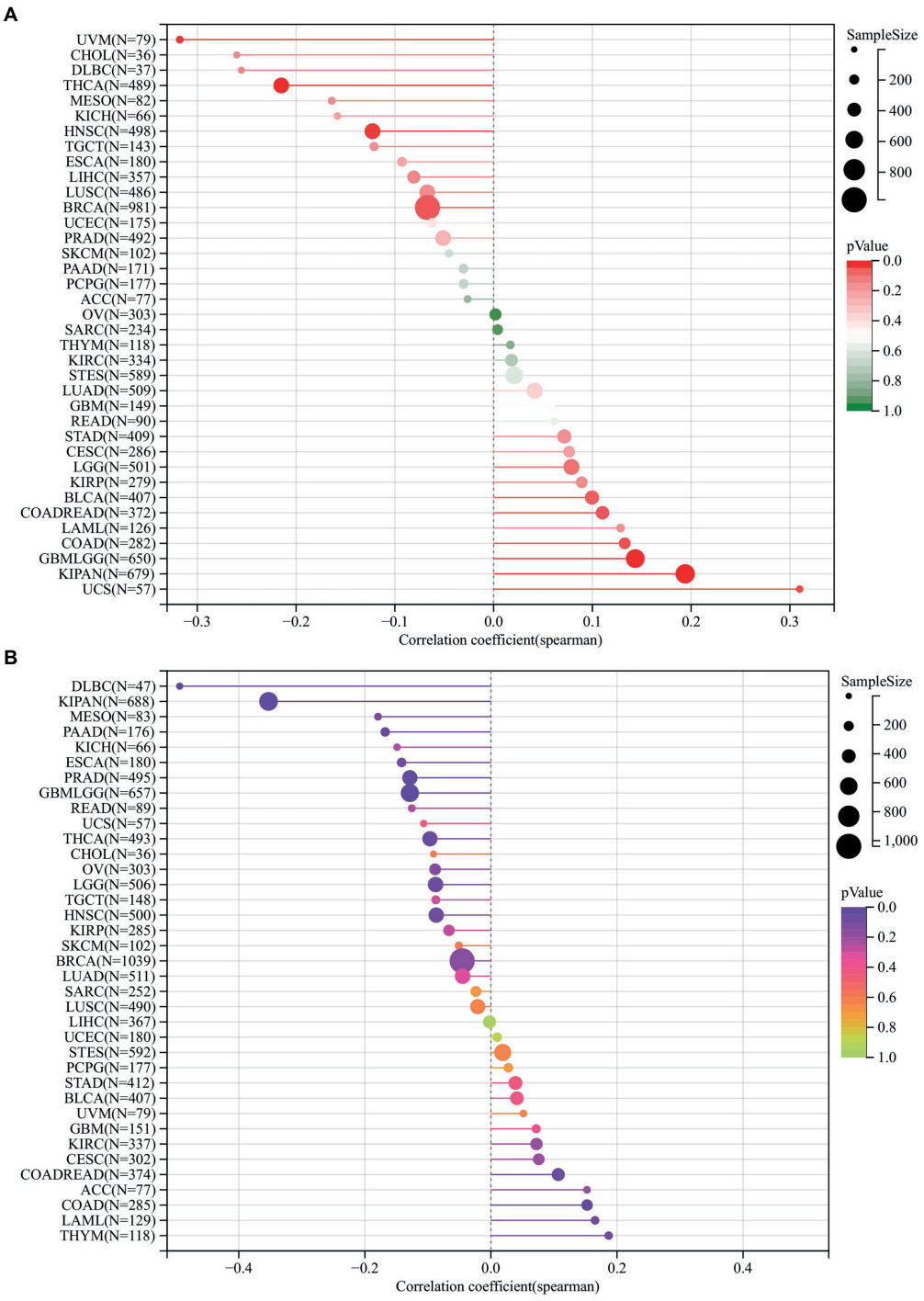


FIGURE 6
The relationship between the mRNA expression levels of TMB, MSI and *DDX58* in various cancers found in TCGA database. TMB was calculated based on the total incidence of mutations per megabase pair in each tumor, and MSI was calculated based on the total incidence of deletions or insertions in repeated sequences per megabase pair. **(A)** Correlation between TMB and DDX58 expression. **(B)** Correlation between MSI and *DDX58* expression. Spearman correlation test, $p<0.05$ is significant.

cytoplasmic viral nucleic acids (Bamming and Horvath, 2009; Shi et al., 2017; Zhao et al., 2017; Cadena et al., 2019). It formed ribonucleoprotein complex with viral RNA, on which homologous polymerization forms silk (Yoneyama et al., 2004; Sumpter et al., 2005). 3pRNA (RIG-1 agonist) treatment could increase cell death in melanoma cell lines and keep most melanoma cells in a non-proliferative state (Thier et al., 2022). In addition, RIG-1 activation inhibited STAT3/CSE pathway activity

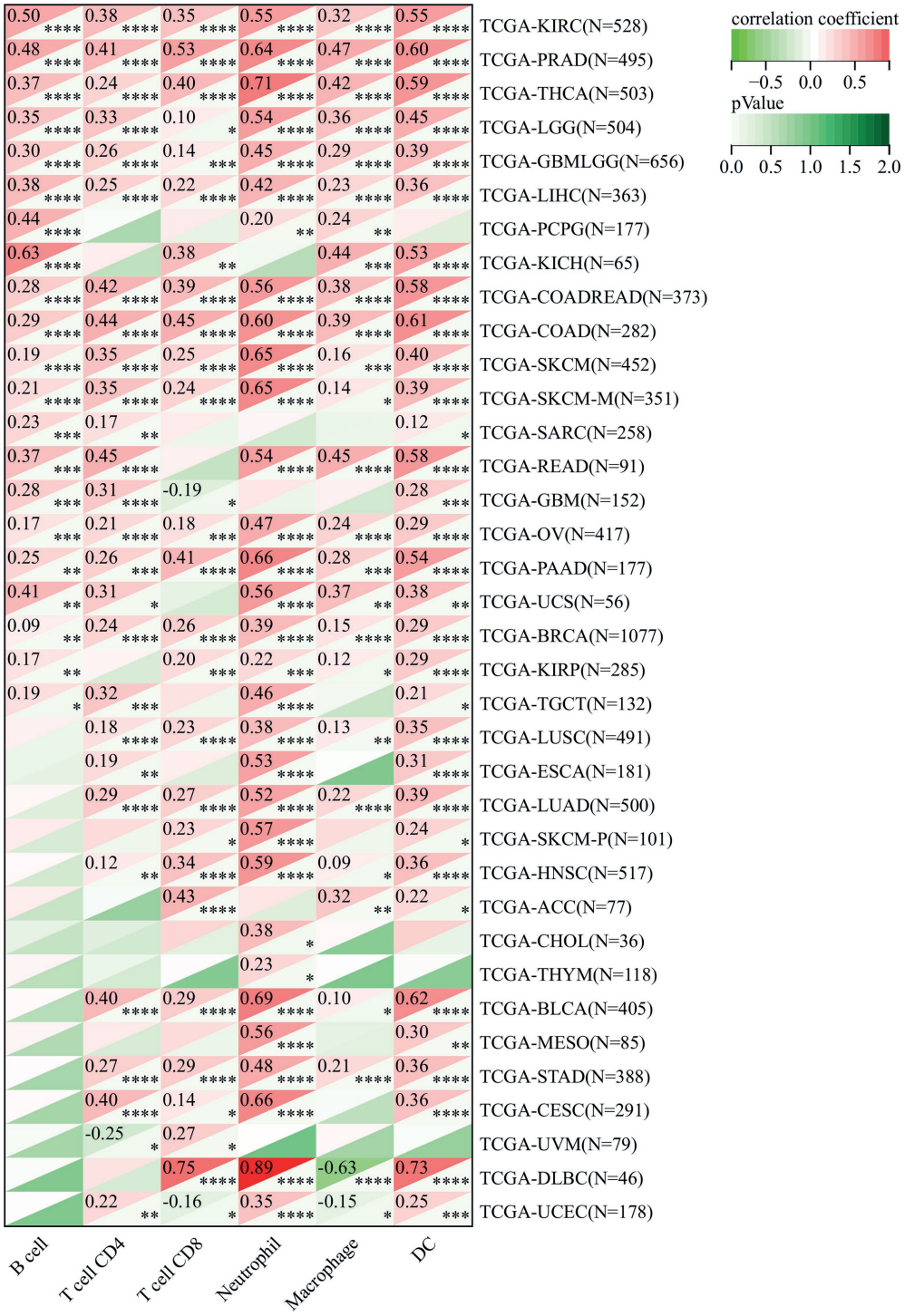


FIGURE 7
The expression level of *DDX58* mRNA calculated by TCGA and TIMER in the database was significantly correlated with the infiltration score of six common immune cells (B cells, CD4+T cells, CD8+T cells, neutrophils, macrophages, dendritic cells). Spearman correlation test, $p<0.05$ is significant.

to restrain the proliferation of colon cancer cells (Deng et al., 2022).

Thus, to clarify how *DDX58* contributes to the pathogenicity of COVID-19, we must examine its relation to *DDX58*, this study systematically analyzed the expression profile of *DDX58* in the entire cancer type spectrum. Using TCGA pan-cancer database and related data resources, we analyzed the expression, survival analysis, methylation expression, mutation status, microsatellite

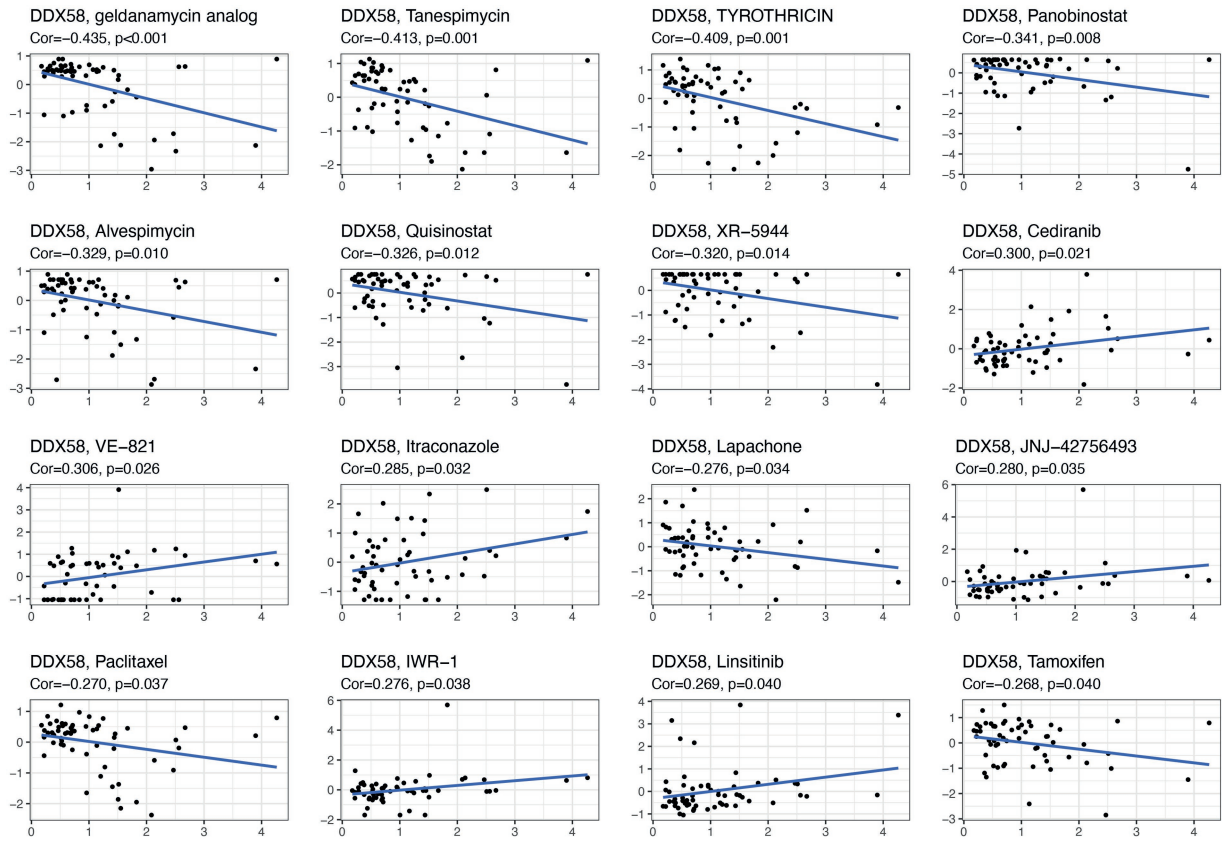


FIGURE 8
DDX58 drug sensitivity analysis.

instability (MSI), immune related microenvironment, gene related network, function and drug sensitivity of *DDX58*. Analysis of the relationship between *DDX58* expression and cancer immune invasion, tumor mutation, microenvironment and drug sensitivity had been finished, in order to determine *DDX58*'s potential for cancer immunotherapy and anti-COVID-19 treatment. We also carried out the correlation analysis between *DDX58* and AEC2 (SARS CoV-2 receptor; Supplementary Figure S2) to better understand the role of *DDX58* in COVID-19 and cancers.

In this study, we found the changes of *DDX58* mRNA in tumors. According to our research, pan-cancer was closely associated with the expression of *DDX58* protein. *DDX58* was highly expressed in BRCA, ESCA, STES, KIPAN, STAD, HNSC, KIRC, LIHC, CHOL, while it was low expressed in LUAD, COAD, READ, KIRP, LUSC, and KICH. *DDX58* was significantly associated with poor prognosis of LGG, TGCT, PAAD, LUAD, but significantly associated with improved prognosis of KIRC, SKCM, MESO patients. It indicated that *DDX58* might play different roles and functions in different cancers.

Eight genes were obtained by crossing the potential genes that interact with *DDX58* in the two databases, and nine genes including *DDX58* were analyzed by GO and KEGG. These 8 genes were ATG5, ATG12, RNF135, NLRC5, MAVS, ISG15, TRIM25, and CYLD, respectively. ATG5 usually combined with ATG12, catalyzed ATG7 and ATG10, played a role in autophagy, and regulates various functions of the body (Cui et al., 2022). It was known that RNF135 regulated the expression of IFN, and it participated in the RIG-I signal pathway by targeting RIG-I (Lai et al., 2019). NLRC5 could combine with LC3 to mediate MHC class I antigen presentation pathway (Zhan et al., 2022). MAVS mediated antiviral innate immunity (Zhang et al., 2022). The protein encoded by ISG15 gene was a ubiquitin like protein, when it was activated by interferon-αand-β, it binded to target proteins in cells. The encoded protein had a variety of functions, including chemotactic activity to neutrophils, orientation of junction target protein to intermediate filament, intercellular signal transduction and antiviral activity during viral infection (Jurczyszak et al., 2022). In responsd to ubiquitin E3 ligase and ISG15 E3 ligase (Zou and Zhang, 2006), TRIM25 played a role in the innate immune response to viruses by ubiquitinating *DDX58* and IFIH1 (Chiang et al., 2021). CYLD was a ubiquitin free enzyme that participates in NFκB activation and TNF-α induced necrosis (Dobson-Stone et al., 2020). Through enrichment analysis, it was found that these genes were associated with interferon related pathways, phagosomes, ubiquitination, RIG-I, NFκB related pathway, suggesting that it may affect the development of cancer through regulating immunity (Overman et al., 2017; Yang et al., 2021).

Disease network analysis found that *DDX58* was related to genetic, family or genetic disease, immune system disease, infectious disease, cancer or disease. This also showed that this gene was closely related to tumor and infectious diseases. Afterwards, we examined the relationship between *DDX58* expression and immune cell infiltration, and found that *DDX58* was significantly correlated with six types of immune cells (B cells, CD4 + T cells, CD8 + T cells, neutrophils, macrophages, and dendritic cells). In addition, abnormal DNA methylation was highly related to the occurrence, growth and carcinogenesis of tumors (Yang et al., 2021). Our study found that compared with their normal counterparts, cancer tissues were significantly hypermethylated, indicating that *DDX58* might promote tumor development by altering DNA methylation. However, the exact mechanism was still unclear. TMB and MSI are effective biomarkers to predict the prognosis of various tumors and indicators of immune response. TMB and MSI had been shown to be indicators of drug response in previous studies, particularly those that target immune checkpoint inhibitors such as CTLA4 and PD-1/PD-L1 (Overman et al., 2017; Mariathasan et al., 2018; Shim et al., 2020). Subsequently, we used the CellMinerTM database to find that the expression of *DDX58* was related to the sensitivity to many drugs, including Cediranib, VE−821, Itraconazole, JNJ−42,756,493, IWR−1, Linsitinib. These results are helpful to promote clinical drug guidance.

However, there were still some deficiencies in our research. First, based on bioinformatics analysis, there was a lack of relevant experimental or clinical data. In addition, although there was a correlation between the expression of *DDX58* in some tumors and survival rates, and *DDX58* changed the infiltration of immune cells, we were unable to establish a direct causal relationship. Future biological research needs to further clarify and confirm the role of *DDX58* in cancer.

In conclusion, the expression level of *DDX58* was significantly different in pan carcinoma.Turning RIG-I Sensor Activation Against Cancer had been used in clinical trails (Iurescia et al., 2020).And it had been proved that SARS CoV-2 M protein could inhibit the expression of IFNb and interferon stimulated genes induced by RIG-1(Sui et al., 2021). However, how *DDX58* played a role in these two diseases had not been reported. As an immune related biomarker, *DDX58* could be used to diagnose and predict the prognosis of COVID-19 cancer patients and their potential therapeutic targets.

5. Conclusion

We found that *DDX58* expression, survival prognosis, methylation, MSI, TMB, tumor immune microenvironment and drug sensitivity were different in pan-cancer. It was expected that *DDX58* might become a potential target for COVID-19 cancer therapy based on its abnormal expression in pan-cancer and significant differences in prognosis and immune environment. As a result, this study provided new insight into *DDX58*'s possible role in drug regulation as well as exploring its multiple roles in pan-cancer.

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

LS, YZ, and YH conceived and designed the study. ZH, LY, and LJ performed the experiments. LY and JC analyzed the data. ZH and LY wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Joint Funds for the innovation of science, Technology, Fujian province (Grant number: 2020Y9039) and Medical Research Fund of Guangdong (No. 2021112015285821).

Acknowledgments

The authors thank reviewers for helpful comments on the manuscript.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.1085086/full#supplementary-material

References

Al-Quteimat, O. M., and Amer, A. M. (2020). The impact of the COVID-19 pandemic on cancer patients. *Am. J. Clin. Oncol.* 43, 452–455. doi: 10.1097/COC.0000000000000712

Bamming, D., and Horvath, C. M. (2009). Regulation of signal transduction by enzymatically inactive antiviral RNA helicase proteins MDA5, RIG-I, and LGP2. *J. Biol. Chem.* 284, 9700–9712. doi: 10.1074/jbc.M807365200

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., et al. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature* 463, 899–905. doi: 10.1038/nature08822

Cadena, C., Ahmad, S., Xavier, A., Willemsen, J., Park, S., Park, J. W., et al. (2019). Ubiquitin-dependent and-independent roles of E3 ligase RIPLET in innate immunity. *Cells* 177, 1187–1200.e16. doi: 10.1016/j.cell.2019.03.017

Chiang, C., Dvorkin, S., Chiang, J. J., Potter, R. B., and Gack, M. U. (2021). The small t antigen of JC virus antagonizes RIG-I-mediated innate immunity by inhibiting TRIM25's RNA binding ability. *MBio* 12, e00620–e00621. doi: 10.1128/mBio.00620-21

Cui, J., Ogasawara, Y., Kurata, I., Matoba, K., Fujioka, Y., Noda, N. N., et al. (2022). Targeting the ATG5-ATG16L1 protein-protein interaction with a hydrocarbon-stapled peptide derived from ATG16L1 for autophagy inhibition. *J. Am. Chem. Soc.* 144, 17671–17679. doi: 10.1021/jacs.2c07648

Deng, Y., Fu, H., Han, X., Li, Y., Zhao, W., Zhao, X., et al. (2022). Activation of DDX58/RIGI suppresses the growth of tumor cells by inhibiting STAT3/CSE signaling in colon cancer. *Int. J. Oncol.* 61, 1–13. doi: 10.3892/ijo.2022.5410

Dobson-Stone, C., Hallupp, M., Shahheydari, H., Ragagnin, A. M. G., Chatterton, Z., Carew-Jones, F., et al. (2020). CYLD is a causative gene for frontotemporal dementia - amyotrophic lateral sclerosis. *Brain* 143, 783–799. doi: 10.1093/brain/awaa039

Gounder, M. M., Agaram, N. P., Trabucco, S. E., Robinson, V., Ferraro, R. A., Millis, S. Z., et al. (2022). Clinical genomic profiling in the management of patients with soft tissue and bone sarcoma. *Nat. Commun.* 13:3406. doi: 10.1038/s41467-022-30496-0

Han, H. J., Nwagwu, C., Anyim, O., Ekweremadu, C., and Kim, S. (2021). COVID-19 and cancer: from basic mechanisms to vaccine development using nanotechnology. *Int. Immunopharmacol.* 90:107247. doi: 10.1016/j.intimp.2020.107247

Iurescia, S., Fioretti, D., and Rinaldi, M. (2020). The innate immune Signalling pathways: turning RIG-I sensor activation against cancer. *Cancers* 12:3158. doi: 10.3390/cancers12113158

Jurczyszak, D., Manganaro, L., Buta, S., Gruber, C., Martin-Fernandez, M., Taft, J., et al. (2022). ISG15 deficiency restricts HIV-1 infection. *PLoS Pathog.* 18:e1010405. doi: 10.1371/journal.ppat.1010405

Lai, Y., Liang, M., Hu, L., Zeng, Z., Lin, H., Yi, G., et al. (2019). RNF135 is a positive regulator of IFN expression and involved in RIG-I signaling pathway by targeting RIG-I. *Fish Shellfish Immunol.* 86, 474–479. doi: 10.1016/j.fsi.2018.11.070

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* 77, e108–e110. doi: 10.1158/0008-5472.CAN-17-0307

Mariathasan, S., Turley, S. J., Nickles, D., Castiglioni, A., Yuen, K., Wang, Y., et al. (2018). TGFbeta attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* 554, 544–548. doi: 10.1038/nature25501

Morelli, M., Galluzzo, M., Madonna, S., Scarponi, C., Scaglione, G. L., Galluccio, T., et al. (2021). HLA-Cw6 and other HLA-C alleles, as well as MICB-DT, DDX58, and TYK2 genetic variants associate with optimal response to anti-IL-17A treatment in patients with psoriasis. *Expert. Opin. Biol. Ther.* 21, 259–270. doi: 10.1080/14712598.2021.1862082

Overman, M. J., McDermott, R., Leach, J. L., Lonardi, S., Lenz, H. J., Morse, M. A., et al. (2017). Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study. *Lancet Oncol.* 18, 1182–1191. doi: 10.1016/S1470-2045(17)30422-9

Reimer, N., Unberath, P., Busch, H., Börries, M., Metzger, P., Ustjanzew, A., et al. (2021). Challenges and experiences extending the cBioPortal for cancer genomics to a molecular tumor board platform. *Stud. Health Technol. Inform.* 287, 139–143. doi: 10.3233/SHTI210833

Shi, Y., Yuan, B., Zhu, W., Zhang, R., Li, L., Hao, X., et al. (2017). Ube2D3 and Ube2N are essential for RIG-I-mediated MAVS aggregation in antiviral innate immunity. *Nat. Commun.* 8:15138. doi: 10.1038/ncomms15138

Shim, J. H., Kim, H. S., Cha, H., Kim, S., Kim, T. M., Anagnostou, V., et al. (2020). HLA-corrected tumor mutation burden and homologous recombination deficiency for the prediction of response to PD-(L)1 blockade in advanced non-small-cell lung cancer patients. *Ann. Oncol.* 31, 902–911. doi: 10.1016/j.annonc.2020.04.004

Sui, L., Zhao, Y., Wang, W., Wu, P., Wang, Z., Yu, Y., et al. (2021). SARS-CoV-2 membrane protein inhibits type I interferon production through ubiquitin-mediated degradation of TBK1. *Front. Immunol.* 12:662989. doi: 10.3389/fimmu.2021.662989

Sumpter, R. Jr., Loo, Y. M., Foy, E., Li, K., Yoneyama, M., Fujita, T., et al. (2005). Regulating intracellular antiviral defense and permissiveness to hepatitis C virus RNA replication through a cellular RNA helicase, RIG-I. *J. Virol.* 79, 2689–2699. doi: 10.1128/JVI.79.5.2689-2699.2005

Talic, S., Shah, S., Wild, H., Gasevic, D., Maharaj, A., Ademi, Z., et al. (2021). Effectiveness of public health measures in reducing the incidence of covid-19, SARS-CoV-2 transmission, and covid-19 mortality: systematic review and meta-analysis. *BMJ* 375:e068302. doi: 10.1136/bmj-2021-068302

Thier, B., Zhao, F., Stupia, S., Brüggemann, A., Koch, J., Schulze, N., et al. (2022). Innate immune receptor signaling induces transient melanoma dedifferentiation while preserving immunogenicity. *J. Immunother. Cancer* 10:e003863. doi: 10.1136/jitc-2021-003863

Yamada, T., Sato, S., Sotoyama, Y., Orba, Y., Sawa, H., Yamauchi, H., et al. (2021). RIG-I triggers a signaling-abortive anti-SARS-CoV-2 defense in human lung cells. *Nat. Immunol.* 22, 820–828. doi: 10.1038/s41590-021-00942-0

Yang, B., Wang, J. Q., Tan, Y., Yuan, R., Chen, Z. S., and Zou, C. (2021). RNA methylation and cancer treatment. *Pharmacol. Res.* 174:105937. doi: 10.1016/j.phrs.2021.105937

Yoneyama, M., Kikuchi, M., Natsukawa, T., Shinobu, N., Imaizumi, T., Miyagishi, M., et al. (2004). The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. *Nat. Immunol.* 5:730. doi: 10.1038/ni1087

Zhan, L., Zhang, J., Zhang, J., Liu, X., Zhu, S., Shi, Y., et al. (2022). LC3 and NLRC5 interaction inhibits NLRC5-mediated MHC class I antigen presentation pathway in endometrial cancer. *Cancer Lett.* 529, 37–52. doi: 10.1016/j.canlet.2021.12.031

Zhang, R., Hou, X., Wang, C., Li, J., Zhu, J., Jiang, Y., et al. (2022). The endoplasmic reticulum ATP13A1 is essential for MAVS-mediated antiviral innate immunity. *Adv. Sci.* 9:e2203831. doi: 10.1002/advs.202203831

Zhao, C., Jia, M., Song, H., Yu, Z., Wang, W., Li, Q., et al. (2017). The E3 ubiquitin ligase TRIM40 attenuates antiviral immune responses by targeting MDA5 and RIG-I. *Cell Rep.* 21, 1613–1623. doi: 10.1016/j.celrep.2017.10.020

Zou, W., and Zhang, D. E. (2006). The interferon-inducible ubiquitin-protein isopeptide ligase (E3) EFP also functions as an ISG15 E3 ligase. *J. Biol. Chem.* 281, 3989–3994. doi: 10.1074/jbc.M510787200

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 22 December 2022
DOI 10.3389/fmicb.2022.1077111

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Qiaoming Liu,
Harbin Institute of Technology, China
Wenyan Wang,
Anhui University of Technology, China

*CORRESPONDENCE
Min Jin
✉ jinmin@hnu.edu.cn
Junlin Xu
✉ xjl@hnu.edu.cn

†These authors share first authorship

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

RECEIVED 22 October 2022
ACCEPTED 30 November 2022
PUBLISHED 22 December 2022

CITATION
Gong H, You X, Jin M, Meng Y, Zhang H, Yang S and Xu J (2022) Graph neural network and multi-data heterogeneous networks for microbe-disease prediction. *Front. Microbiol.* 13:1077111. doi: 10.3389/fmicb.2022.1077111

Graph neural network and multi-data heterogeneous networks for microbe-disease prediction

Houwu Gong[1,2†], Xiong You[3†], Min Jin[1*], Yajie Meng[4], Hanxue Zhang[1], Shuaishuai Yang[1] and Junlin Xu[1*]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, [2]Academy of Military Sciences, Beijing, China, [3]Center of Rehabilitation Diagnosis and Treatment, Hunan Provincial Rehabilitation Hospital, Changsha, China, [4]School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, China

The research on microbe association networks is greatly significant for understanding the pathogenic mechanism of microbes and promoting the application of microbes in precision medicine. In this paper, we studied the prediction of microbe-disease associations based on multi-data biological network and graph neural network algorithm. The HMDAD database provided a dataset that included 39 diseases, 292 microbes, and 450 known microbe-disease associations. We proposed a Microbe-Disease Heterogeneous Network according to the microbe similarity network, disease similarity network, and known microbe-disease associations. Furthermore, we integrated the network into the graph convolutional neural network algorithm and developed the GCNN4Micro-Dis model to predict microbe-disease associations. Finally, the performance of the GCNN4Micro-Dis model was evaluated *via* 5-fold cross-validation. We randomly divided all known microbe-disease association data into five groups. The results showed that the average AUC value and standard deviation were 0.8954 ± 0.0030. Our model had good predictive power and can help identify new microbe-disease associations. In addition, we compared GCNN4Micro-Dis with three advanced methods to predict microbe-disease associations, KATZHMDA, BiRWHMDA, and LRLSHMDA. The results showed that our method had better prediction performance than the other three methods. Furthermore, we selected breast cancer as a case study and found the top 12 microbes related to breast cancer from the intestinal flora of patients, which further verified the model's accuracy.

KEYWORDS

graph neural network, multi-data heterogeneous networks, microbe-disease association, biological network, graph convolution neural network

Introduction

In microecology, human microbes, especially intestinal microbes, have been found to play a key role in the generation and development of human complex diseases (Baron, 1996). This discovery provided a new perspective for revealing the inherent pathological mechanism of complex diseases. Microbes are responsible for the development of infectious diseases, such as SARS, MERS, and COVID-19 (Singh et al., 2014; Gong et al., 2022). According to the latest real-time statistics from WHO, 618 million confirmed cases and 6.5 million deaths have been reported globally between the outbreak of COVID-19 up until 9 October 2022 (World Health Organization, 2022). Although the composition, morphology, and functions of microbial communities are well understood and thoroughly studied, systematically analyzing the mechanisms by which human microbes initiate and drive diseases is still a major challenge (Karstens et al., 2018). Generally, the interaction between microbes and diseases can be verified to high accuracy using traditional experimental techniques, which can determine whether a certain microbe is directly or indirectly related to diseases. However, this method requires advanced experimental setup, environmental conditioning, and scientific research skill (Teh et al., 2021). Experimentally identifying the relationship between millions of microbes and human diseases takes a lot of time, highly-skilled human labor, and financial resources. This pinch could be obliviated by combining deep learning methods and biological network methods to identify the potential interactions between microbes and diseases on a large scale, allowing us to systemically understand the pathogenic mechanism of complex human diseases and provide a reference for the prevention, diagnosis, and treatment of diseases (Liu et al., 2021).

To address the challenges above, we propose a graph convolutional neural network approach, termed GCNN4Micro-Dis, for microbe-disease prediction. The key motivation is to model associations between diverse biological domains through a graph neural network.

Related work

In 2016, Ma et al. (2017) established the Human Microbe-Disease Association Database (HMDAD) by collecting published literature and collating 483 pairs of human microbe-disease association information. These highly-accurate data sources have attracted the attention of the bio information field. Researchers have successively proposed microbe-disease prediction models based on different theories, which can be roughly divided into the following three categories: (1) methods based on network algorithms, (2) methods based on dichotomous local features, (3) Machine learning-based methods.

In network algorithm-based methods, the similarity or heterogeneous network is first constructed, then the association probability is calculated based on the network and the specific network algorithm. In 2017, Chen et al., (2018) proposed the first KATZHMDA, which used the known topological information of microbe-disease association network to infer the potential relationship between microbes and diseases by using the social network relationship prediction method. In this model, the problem of predicting potential associations is transformed into the calculation of the similarity between corresponding nodes according to the length and number of paths connecting two nodes in the network. This model not only exhibited excellent predictive power, but also pioneered the field of microbe-disease prediction. Huang et al. (2017) proposed the path-based human microbe-disease association prediction computing model (PBHMDA), which used a special depth-first search algorithm to traverse all the paths communicated between nodes in the heterogenous network, thereby obtaining the prediction score of each pair of microbe-disease association. Shen et al. (2016) used the restart random walk algorithm to score each candidate microbe-disease pair in the microbe network based on Spearman correlation and the disease network based on symptom similarity. The main advantage of these models is their ability to make full use of the network's topological information. They also involve few parameters, which greatly reduces the difficulty of parameter selection.

The second type of method is based on dichotomous local features. It considers microbes and diseases as local objects and calculates the final prediction by combining their characteristics. Huang et al., (2017) integrated two independent recommendation models and developed NGRHMDA to infer disease-related microbes. NGRHMDA considers diseases that share the same associated microbes or microbes that share the same associated diseases as neighbors. It then considers microbes and diseases as users and items, respectively, and adopts a collaborative filtering recommendation algorithm for local recommendation to make association predictions. Shen et al. (2018) proposed BiRWMP to predict microbe-disease association. The model first builds the microbe-disease associated-network, then it calculates the correlation between microbes and diseases based on the random walk algorithm, using the disease-to-microbe node as the initial starting point. Since the model is a combination of random walks, the local information of microbes, and the random walk of disease information, it can make better predictions than the one-way random walk model. This method improves the local feature bias by considering different perspectives, solving the noise problem caused by the known uneven distribution of associations in the data set to a certain extent and improving the model's overall predictive power.

The third category is machine learning-based methods. Wang et al. (2017) proposed LRLSHMDA for predicting potential disease-related microbes. Two objective functions were constructed using the Laplacian Regularized Least Squares classification method. An optimal classifier was trained by combining the known topological information of the microbe-disease association network. Potential disease-associated microbes are eventually inferred. Peng et al. developed

TABLE 1 Data features of verified microbe-disease association.

Number of diseases	Number of microbes	Number of microbe-disease association
39	292	450

ABHMDA, which reveals disease-related microbes through a strong classifier consisting of weak classifiers with corresponding weights. ABHMDA assigns different weights to multiple weak classifiers, which proves that the computational method can achieve satisfactory performance in identifying potential associations between microbes and diseases. This work inspired researchers to further explore more novel and effective computational methods to predict the association between microbes and diseases.

Materials and methods

Dataset

The dataset used in this study was downloaded from the newly built Human Microbe-Disease Association Database (HMDAD[1]), which collects human microbe-disease association data from 61 published studies. HMDAD contains 450 verified microbe-disease association records between 292 microbes and 39 diseases (Ma et al., 2017; Table 1).

Microbe-disease heterogeneous network

HMDAD allows the download of data on 39 diseases, 292 microbes, and 450 microbes with known association and disease data. This data can be represented as a microbe-disease binary network, which combines all microbe species ($M = \{m_1, m_2, m_3, \ldots, m_x\}$) and diseases ($D = \{d_1, d_2, d_3, \ldots, d_y\}$) as A network node. If the microbe m_j is known to be associated with disease d_i, add an edge between node m_j and d_i. Using the adjacency matrix $A \in R_{x*y}$, where x and y represent the database of different kinds of diseases and the number of microbes, an adjacency matrix A may be constructed. If d_i has been proven to be linked with m_j, then $A_{(i,j)} = 1$, or 0, resulting in an adjacency matrix A with 39 rows and 292 columns containing 1 s and 0 s.

A microbe-disease heterogeneous network is illustrated in Figure 1. The network is constructed from microbe similarity network, disease similarity network, and known microbe-disease associations. The heterogeneous network contains two node types: microbe nodes and disease nodes, and three types of connecting edges: microbe connecting edges, disease connecting edges, and microbe-disease association edges. The present study aimed to predict the potential association between microbes and diseases using the constructed microbe-disease heterogeneous network, and subsequently find new microbe-disease association pairs with high association possibility from it.

Graph convolutional neural network

Graph convolutional neural network (GCNN) is a model that applies convolution to the field of graph data (Wu et al., 2021). Its core idea is to learn a mapping function $f(x)$ by which the characteristics of a node x and its neighbors can be aggregated together, resulting in the representation vector of node x. In CNN, the image processing method is to further convolve and pool the matrix data by arranging the image pixels into a matrix (LeCun and Bengio, 1995). In GCNN, the image is processed by establishing a topological graph of corresponding relationships between vertices and edges. The spatial features on the topological graph are then extracted (Shou et al., 2022). The structure of GCNN is shown in Figure 2. The biggest difference between GCNN and CNN is that GCNN is stacked at multiple layers, and the parameters between layers are different. The parameters of each layer are shared iteratively. The biggest advantage of GCNN is its introduction of an optimized convolution parameter that extracts graph structure data features. This function is realized through a Laplace matrix in GCNN (Zhang et al., 2022).

GCNNs are divided into two major forms: spatial domain and spectral domain. Spatial domain GCNNs are similar to the application of convolution in deep learning and are optimized to collect information from adjacent nodes. Although this class of network intuitively borrows image convolution operations, it lacks a specific theoretical basis (He et al., 2022). In contrast, spectral domain GCNNs can extract features from nonlinear data more easily. They do so in three steps: (1) perform graphic Fourier transform on input data, (2) convolve the transform result in the spectral domain, (3) inverse Fourier transform convolution result.

Based on graph theory, the coefficient matrix obtained is defined as a graph with nodes and edges. Any graph composed of multiple nodes and edges can be expressed as $G = (V, E, W)$, where V is a node, E is the edge between two nodes, and W is the weighted adjacency matrix of connection weights between two vertices. It is usually represented by a Laplace matrix defined as $L = D - A$, where D and A represent the degree matrix and adjacency matrix, respectively. The degree matrix is a diagonal matrix representing the number of connected nodes. The adjacency matrix represents the relationship between nodes. Connected nodes are represented as 1, and unconnected nodes are represented as 0. The formula of the Laplace matrix is as follows:

1 http://www.cuilab.cn/hmdad

$$L = U\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}U^{-1} = UAU^{-1} \quad (1)$$

In Equation 1, U is a matrix composed of unit eigenvectors, and A is a diagonal matrix composed of the eigenvalues of the Laplace matrix.

Model performance evaluation metrics

For a prediction model, the model is under-fitted if the deviation is too large, and over-fitted if the variance is too large. A model's output is strongly distorted when it is under-fitted or over-fitted. To solve these two thorny problems, a set of evaluation methods and performance indicators are needed to comprehensively evaluate the prediction effect of the model. Evaluation methods evaluate the generalizability of the model. Performance indicators evaluate the performance of a single model. The evaluation methods and performance indicators are described in detail below.

Selecting appropriate evaluation methods and performance indicators is important for the evaluation of the model. In this study, common performance index parameters such as accuracy (Acc), recall (Rec), and F1 score (F1) are used (Zhou and Li, 2010). Their definitions are as follows:

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \quad (2)$$

$$\mathrm{Re}c = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = \frac{2^{*}TP}{2^{*}TP + FN + FP} \quad (4)$$

TP represents the number of known microbe-disease association data that can be correctly identified; FP represents the number of unknown microbe-disease association data that have not been correctly identified; TN represents the number of unknown microbe-disease association data that can be correctly identified; FN represents the number of known microbe-disease association data that have not been correctly identified.

The ROC and PR curves were widely used in model evaluation. In the microbe-disease association prediction literature, researchers used the area under the ROC curve (AUC

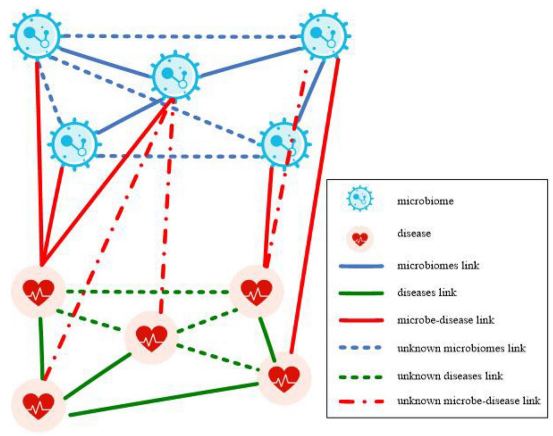


FIGURE 1
Microbe-disease heterogeneous network.

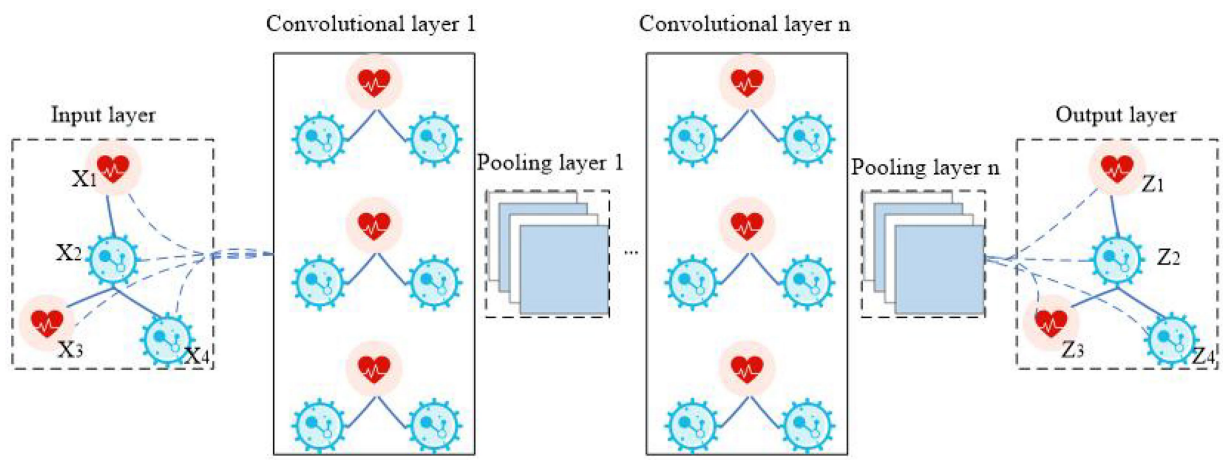


FIGURE 2
The flowchart of GCNN.

value) and the area under the PR curve (AUPR value) as the comprehensive evaluation indicators of the model. The larger the AUC and AUPR values, the better the predictive power of the model (Zhou and Washio, 2009).

ROC stands for "receiver operating characteristic." Its vertical axis is the true positive rate (TPR), while its horizontal axis is the false positive rate (FPR). FPR and TPR are calculated using the following formulae:

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

TPR represents the proportion of correctly identifying the known microbe-disease associations. FPR represents the proportion of incorrectly identifying the unknown microbe-disease associations. The meanings of TP, FN, FP, and TN have been described in detail in the literature. TP + FN represents all known microbe-disease associations, while FP + TN represents all unknown microbe-disease associations.

PR stands for Precision-Recall. Its vertical axis is Precision (Pre), while its horizontal axis is Recall (Rec). Precision is calculated as follows:

$$\Pr e = \frac{TP}{TP + FP} \tag{7}$$

Precision represents the proportion of correctly predicted known microbe-disease associations in all predicted known microbe-disease associations. Recall represents the proportion of correctly predicted known microbe-disease associations in all known microbe-disease associations.

To sum up, the ROC curve considers both positive and negative samples in the data set: the known microbe-disease associations and the unknown microbe-disease associations. This parameter can be applied to evaluate the overall performance of the model. The PR curve covers only the positive samples, the known microbe-disease associations. It is an indispensable indicator when there is an imbalance between positive and negative samples.

Results

Data preprocessing

The positive samples comprise 450 known interactions. The negative samples comprise 450 randomly selected data from the unknown interactions. If the node code of the disease is d_i and the microbe node code is m_j, then the sample code of the interaction between the disease and the microbe is $d_i + m_j$.

Dataset partition

When evaluating the merits and demerits of a prediction model, the choice of evaluation method is very important. In model evaluation, data sets are commonly divided into training and test sets. The partitioning should satisfy two conditions: the data in the respective sets follow the real distribution, and the data in the sets are mutually exclusive. Considering the different partitioning methods, the evaluation methods are mainly divided into three types: cross-validation, self-help, and set-aside (Zhou and Washio, 2009).

The present study utilized the same assessment method as the existing microbe-disease association predictive models. The proposed model was evaluated using the cross-validation method, specifically 5-fold cross-validation (5-fold CV). For the microbe-disease association data, these three datasets contained only known microbe-disease association data and unknown microbe-disease association data. The known microbe-disease association data were used as positive samples, while the unknown microbe-disease association data were used as negative samples.

Based on the 5-fold CV, all known microbe-disease associations were randomly divided into five groups.

1. Divide the positive samples into five subsets of equal size.
2. Divide the negative samples into five subsets of equal size.
3. One of the five subsets of positive and negative samples takes turns as the test set.
4. Remove the positive samples in the test set from the adjacency matrix by deleting their links with known interactions in the test set network.
5. In the remaining four subsets of positive and negative samples, the training set is 0.875, and the validation set is 0.125.
6. Randomly generate the initialization code of each node.
7. Repeat all experiments five times, with iteration set to 5, and average the final results to reduce the bias caused by random grouping.

Hyper-parameters selection

Convolutional neural network training can be regarded as a process of minimizing the loss function. The training network must initialize the parameters, set the appropriate learning rate, select the appropriate batch normalization method, and continuously iterate and update the parameters according to the optimization algorithm and strategy, including hyper parameters like Epoch, Batch, Batch_size, iteration, learning rate, etc.

In this experiment, we set Epoch to 100, learning rate to 0.001, coding dimension to 256, and the number of GCN coding layers to 3. Epoch refers to the complete training of the model using all the data in the training set, called "generation training." Iteration is the process of updating the model parameters using a Batch of

TABLE 2 The summary of model performance under 5-fold CV.

		Iter1	Iter2	Iter3	Iter4	Iter5
Fold0	Acc	0.7556	0.7722	0.7722	0.7722	0.7833
	Rec	0.7444	0.7556	0.7444	0.7333	0.7778
	F1	0.7528	0.7684	0.7657	0.7630	0.7821
	AUC	0.8121	0.8169	0.8223	0.8254	0.8328
	AUPR	0.7866	0.8071	0.8148	0.8223	0.8065
Fold1	Acc	0.7444	0.7333	0.7333	0.7556	0.7722
	Rec	0.7444	0.7667	0.7889	0.8111	0.7778
	F1	0.7444	0.7419	0.7474	0.7684	0.7735
	AUC	0.8020	0.8137	0.8230	0.8181	0.8207
	AUPR	0.7661	0.8146	0.8138	0.7945	0.7913
Fold2	Acc	0.7444	0.7222	0.7444	0.7278	0.7556
	Rec	0.7333	0.7556	0.7556	0.7667	0.7556
	F1	0.7416	0.7312	0.7473	0.7380	0.7556
	AUC	0.8258	0.8084	0.8226	0.7947	0.8126
	AUPR	0.8279	0.8282	0.8283	0.7794	0.8125
Fold3	Acc	0.7389	0.6833	0.7278	0.7333	0.7278
	Rec	0.7444	0.6667	0.7444	0.7222	0.7222
	F1	0.7403	0.6780	0.7322	0.7303	0.7263
	AUC	0.7795	0.7670	0.7985	0.7968	0.7974
	AUPR	0.7906	0.7539	0.7866	0.7919	0.7713
Fold4	Acc	0.7722	0.7556	0.7611	0.7556	0.7611
	Rec	0.7333	0.7111	0.7333	0.7000	0.6889
	F1	0.7630	0.7442	0.7543	0.7412	0.7425
	AUC	0.8485	0.8204	0.8338	0.8260	0.8190
	AUPR	0.8468	0.8250	0.8164	0.8237	0.7981

data, called "a training session." The learning rate determines how fast the parameters move to the optimal value. If the learning rate is too large, it is likely to cross the optimal value and lead to function convergence failure or even divergence. On the contrary, if the learning rate is too low, the optimization becomes inefficient, the convergence is too slow, and the algorithm can easily fall into a local optimum. The appropriate learning rate should converge as soon as possible on the premise of ensuring convergence.

Model effects

Samples with the same number of positive samples were randomly selected as negative samples from the unknown samples to ensure the balance of positive and negative samples. The 5-fold CV method was used to ensure that each sample data was used as a test set. The experiment was repeated five times, which greatly reduced the influence of randomness. The 25 experimental results reported 19 AUC values that are mostly above 0.8 with an average value of 0.8154, indicating that the model can be well applied to predict the link between diseases and microbes.

There is still a lot of room to improve the model's performance. Its results are largely limited by the amount of data, with only 450 positive samples utilized in this study. Furthermore, the node initialization coding adopted random initialization coding, which cannot express the inherent attribute characteristics of different node entities well.

The average AUC value and standard deviation given by the model was 0.8954 ± 0.0030. Our model evidently performed well and can help identify novel disease-microbe associations (Table 2).

The ROC and AUPR curves of the fifth experiment (Iter5) are shown in Figure 3.

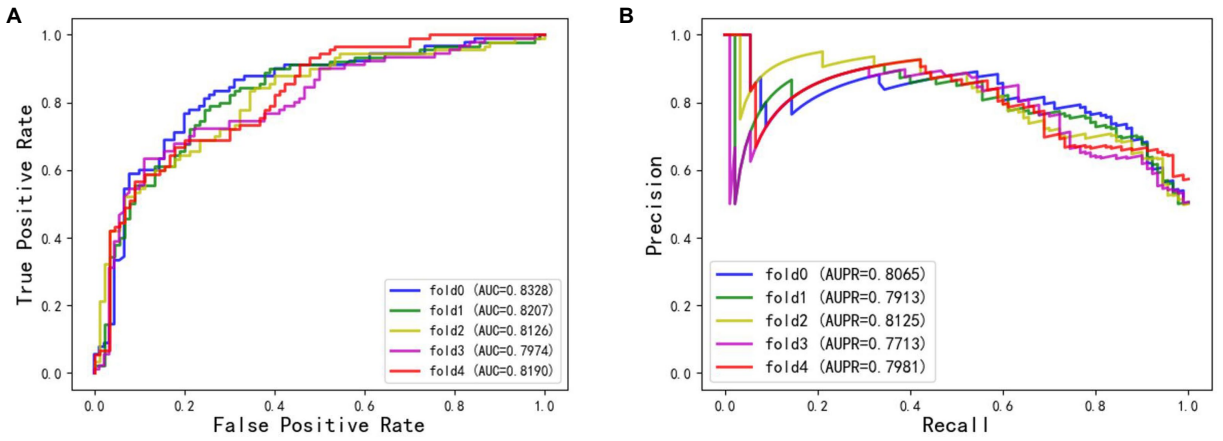


FIGURE 3
(A) The ROC curves of Iter5. **(B)** The AUPR curves of Iter5.

TABLE 3 Comparison of AUC and AUPR for different microbe-disease association predictions methods.

Methods	AUC	AUPR
GCNN4Micro-Dis	**0.8154**	**0.8092**
LRLSHMDA (Wang et al., 2017)	0.8410	0.5045
KATZHMDA (Chen et al., 2018)	0.8428	0.4782
BiRWHMDA (Zou et al., 2017)	0.7984	0.4363

Bold values represent the effect of our model.

TABLE 4 Top 12 potential microbes related to breast cancer.

BRCA subtypes	Rank	Microbes
HER2 positive	1	Megasphaera
	2	Barnesiellaceae
	3	Alloprevotella
ER positive	1	Megasphaera
	2	Roseburia
	3	Prevotellaceae
PR positive	1	Prevotellaceae
	2	Tyzzerella
	3	Enorma
Ki67 positive	1	Tenericutes
	2	Izimaplasmatales
	3	Sporobacter

Comparison with other methods

To verify the superiority of the GCNN4Micro-Dis model proposed in this study, it is compared with three advanced methods used to predict microbe-disease associations: KATZHMDA (Chen et al., 2018), BiRWHMDA (Zou et al., 2017), and LRLSHMDA (Wang et al., 2017).

- The KATZ measure for Human Microbe-Disease Association (KATZHMDA) is a novel computational model based on the assumption that functionally similar microbes tend to have similar interaction and non-interaction patterns with non-infectious diseases and vice versa (Chen et al., 2018).
- BiRWHMDA is a novel computational model to predict potential microbe-disease associations using bi-random walk on the heterogeneous network (Zou et al., 2017).
- The Laplacian Regularized Least Squares for Human-Microbe Disease Association (LRLSHMDA) is a semi-supervised computational model using the Gaussian interaction profile kernel similarity calculation and Laplacian regularized least squares classifier (Wang et al., 2017).

The AUC of BiRWHMDA reached 0.7984, while the AUCs of LRLSHMDA and KATZHMDA were 0.8410 and 0.8428, respectively. The AUC of GCNN4Micro-Dis was better than that of BiRWHMDA. Therefore, the performance of GCNN4Micro-Dis was not different from the other three methods in terms of prediction accuracy.

The data set used in this study was unbalanced, making the AUPR value an indispensable model evaluation index. The AUPR of LRLSHMDA, KATZHMDA, and BiRWHMDA were 0.5045, 0.4782, and 0.4363, respectively. The AUPR of GCNN4Micro-Dis was 0.8092, better than the other three competitors. The experimental data conclusively demonstrated that GCNN4Micro-Dis had a better prediction performance than the other three methods (Table 3).

Case study

In this section, a prevalent human disease, breast cancer, was selected as a case study to further analyze the performance of GCNN4Micro-Dis. Given that the role of gut microbiome in health and disease has recently attracted more and more attention, many observations and *in vitro* studies depict that it may be involved in the development of breast cancer. The 12 microbes most related to breast cancer were selected from the intestinal flora of patients as case studies. The result has been verified in the literature (Liu et al., 2020; Huang et al., 2021). Some fecal intestinal bacteria were found to be associated with breast cancer and are expected to become new targets for breast cancer treatment (Wu et al., 2016; Zheng et al., 2018; Table 4).

Conclusion

A heterogeneous network of microbe-disease association was constructed from data extracted from the HMDAD database. A graph neural network algorithm was proposed, and the accuracy of our algorithm was evaluated using a 5-fold cross-validation. The main parameters involved in the algorithm were verified, proving the effectiveness of the prediction method. The main research results of this paper are as follows.

GCNN4Micro-Dis, a microbe-disease prediction method based on the Graph Neural Network and Multi-Data Heterogeneous Networks, was proposed. The heterogeneous network was obtained by integrating the known microbe-disease networks. The network was applied to the Graph Neural Network model for prediction. The methods proposed in this study predicted the association between potential microbes and diseases. Although these methods performed well in experimental verification and analysis, there are still some limitations that could be addressed in future works:

(1) The known microbe-disease association dataset was too small, which reduced its accuracy to some extent. In the future, the method's predictive power will improve with more data available. (2) More similarity data can be added. The microbe and disease similarity in this paper are calculated from the known microbe-disease associations, which were inadequate. The

prediction could be more accurate if more similarity data could be integrated into the heterogeneous networks. (3) More network information can be added. The current prediction methods require known microbe disease association data. Without this information, most methods cannot be implemented. More information may be mined if the potential microbe disease association can be predicted without this information. For example, the correlation data between microbes and RNA and between RNA and microbes allows the use an RNA network as an intermediate layer to build a three-layer microbe RNA disease network. The three-layer heterogeneous network can mine more unknown information.

Due to the relatively late development of microbe-disease association prediction, there are still many deficiencies and challenges at the present stage. Nevertheless, many studies have made preliminary exploration on the design of the prediction model (Peng et al., 2017, 2021, 2022a,b; Shen et al., 2022), which can be summarized as follows:

1. There are relatively few validated microbe-disease association data. Relatively few microbe-disease associations have been demonstrated through biological experiments compared to other biomarkers, such as non-coding RNAs. Since current computational methods often infer possible microbe-disease associations based on known association data, more known associations are needed to enrich the training set of the prediction models and improve their prediction power. Therefore, more accurate microbe-disease associations should be mined, using biological experiments as the fundamental data source for the calculation methods.
2. Few available datasets. The number of publicly available microbe-disease association databases is limited, yet few researchers have constructed new data sets, forcing a broad consensus of data sets used in the field. Most of the data sets used currently are microbe-disease associations provided by the HMDAD database. Although they are true and reliable associations verified by biological experiments, the number is small. Small and single data sets cannot fully depict the performance of the prediction model and render the prediction model unreliable. Therefore, there is an urgent need to build a larger microbe-disease association database.
3. The design of some methods should be improved. Methods based on network algorithms usually make assumptions about probability distributions, which fail if the data sources are not conformant. For example, this part of the model constructs similarity networks by assuming that functionally similar microbes have similar interaction patterns with diseases, which is more beneficial for microbes with more known related diseases. Optimizing the network structure by introducing local features is expected to improve this deficiency.
4. The prediction performance must be improved. Microbe-disease association prediction is a relatively new research field, so the performance of the proposed prediction models must be improved. In the future, more diverse biological information and more effective computational methods (such as neural networks) can be used to design prediction models with superior performances.

As an unsupervised deep neural network, GCN can learn and extract features from unlabeled data, obtain low-dimensional feature expressions from high-dimensional original data, simplify the classification work, and overcome the randomness of weight coefficient initialization in traditional neural networks. In future works, biological information features, such as functional similarity of microbes and semantic similarity of diseases, will be considered for addition to GCNN4Micro-Dis to more accurately predict the associations between microbes and diseases and help prevent, diagnose, treat, and prognose diseases.

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

HG and MJ: conceptualization. JX and XY: methodology. YM and HZ: software. HZ and SY: validation. XY: resources, supervision, and funding acquisition. SY: data curation. HG: visualization and writing-original draft preparation. HG and JX: writing-review and editing. MJ and JX: project administration. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the Natural Sciences Foundation of Hunan Province (Grant No. 2021JJ30139), the National Natural Science Foundation of China (Grant No. 61773157), the Key Project of R & D plan of Changsha (Grant No. kq2004011), the China Postdoctoral Science Foundation (Grant No. 2022M711113), the Rehabilitation Project of Hunan Disabled Persons' Federation in 2022 (Grant No. 2022XK0305), and the Excellent Youth Fund of Hunan Provincial Department of Education (Grant No. 22B0021).

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

References

Baron, S. (1996). *Clostridia: Sporeforming Anaerobic Bacilli—Medical Microbiology*. Galveston, TX: University of Texas Medical Branch at Galveston.

Chen, X., Huang, Y. A., You, Z. H., Yan, G. Y., and Wang, X. S. (2018). A novel approach based on KATZ measure to predict associations of human microbiota with non infectious diseases. *Bioinformatics* 34:1440. doi: 10.1093/bioinformatics/btx773

Gong, H., Wang, M., Zhang, H., Elahe, M. F., and Jin, M. (2022). An explainable AI approach for the rapid diagnosis of COVID-19 using ensemble learning algorithms. *Front. Public Health* 10, 1–12. doi: 10.3389/fpubh.2022.874455

He, J., Xiao, P., Chen, C. Y., Zhu, Z., Zhang, J., and Deng, L. (2022). GCNCMI: a graph convolutional neural network approach for predicting circ RNA-mi RNA interactions. *Front. Genet.* 13:959701. doi: 10.3389/fgene.2022.959701

Huang, Z., Pan, J., Wang, H., Du, X., Xu, Y., Wang, Z., et al. (2021). Prognostic significance and tumor immune microenvironment Heterogenicity of m5C RNA methylation regulators in triple-negative breast cancer. *Front. Cell Dev. Biol.* 9:657547. doi: 10.3389/fcell.2021.657547

Huang, Z. A., Chen, X., Zhu, Z. X., Liu, H., Yan, G. Y., You, Z. H., et al. (2017). PBHMDA: path-based human microbe-disease association prediction. *Front. Microbiol.* 8:233. doi: 10.3389/fmicb.2017.00233

Huang, Y. A., You, Z. H., Chen, X., Huang, Z. A., Zhang, S., and Yan, G. Y. (2017). Prediction of microbe- -disease association from the integration of neighbor and graph with collaborative recommendation model. *J. Transl. Med.* 15, 1–11. doi: 10.1186/s12967-017-1304-7

LeCun, Y., and Bengio, Y. (1995). "Convolutional networks for images, speech, and time series" in *The Handbook of Brain Theory and Neural Networks* (Cambridge, MA: MIT Press)

Karstens, L., Asquith, M., Caruso, V., Rosenbaum, J. T., Fair, D. A., Braun, J., et al. (2018). Community profiling of the urinary microbiota: considerations for low-biomass samples. *Nat. Rev. Urol.* 15, 735–749. doi: 10.1038/s41585-018-0104-z

Liu, Y., Wang, S. L., Zhang, J. F., Zhang, W., Zhou, S., and Li, W. (2021). DMFMDA: prediction of microbe-disease associations based on deep matrix factorization using Bayesian personalized ranking. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 1763–1772. doi: 10.1109/TCBB.2020.3018138

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). Improved anticancer drug response prediction based on an ensemble method integrating matrix completion and ridge regression. *Mol. Ther.* 21, 676–686. doi: 10.1016/j.omtn.2020.07.003

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2017). An analysis of human microbe-disease associations. *Brief. Bioinform.* 18, 85–97. doi: 10.1093/bib/bbw005

Peng, L. H., Liao, B., Zhu, B., Li, Z., and Li, K. (2017). Predicting drug-target interactions with multi-information fusion. *IEEE J. Biomed. Health Inform.* 21, 561–572. doi: 10.1109/JBHI.2015.2513200

Peng, L. H., Wang, C., Tian, G., Liu, G., Li, G., Lu, Y., et al. (2022a). Analysis of CT scan images for COVID-19 pneumonia based on a deep ensemble framework with dense net, Swin transformer, and RegNet. *Front. Microbiol.* 13:995323. doi: 10.3389/fmicb.2022.995323

Peng , L. H., Wang, C., Tian, X. F., Zhou, L., and Li, K. (2021). Finding lncRNA-protein interactions based on deep learning with Dual-net neural architecture. *IEEE/ACM Trans Comput Biol Bioinform.* 19, 3456–3468. doi: 10.1109/TCBB.2021.3116232

Peng, L. H., Wang, F. X., Wang, Z., Tan, J., Huang, L., Tian, X., et al. (2022b). Cell-cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies. *Brief. Bioinform.* 23:bbac 234. doi: 10.1093/bib/bbac234

Shen, X., Chen, Y., Jiang, X., and Yang, J. (2016). "Predicting disease-microbe association by random walking on the heterogeneous network." in *Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 15–18 December.

Shen, L., Liu, F., Huang, L., Liu, G., Zhou, L., and Peng, L. (2022). VDA-RWLRLS: an anti-SARS-CoV-2 drug prioritizing framework combining an unbalanced bi-random walk and Laplacian regularized least squares. *Comput. Biol. Med.* 140:105119. doi: 10.1016/j.compbiomed.2021.105119

Shen, X., Zhu, H., Jiang, X., Hu, X., and Yang, J. (2018). A novel approach based on bi-random walk to predict microbe-disease associations. in *Proceedings of the International Conference on Intelligent Computing, F.* Cham: Springer.

Shou, Y. T., Meng, T., Ai, W., Yang, S., and Li, K. (2022). Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis. *Neurocomputing* 501, 629–639. doi: 10.1016/j.neucom.2022.06.072

Singh, S. R., Krishnamurthy, N. B., and Mathew, B. B. (2014). A review on recent diseases caused by microbes. *JAEM* 2, 106–115.

Teh, J. J., Berendsen, E. M., Hoedt, E. C., Kang, S., Zhang, J., Zhang, F., et al. (2021). Novel strain-level resolution of Crohn's disease mucosa-associated microbiota via an ex vivo combination of microbe culture and metagenomic sequencing. *ISME J.* 15, 3326–3338. doi: 10.1038/s41396-021-00991-1

Wang, F., Huang, Z. A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., et al. (2017). Lrlshmda: Laplacian regularized least squares for human microbe—disease association prediction. *Sci. Rep.* 7, 1–11. doi: 10.1038/s41598-017-08127-2

World Health Organization. (2022). Weekly epidemiological update on COVID-19 - 12 October 2022[EB/OL]. Available at: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19 (Accessed October 12, 2022).

Wu, C., Chen, L., and Li, L. (2016). Apelin/APJ system: a novel promising therapy target for pathological angiogenesis. *Clin. Chim. Acta* 466, 78–84. doi: 10.1016/j.cca.2016.12.023

Wu, Z. H., Pan, S. R., Chen, F. W., Long, G., Zhang, C., and Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 4–24. doi: 10.1109/TNNLS.2020.2978386

Zhang, P., Tu, S. K., Zhang, W., and Xu, L. (2022). Predicting cell line-specific synergistic drug combinations through a relational graph convolutional network with attention mechanism. *Brief. Bioinform.* 23:bbac403. doi: 10.1093/bib/bbac403

Zheng, R., Zhao, Y., Wu, J., Wang, Y., Liu, J. L., Zhou, Z. L., et al. (2018). A novel PNPLA6 compound heterozygous mutation identified in a Chinese patient with Boucher-Neuhauser syndrome. *Mol. Med. Rep.* 18, 261–267. doi: 10.3892/mmr.2018.8955

Zhou, Z. H., and Li, H. (2010). Preface [special section on advances in machine learning and applications]. *J. Comput. Sci. Tech.* 4, 651–652. doi: 10.1007/s11390-010-9354-9

Zhou, Z. H., and Washio, T.. (2009). Advances in machine learning. in *Proceedings of the 1st Asian Conference on Machine learning: Advances in Machine Learning[C]// Asian Conference on Machine Learning*. Heidelberg: Springer-Verlag.

Zou, S., Zhang, J., and Zhang, Z. (2017). A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network. *PLoS One* 12:e0184394. doi: 10.1371/journal.pone.0184394

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 05 January 2023
DOI 10.3389/fmicb.2022.1092467

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Yuhua Yao,
Hainan Normal University,
China
Yi Xiong,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Zhixiang Yin
✉ zxyin66@163.com

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

RECEIVED 08 November 2022
ACCEPTED 07 December 2022
PUBLISHED 05 January 2023

CITATION
Peng Y, Zhao S, Zeng Z, Hu X and Yin Z (2023) LGBMDF: A cascade forest framework with LightGBM for predicting drug-target interactions.
Front. Microbiol. 13:1092467.
doi: 10.3389/fmicb.2022.1092467

LGBMDF: A cascade forest framework with LightGBM for predicting drug-target interactions

Yu Peng, Shouwei Zhao, Zhiliang Zeng, Xiang Hu and Zhixiang Yin*

School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai, China

Prediction of drug-target interactions (DTIs) plays an important role in drug development. However, traditional laboratory methods to determine DTIs require a lot of time and capital costs. In recent years, many studies have shown that using machine learning methods to predict DTIs can speed up the drug development process and reduce capital costs. An excellent DTI prediction method should have both high prediction accuracy and low computational cost. In this study, we noticed that the previous research based on deep forests used XGBoost as the estimator in the cascade, we applied LightGBM instead of XGBoost to the cascade forest as the estimator, then the estimator group was determined experimentally as three LightGBMs and three ExtraTrees, this new model is called LGBMDF. We conducted 5-fold cross-validation on LGBMDF and other state-of-the-art methods using the same dataset, and compared their Sn, Sp, MCC, AUC and AUPR. Finally, we found that our method has better performance and faster calculation speed.

KEYWORDS

drug-target interactions, machine learning, LightGBM, deep forest, prediction

1. Introduction

In recent years, with the rapid development of computer data processing capabilities, the continuous enrichment of data content, and the improvement of algorithm models, more and more researches on artificial intelligence in the fields of biology and medicine have been carried out (Guo et al., 2021; Chen and Yin, 2022; Zhou et al., 2022). Many computational methods based on machine learning have been proposed to solve biological problems (Lihong et al., 2021; Zhou et al., 2021; Peng et al., 2022; Shen et al., 2022). Especially in drug development, the prediction of drug-target interactions (DTIs) played an important role in drug development and drug repositioning, so using machine learning methods to predict DTIs became a research hotspot.

Over the past decade, a large number of machine learning-based methods were proposed for identifying DTI (Zhou et al., 2019). Among them, binary classification

methods account for the majority. Some methods identify drug-target pairs based on drug and protein information, Li et al. (2020) used protein sequences and drug substructure fingerprint information to predict DTIs. In addition, there were many models (Mousavian et al., 2016; Li et al., 2020; Zhan et al., 2020; Tanoori et al., 2021) that predicted new DTIs based on information similarity.

In fact, there are more methods based on network inference, Yamanishi et al. (2010) integrated chemical, genomic and pharmacological information in bipartite graph to uncover potential DTIs. Mei J. P et al. (Mei et al., 2013) proposed Neighbor-based Interaction-profile Inferring (NII) based on bipartite local model (BLM). Chen et al. (2012) proposed the method of Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) which integrates three different networks into a heterogeneous network through known DTIs, and achieves random wandering on this heterogeneous network. Cao et al. (2014) proposed a computational method for DTI prediction by combining the information from chemical, biological, and network properties. Ding et al. (2017) used molecular substructure fingerprints, multivariate mutual information (MMI) of proteins and network topology to represent drugs, targets and their relationships, and employ SVM and Feature Selection (FS) to build predictive models. Thereafter, scholars began to extract features from more complex networks. SNF-CVAE (Jarada et al., 2021) integrates similarity network fusion (SNF) and collective variational autoencoder (CVAE) to improve prediction accuracy. An and Yu (2021) proposed a Network Embedding framework in mulTiPlex networks (NEDTP) to predict DTIs. Jin et al. (2021) proposed a machine learning model called HeTDR, the method combines drug features in multiple networks and disease features in biomedical corpora to predict the degree of association between drugs and diseases. In addition, there are some computational methods based on matrix factorization (Gönen, 2012; Liu et al., 2016; Bagherian et al., 2021) and multi-label learning (Yuan et al., 2016; Pliakos et al., 2019; Chu et al., 2021b).

Moreover, with the rise of deep learning methods, people have made a lot of achievements in the field of DTI prediction based on deep learning methods. Many scholars consider graph analysis (Olayan et al., 2018; Peng et al., 2021; Yang et al., 2022) as an important means to predict DTIs. Many models apply deep neural networks (DNN) to DTI prediction, LASSO-DNN (You et al., 2019) combines LASSO with DNN, deepDTnet (Zeng et al., 2020b) applies DNN algorithm to network embedding, DeepFusionDTA (Pu et al., 2021) proposes a two-stage deep neural network ensemble model, based on DNN, DNN-DTIs (Chen et al., 2021) employs layer-by-layer learning method to predict DTIs. Besides, DeepACTION (Hasan Mahmud et al., 2020), AutoDTI++ (Sajadi et al., 2021), GCNMK (Wang et al., 2022) and DeepStack-DTIs (Zhang et al., 2022) also use deep learning methods.

Specially, inspired by DNN, Zhou and Feng (2017) proposed Deep Forest, and some DTI prediction methods based on Deep Forest showed good performance. Such as AOPEDF (Zeng et al., 2020a), DTI-CDF (Chu et al., 2021a) and EC-DFR (Lin et al., 2022).

In this study, we make some improvements based on the AOPEDF model, thus proposing a new method termed LGBMDF. We add LightGBM (Ke et al., 2017), which outperforms XGBoost and CatBoost in another work (Al Daoud, 2019), to Cascade Forest as a new estimator. For the convenience of comparison, we used the same feature extraction method as AOPEDF. For the obtained vector features, we input them into a modified Cascade Forest for predicting DTIs. Finally, we compared our model with other models in terms of performance and speed, our model is comparable to and in some way ahead of the state-of-the-art models. In conclusion, LGBMDF is a very practical method for DTI prediction, which can help new drug development and some other fields, such as identifying miRNA-disease associations or the associations between cancers and microbes.

2. Materials and methods

2.1. Data resource

DTI-related information was collected from DrugBank (v4.2) (Wishart et al., 2018), the Therapeutic Target Database (Yang et al., 2016), and the PharmGKB (Hernandez-Boussard et al., 2007) database. Bioactivity data for drug–target pairs are collected from ChEMBL (v20) (Gaulton et al., 2012), BindingDB (Liu et al., 2007), and IUPHAR/BPS Guide to PHARMACOLOGY (Pawson et al., 2014). The chemical structure of each drug with SMILES format is extracted from DrugBank (v4.0) (Law et al., 2014). Here, only DTIs meeting the following three criteria are used: (i) the human target is represented by a unique UniProt (Apweiler et al., 2004) accession number; (ii) the target is marked as 'reviewed' in the UniProt database; (iii) binding affinities, all the $K_i, K_d, IC50\ or\ EC50 \leq 10\mu M$. In short, we constructed a DTI network by using 732 FDA-approved drugs and 1915 targets. In addition, we used 9 drug-related networks and 6 protein-related networks (Cheng et al., 2019a,b; Zeng et al., 2020a). For the feature extraction approach, in order to facilitate comparison, we referred to the previous studies (Zhang et al., 2018; Zeng et al., 2020a).

2.2. Deep forest

The deep neural network has shown good performance in many works. Inspired by DNN, Zhou and Feng (2017) proposed an ensemble algorithm with deep structure based on decision tree. It has much fewer hyperparameters than DNNs, and the complexity of the model can be automatically determined based on the input variables.

After obtaining low-dimensional vector representations of drugs and proteins (targets), we input them into Cascade Forest to predict DTIs. In the cascade structure, the output features vector of the previous layer and the original features vector is used as the

input features vector of the next layer. Furthermore, when a new layer is generated, the performance of the entire cascade is estimated on the validation set, and the training process is terminated if there is no significant increase in performance. The estimators setting at each layer are also important, after experimental testing, we set up three ExtraTrees and three LightGBMs (Figure 1).

To prevent overfitting, class vectors for each estimator are generated by *k*-fold cross-validation. Specifically, the average of the generated *k*-1 class vectors is obtained to obtain the final class vector as the enhanced feature of the next layer.

2.3. LightGBM classifier

2.3.1. Histogram algorithm

The basic idea: First, the continuous floating-point feature values are discretized into k integers, and a histogram of width k is constructed (Figure 2). When the samples are traversed once, the histogram accumulates the required statistics and then traverses the histogram to find the optimal partition point based on the discrete values of the histogram.

Another improved speedup of LightGBM is to subtract the histogram of sibling nodes from the histogram of the parent node so that the speed can be doubled (Figure 3). Usually, when constructing a histogram, it is necessary to traverse all the data on that leaf, but histogram differencing only requires traversing k bins of the histogram. In the actual process of constructing the tree, LightGBM can also calculate the smaller leaf nodes of the histogram first, and then use histogram difference to obtain the larger leaf nodes of the histogram, so that we can get the histogram of its sibling leaf at a very small cost.

2.3.2. Leaf-wise algorithm with depth restriction

Based on the histogram algorithm, LightGBM is further optimized. First, it abandons the level-wise (Figure 4A) tree growth strategy used by most GBDT algorithms and applies the leaf-wise tree growth (Figure 4B) with depth restriction. XGBoost uses level-wise growth strategy, which can split the leaves of the same level at the same time by traversing the data once, making it easy to perform multi-threaded optimization and control the model complexity without overfitting. However, level-wise is an inefficient algorithm because it treats the leaves of the same layer indiscriminately, and in fact, many leaves have low splitting gain, so there is no need to split, thus bringing a lot of unnecessary computational overhead. LightGBM uses

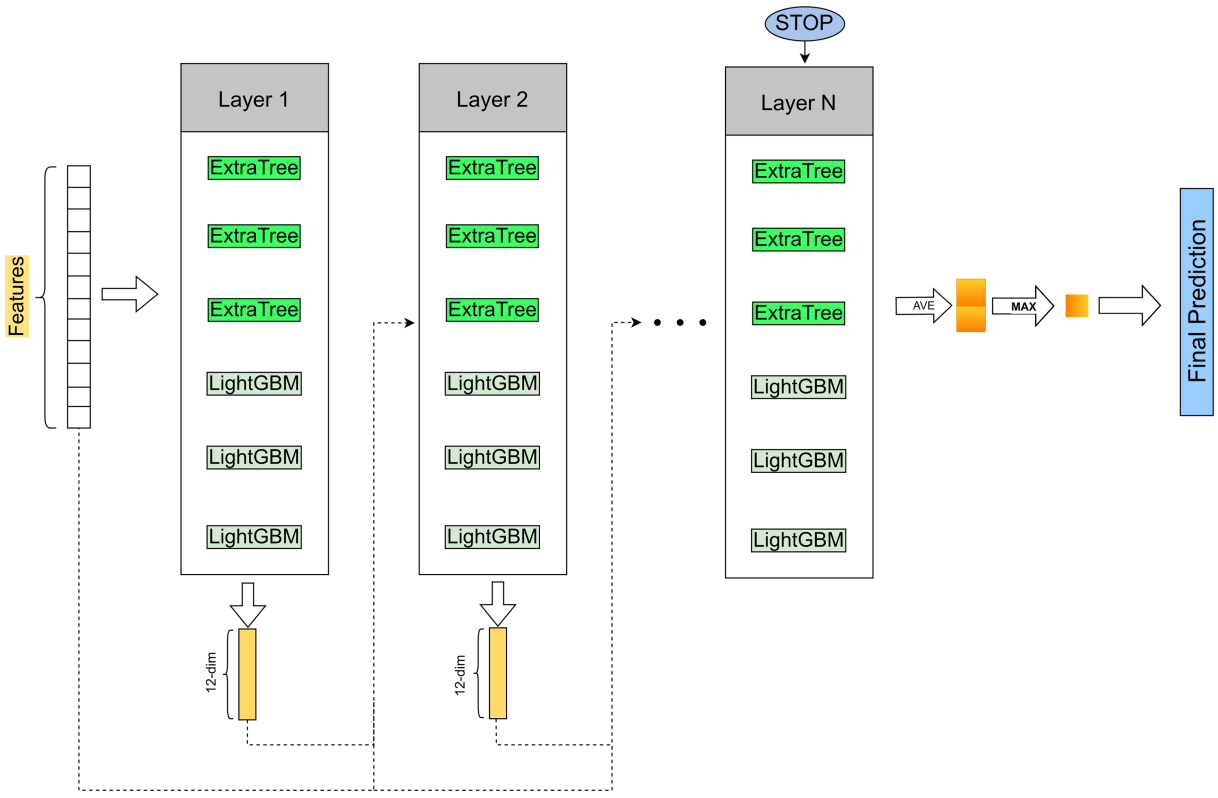


FIGURE 1
The pipeline of LGBMDF. After getting the features of drugs and targets, we process these features with cascade forest, and set 3 LightGBMs and 3 ExtraTrees for each level as estimators, each estimator outputs a 2-dimensional class vector, and then concatenate the output class vector and the original feature vector as the input vector for the next layer.

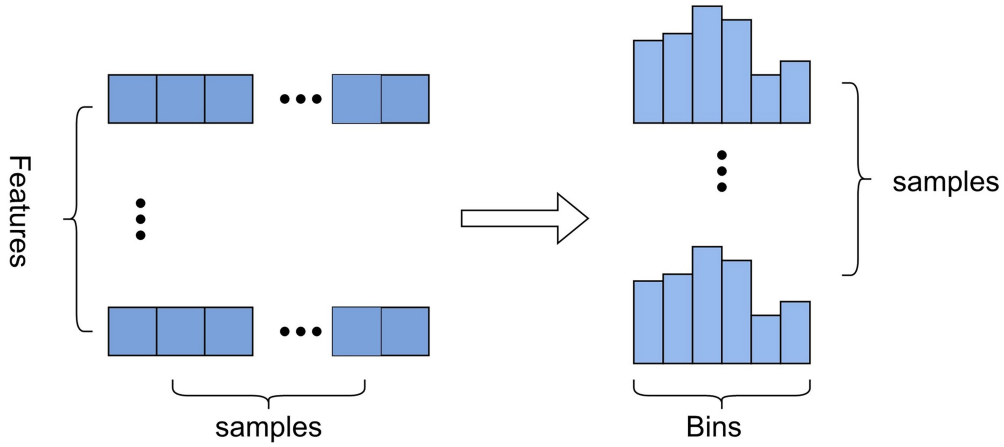


FIGURE 2
The construction of histogram.

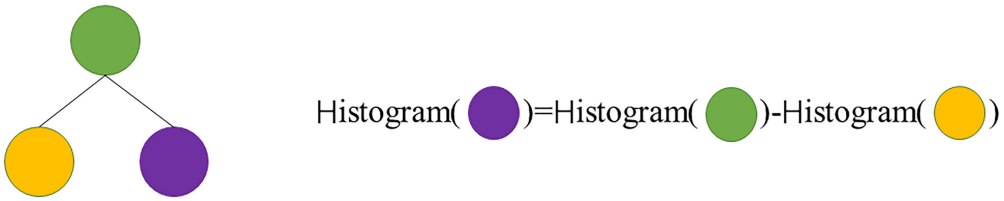


FIGURE 3
Subtract the histogram of sibling node from the histogram of the parent node so that the speed can be doubled.

leaf-wise tree growth strategy, which can locate the leaf with the largest splitting gain from all the current leaves, and then splits it, cycling as this way. Therefore, compared with level-wise, the advantage of leaf-wise is that it can reduce more errors and get better accuracy with the same number of splits; the disadvantage of leaf-wise is that it may grow a deeper decision tree and produce overfitting. For this reason, LightGBM adds a maximum depth limit to leaf-wise to ensure high efficiency and prevent overfitting at the same time.

2.3.3. Gradient-based one-side sampling

The feature vector in Adaboost can represent the importance of a sample well, but there is not a weight vector like this one in GBDT. Fortunately, we found that the sample gradient of GBDT is a good indicator, and samples with small gradients will have small training errors and have been well-trained. Generally, the simpler idea is to discard samples with small gradients, but this will affect the model performance, thus we propose a new method named gradient-based one-side sampling (GOSS).

The basic idea of GOSS is to reduce the complexity of the model by reducing the sample size. GOSS first sorts the samples by the gradient from largest to smallest, uses the top-ranked $a\times100\%$, and then randomly samples the rest data with small gradients $b\times100\%$. Then GOSS amplifiers the data with a small gradient by a constant $\frac{1-a}{b}$ when calculating the information gain.

In GBDT, we assume the input space as X^s, the gradient space as G. Suppose that there are n i.i.d instances $\{x_1,x_2,\cdots,x_n\}$, x_i is a vector of dimension s in X^s. The negative gradient of the loss function is represented as $\{g_1,g_2,\cdots,g_n\}$. The Decision tree model splits nodes where information gain is the largest, and the information gain is usually determined by the variance after the split.

Let $\mathcal{O}$ be the training set of a node d on the decision tree, and the variance of the split feature j at this point is defined as:

$$V_{j|\mathcal{O}}(d)=\frac{1}{n_{\mathcal{O}}}\left(\frac{\left(\sum_{\{x_i\in\mathcal{O}:x_{ij}\le d\}}g_i\right)^2}{n^j_{l|\mathcal{O}}(d)}+\frac{\left(\sum_{\{x_i\in\mathcal{O}:x_{ij}>d\}}g_i\right)^2}{n^j_{r|\mathcal{O}}(d)}\right) \quad (1)$$

Where $n_{\mathcal{O}}=\sum I[x_i\in\mathcal{O}], n^j_{l|\mathcal{O}}(d)=\sum I[x_i\in\mathcal{O}:x_{ij}\le d]$ and $n^j_{r|\mathcal{O}}(d)=\sum I[x_i\in\mathcal{O}:x_{ij}>d]$.

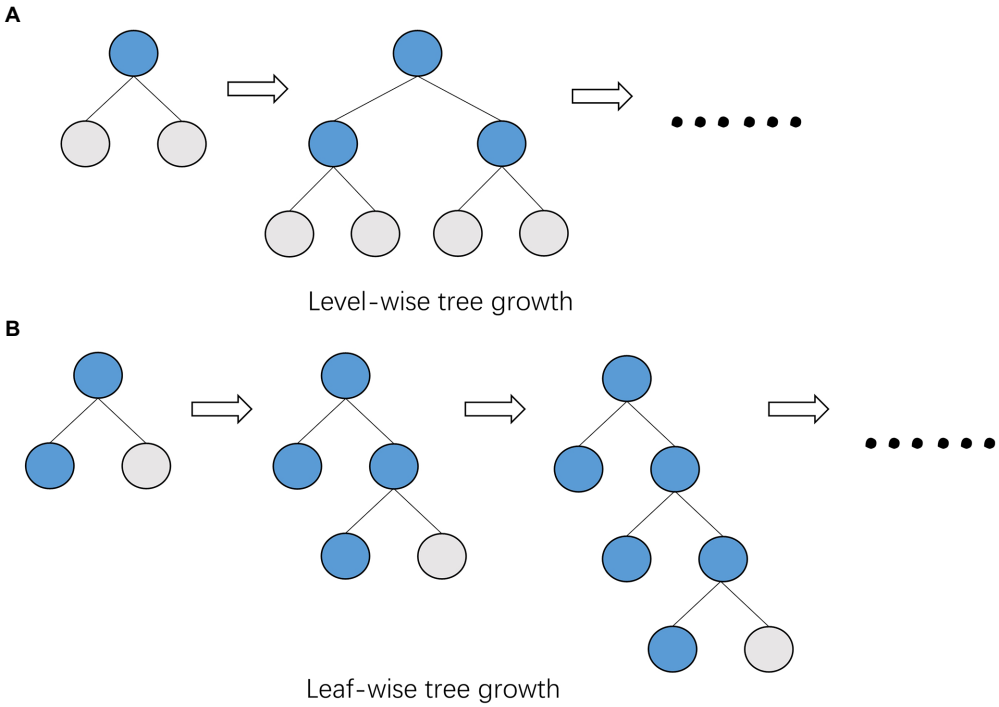


FIGURE 4
Comparison of tree growth patterns between XGBoost and LightGBM. **(A)** XGBoost uses the level-wise growth strategy, which can split the leaves of the same level at the same time by traversing the data once. **(B)** LightGBM uses the leaf-wise growth strategy, which finds the leaf with the largest splitting gain from all the current leaves, and then splits it.

In GOSS, First, all instances absolute values of gradients are sorted in descending order. We select the first $a \times 100\%$ samples as set A, and then randomly sample B of size $b \times \#A^c$ | from the remaining instance set A^c. Finally, we split the instance *via* estimated variance $\widetilde{V}_j(d)$ on $A \cup B$.

$$\widetilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i\right)^2}{n_r^j(d)}\right) \quad (2)$$

Where $A_l = \{x_i \in A : x_{ij} \le d\}, A_r = \{x_i \in A : x_{ij} > d\}$, $B_l = \{x_i \in B : x_{ij} \le d\}, B_r = \{x_i \in B : x_{ij} > d\}$. $\frac{1-a}{b}$ is to normalize the size of B to the size of A^c.

2.3.4. Exclusive feature bundling

High-dimensional space is always sparse, and in a sparse feature space, many features are mutually exclusive, so we can bind mutually exclusive features into a single feature (Figure 5). Through the feature scanning algorithm, we can use the designed feature scanning algorithm to construct the same histogram from the feature bundles as the original single feature. In this way, we can decrease the complexity of histogram building from $O(\#sample \times \#feature)$ to $O(\#sample \times \#bundle)$, while $\#bundle \ll \#feature$, thus we can greatly improve the training speed of GBDT.

In general, compare to XGBoost, LightGBM has the advantages of faster speed and smaller memory usage. LightGBM uses the histogram algorithm to transform the traversal samples into traversal histograms, which greatly reduces the time complexity; applies the GOSS algorithm to filter out many samples with small gradients and adopts leaf-wise growth strategy to build the trees, which reduces a lot of unnecessary calculations. In addition, LightGBM utilizes EFB algorithm to decrease the number of features.

2.4. Evaluation metric

To compare with other methods, we perform a 5-fold cross-validation and adopt Sn, Sp, MCC, AUC and AUPR as evaluation metrics.

Sn, Sp and MCC are commonly used evaluation indicators for binary classification problems, and their calculations are based on the confusion matrix.

$$S_n = \frac{TP}{TP + FN} \quad (3)$$

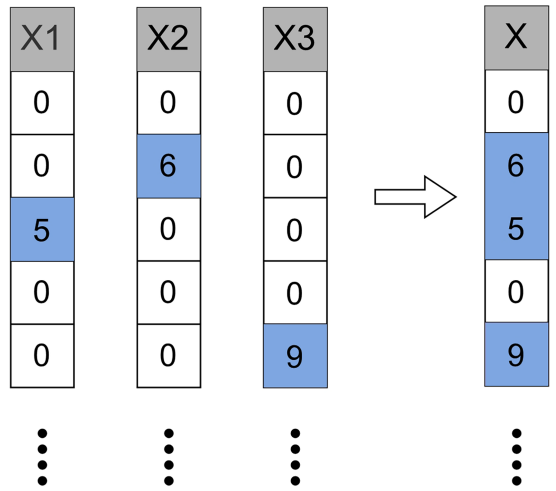


FIGURE 5
Bind mutually exclusive features into a single feature.

$$S_p = \frac{TN}{TN + FP} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

Receiver operating characteristic (ROC) curve is often used to evaluate the model's prediction performance. It is calculated based on the confusion matrix. The higher the curve on the upper left, the better the performance of the model. The vertical axis of the ROC curve is the "True Positive Rate," and the horizontal axis is the "False Positive Rate," which are, respectively, defined as:

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{TN + FP} \tag{7}$$

However, the ROC curves of some models will cross, so we generally choose the AUC (Area Under ROC Curve) for comparison. We assume that the points of the ROC curve are connected in order by the points of $\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$, then the AUC can be estimated as:

$$AUC = \frac{1}{2}\sum_{i=0}^{m-1}(x_{i+1} - x_i) \bullet (y_{i+1} - y_i) \tag{8}$$

The PR curve represents the relationship between Precision and Recall. In general, Recall is set to the abscissa and Precision is set to the ordinate. Precision and Recall can be calculated according to the confusion matrix.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

AUPR is the Area Under PR curve. In such a highly imbalanced dataset, AUPR can provide better performance evaluation because it penalizes false positives more severely.

3. Results

3.1. Parameter optimization

We optimized the parameters of the estimators, considering the impact of parameters on model performance. By the means of employing GridSearchCV function, we set the interval of the parameter, the "scoring" is set as "accuracy." The parameter optimization results are shown in Table 1.

3.2. Estimators setting for each layer

When reproducing the AOPEDF model, we noticed that the XGBoost in cascade is time-consuming, so we chose LightGBM, a classifier that performs better than XGBoost in another work (Al Daoud, 2019), as estimator to accelerate the calculation speed of the model and reduce the computing cost and time cost. We tested five combinations and compared their Sn, Sp, MCC, AUC, AUPR (Table 2) and running time. The experiments are run in the environment of Python3.9, CPU: 2* Intel (R) Xeon (R) Gold 6320R, RAM: 128G.

The names of each combination in the Figure 6 are explained as follows:

- *AOPEDF*: 2 ExtraTrees, 2 RFs and 2 XGBoosts
- *2LGB-2RF-2ET*: 2 LightGBMs, 2 RFs and 2 ExtraTrees
- *3LGB-3RF*: 3 LightGBMs and 3 RFs
- *3LGB-3ET*: 3 LightGBMs and 3 ExtraTrees.

After experiments, we found that the MCC, AUC and AUPR values of 3LGB-3ET are higher than that of the others. Moreover, the calculation speed of 3LGB-3ET is more than twice as fast as AOPEDF. Therefore, we choose the combination of 3LGB-3ET to set the estimators for each layer finally.

3.3. Model comparison

The following 4 models were adopted as baseline methods.

NEDTP (An and Yu, 2021): A node similarity network is constructed based on 15 heterogeneous information networks, and then random walks are applied to extract the topology information of each node in the network and learn it as a low-dimensional vector. Finally, employ LightGBM algorithm to complete the classification task.

AOPEDF (Zeng et al., 2020a): It integrates 15 biological networks to construct a heterogeneous network, and then learns low-dimensional vector representations of features from this heterogeneous network that keep arbitrary-order proximity. Then use the deep forest to predict new DTIs.

Random Forest (Breiman, 2001): It is a combination of tree predictors such that each tree depends on the value of an independently sampled random vector and all trees in the forest have the same distribution.

Support Vector Machine, SVM (Vapnik and Chervoneva, 1964): It is a class of generalized linear classifiers for binary classification of data in a supervised learning manner.

TABLE 1 The result of parameter optimization.

Model	Parameter	Range	Used
RandomForest	n_estimators	[100, 200, 400, 500, 600]	400
LightGBM	n_estimators	[100, 200, 400, 500]	400
	max_depth	[7, 8, 9, 10, 11]	11
	num_leaves	[100, 200, 300, 400, 500]	200
ExtraTree	n_estimators	[100, 200, 400, 500, 600]	500

TABLE 2 Performance comparison under each estimator setting.

Estimators	Sn	Sp	MCC	AUC	AUPR
AOPEDF	**0.9463**	0.9447	0.8911	0.9842	0.9855
2LGB-2RF-2ET	0.9439	**0.9477**	0.8918	0.9841	0.9854
3LGB-3RF	0.9443	0.9453	0.8898	0.9839	0.9849
3LGB-3ET	0.9451	0.9471	**0.8924**	**0.9844**	**0.9857**

The bold values represent the maximum value of each estimator setting under each evaluation metric.

We took drug-protein pairs with known interactions as positive samples, and pairs with unknown interactions as negative samples, and then selected all positive samples and randomly sampled negative samples with the same number of positive samples for 5-fold cross-validation to evaluate model performance (Figure 7, Table 3). For each 5-fold cross-validation, we select 80% positive pairs and the corresponding number of randomly sampled negative pairs as the training set, and the remaining 20% positive pairs and the corresponding number of randomly sampled negative pairs as the test set. We found that the Sp, MCC, AUC, and AUPR of LGBMDF are all higher than those of other methods. In addition, in previous experiments, we have found that LGBMDF is faster than AOPEDF. An excellent model needs to consider both the accuracy and the computing power cost of the model. Therefore, our model is better than the current advanced model in general.

4. Discussion

This paper investigated the application of machine learning methods for DTI prediction. Traditional drug-target effect testing methods are time-consuming and labor-intensive. And Machine learning methods have attracted the attention of many researchers due to these methods can greatly reduce the related costs. We chose the same feature extraction method as AOPEDF, and used this method to extract low-dimensional representations of drug and protein features from 15 biological networks, and these features maintain arbitrary order proximity.

After obtaining low-dimensional feature representations of drugs and targets, we used cascaded deep forests for DTI prediction. Specifically, we used LightGBM as the estimator in the cascade to reduce the computational cost. And the LightGBM has shown better performance and computational speed than XGBoost in other experiments. Considering the effect of estimator diversity in the cascade, we also chose ExtraTree as the estimator.

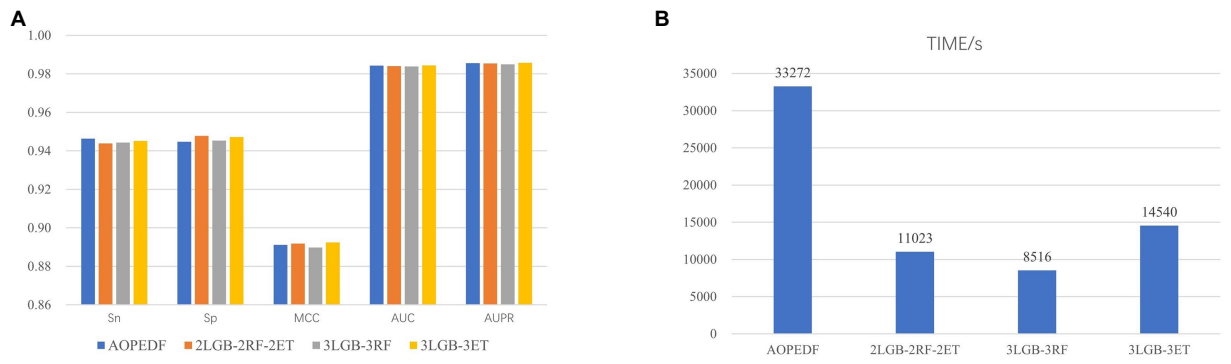


FIGURE 6
Model performance comparison under each estimator setting. **(A)** AUC and AUPR for 4 estimator combinations. **(B)** Computational time for 4 estimator combinations.

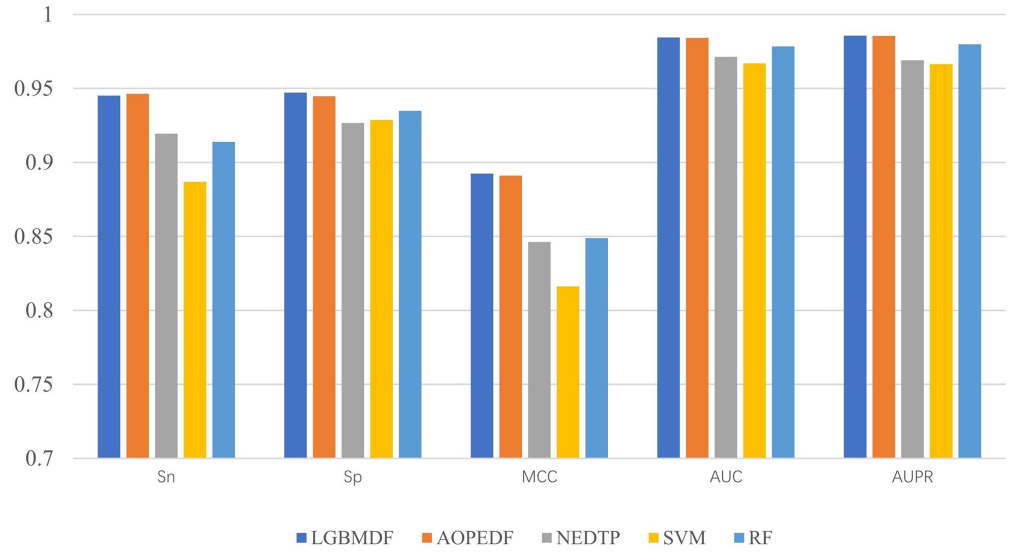


FIGURE 7
Sn, Sp, MCC, AUC and AUPR of LGBMDF, AOPEDF, NEDTP, RF, SVM.

TABLE 3 Performance of LGBMDF and baseline methods.

Model	Sn	Sp	MCC	AUC	AUPR
LGBMDF	0.9451	**0.9471**	**0.8924**	**0.9844**	**0.9857**
AOPEDF	**0.9463**	0.9447	0.8911	0.9842	0.9855
NEDTP	0.9194	0.9267	0.8462	0.9714	0.9690
SVM	0.8869	0.9286	0.8162	0.9668	0.9664
RF	0.9138	0.9348	0.8488	0.9784	0.9798

The bold values represent the maximum value of each estimator setting under each evaluation metric.

By comparing the Sn, Sp, MCC, AUC, AUPR and computation time of the 4 estimator combinations, we chose three ExtraTrees and three LightGBMs as estimators at each layer, and then utilized this cascade forest for DTI prediction. To demonstrate the merits of our model, we compared it with other four baseline models on the same dataset. After 5-fold cross-validation, we obtained the Sn, Sp, MCC, AUC and AUPR of the five models, the Sp (0.9471), MCC (0.8924), AUC (0.9844) and AUPR (0.9857) of LGBMDF were higher than AOPEDF, NEDTP, RF and SVM. The Sn (0.9451) was slightly inferior to AOPEDF, but higher than other three methods. Furthermore, the calculation time of LGBMDF was less than half of that of AOPEDF.

In summary, the method proposed in this paper shows higher prediction accuracy with the current state-of-the-art methods, and greatly improves the computational speed. We believe this will accelerate the drug development process to a certain extent. Certainly, there are still some shortcomings in this paper, such as feature extraction method. We believe that if there is a better way to extract features, the prediction accuracy will also be improved. Moreover, our method could also be applied in other studies, such as in exploring the link between microbes and cancer.

Data availability statement

The data and code for LGBMDF is available at https://github.com/TLanCZ/LGBMDF.

Author contributions

YP proposed the model and completed the manuscript writing. ZZ and XH assisted in completing the model construction. SZ and ZY reviewed and revised the manuscript. ZY provided financial support. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by National Natural Science Foundation of China (no: 62072296).

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

References

Al Daoud, E. (2019). Comparison between XGBoost, light GBM and cat boost using a home credit dataset. *Int. J. Comput. Inf. Eng.* 13, 6–10. doi: 10.5281/zenodo.3607805

An, Q., and Yu, L. (2021). A heterogeneous network embedding framework for predicting similarity-based drug-target interactions. *Brief. Bioinform.* 22:bbab275. doi: 10.1093/bib/bbab275

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). Uni Prot: the universal protein knowledgebase. *Nucleic Acids Res.* 32, 115D–1119D. doi: 10.1093/nar/gkh131

Bagherian, M., Kim, R. B., Jiang, C., Sartor, M. A., Derksen, H., and Najarian, K. (2021). Coupled matrix–matrix and coupled tensor–matrix completion methods for predicting drug–target interactions. *Brief. Bioinform.* 22, 2161–2171. doi: 10.1093/bib/bbaa025

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cao, D. S., Zhang, L. X., Tan, G. S., Xiang, Z., Zeng, W. B., Xu, Q. S., et al. (2014). Computational prediction of drug target interactions using chemical, biological, and network features. *Mol Inform* 33, 669–681. doi: 10.1002/minf.201400009

Chen, X., Liu, M.-X., and Yan, G.-Y. (2012). Drug–target interaction prediction by random walk on the heterogeneous network. *Mol. Bio Syst.* 8, 1970–1978. doi: 10.1039/c2mb00002d

Chen, C., Shi, H., Jiang, Z., Salhi, A., Chen, R., Cui, X., et al. (2021). DNN-DTIs: improved drug-target interactions prediction using XGBoost feature selection and deep neural network. *Comput. Biol. Med.* 136:104676. doi: 10.1016/j.compbiomed.2021.104676

Chen, M., and Yin, Z. (2022). Classification of Cardiotocography based on Apriori algorithm and multi-model ensemble classifier. *Front. Cell Dev. Biol.* 10:888859. doi: 10.3389/fcell.2022.888859

Cheng, F., Kovács, I. A., and Barabási, A.-L. (2019a). Network-based prediction of drug combinations. *Nat. Commun.* 10, 1–11.

Cheng, F., Lu, W., Liu, C., Fang, J., Hou, Y., Handy, D. E., et al. (2019b). A genome-wide positioning systems network algorithm for in silico drug repurposing. *Nat. Commun.* 10, 1–14. doi: 10.1038/s41467-019-10744-6

Chu, Y., Kaushik, A. C., Wang, X., Wang, W., Zhang, Y., Shan, X., et al. (2021a). DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. *Brief. Bioinform.* 22, 451–462. doi: 10.1093/bib/bbz152

Chu, Y., Shan, X., Chen, T., Jiang, M., Wang, Y., Wang, Q., et al. (2021b). DTI-MLCD: predicting drug-target interactions using multi-label learning with community detection method. *Brief. Bioinform.* 22:bbaa205. doi: 10.1093/bib/bbaa205

Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. *Inf. Sci.* 418-419, 546–560. doi: 10.1016/j.ins.2017.08.045

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107.

Gönen, M. (2012). Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* 28, 2304–2310. doi: 10.1093/bioinformatics/bts360

Guo, F., Yin, Z., Zhou, K., and Li, J. (2021). PLncWX: a machine-learning algorithm for plant lncRNA identification based on WOA-XGBoost. *J. Chem.* 2021, 1–11. doi: 10.1155/2021/6256021

Hasan Mahmud, S. M., Chen, W., Jahan, H., Dai, B., Din, S. U., and Dzisoo, A. M. (2020). DeepACTION: a deep learning-based method for predicting novel drug-target interactions. *Anal. Biochem.* 610:113978. doi: 10.1016/j.ab.2020.113978

Hernandez-Boussard, T., Whirl-Carrillo, M., Hebert, J. M., Gong, L., Owen, R., Gong, M., et al. (2007). The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res.* 36, D913–D918. doi: 10.1093/nar/gkm1009

Jarada, T. N., Rokne, J. G., and Alhajj, R. (2021). SNF–CVAE: computational method to predict drug–disease interactions using similarity network fusion and collective variational autoencoder. *Knowl. Based Syst.* 212:106585. doi: 10.1016/j.knosys.2020.106585

Jin, S., Niu, Z., Jiang, C., Huang, W., Xia, F., Jin, X., et al. (2021). HeTDR: drug repositioning based on heterogeneous networks and text mining. *Patterns (N Y)* 2:100307. doi: 10.1016/j.patter.2021.100307

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Proces. Syst.* 30, 3149–3157.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42, D1091–D1097. doi: 10.1093/nar/gkt1068

Li, Y., Liu, X. Z., You, Z. H., Li, L. P., Guo, J. X., and Wang, Z. (2020). A computational approach for predicting drug–target interactions from protein sequence and drug substructure fingerprint information. *Int. J. Intell. Syst.* 36, 593–609. doi: 10.1002/int.22332

Lihong, P., Wang, C., Tian, X., Zhou, L., and Li, K. (2021). Finding lncRNA-protein interactions based on deep learning with dual-net neural architecture. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 19, 3456–3468. doi: 10.1109/TCBB.2021.3116232

Lin, W., Wu, L., Zhang, Y., Wen, Y., Yan, B., Dai, C., et al. (2022). An enhanced cascade-based deep forest model for drug combination prediction. *Brief. Bioinform.* 23:bbab562. doi: 10.1093/bib/bbab562

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., and Gilson, M. K. (2007). BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* 35, D198–D201. doi: 10.1093/nar/gkl999

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

Mei, J. P., Kwoh, C. K., Yang, P., Li, X. L., and Zheng, J. (2013). Drug-target interaction prediction by learning from local information and neighbors. *Bioinformatics* 29, 238–245. doi: 10.1093/bioinformatics/bts670

Mousavian, Z., Khakabimamaghani, S., Kavousi, K., and Masoudi-Nejad, A. (2016). Drug-target interaction prediction from PSSM based evolutionary information. *J. Pharmacol. Toxicol. Methods* 78, 42–51. doi: 10.1016/j.vascn.2015.11.002

Olayan, R. S., Ashoor, H., and Bajic, V. B. (2018). DDR: efficient computational method to predict drug-target interactions using graph mining and machine learning approaches. *Bioinformatics* 34, 1164–1173. doi: 10.1093/bioinformatics/btx731

Pawson, A. J., Sharman, J. L., Benson, H. E., Faccenda, E., Alexander, S. P., Buneman, O. P., et al. (2014). The IUPHAR/BPS guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Res.* 42, D1098–D1106. doi: 10.1093/nar/gkt1143

Peng, J., Wang, Y., Guan, J., Li, J., Han, R., Hao, J., et al. (2021). An end-to-end heterogeneous graph representation learning-based framework for drug-target interaction prediction. *Brief. Bioinform.* 22:bbaa430. doi: 10.1093/bib/bbaa430

Peng, L., Wang, F., Wang, Z., Tan, J., Huang, L., Tian, X., et al. (2022). Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies. *Brief. Bioinform.* 23:bbac234. doi: 10.1093/bib/bbac234

Pliakos, K., Vens, C., and Tsoumakas, G. (2019). Predicting drug-target interactions with multi-label classification and label partitioning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 1596–1607. doi: 10.1109/TCBB.2019.2951378

Pu, Y., Li, J., Tang, J., and Guo, F. (2021). DeepFusionDTA: drug-target binding affinity prediction with information fusion and hybrid deep-learning ensemble model. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 19, 2760–2769. doi: 10.1109/TCBB.2021.3103966

Sajadi, S. Z., Zare Chahooki, M. A., Gharaghani, S., and Abbasi, K. (2021). AutoDTI++: deep unsupervised learning for DTI prediction by autoencoders. *BMC Bioinformatics* 22:204. doi: 10.1186/s12859-021-04127-2

Shen, L., Liu, F., Huang, L., Liu, G., Zhou, L., and Peng, L. (2022). VDA-RWLRLS: An anti-SARS-CoV-2 drug prioritizing framework combining an unbalanced bi-random walk and Laplacian regularized least squares. *Comput. Biol. Med.* 140:105119. doi: 10.1016/j.compbiomed.2021.105119

Tanoori, B., Jahromi, M. Z., and Mansoori, E. G. (2021). Drug-target continuous binding affinity prediction using multiple sources of information. *Expert Syst. Appl.* 186:115810. doi: 10.1016/j.eswa.2021.115810

Vapnik, V. N., and Chervoneva, A. (1964). On class of perceptrons. *Autom. Remote. Control.* 25:103.

Wang, F., Lei, X., Liao, B., and Wu, F. X. (2022). Predicting drug-drug interactions by graph convolutional network with multi-kernel. *Brief. Bioinform.* 23:bbab511. doi: 10.1093/bib/bbab511

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Yamanishi, Y., Kotera, M., Kanehisa, M., and Goto, S. (2010). Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 26, i246–i254. doi: 10.1093/bioinformatics/btq176

Yang, H., Qin, C., Li, Y. H., Tao, L., Zhou, J., Yu, C. Y., et al. (2016). Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res.* 44, D1069–D1074. doi: 10.1093/nar/gkv1230

Yang, Z., Zhong, W., Zhao, L., and Yu-Chian Chen, C. (2022). MGraphDTA: deep multiscale graph neural network for explainable drug-target binding affinity prediction. *Chem. Sci.* 13, 816–833. doi: 10.1039/d1sc05180f

You, J., McLeod, R. D., and Hu, P. (2019). Predicting drug-target interaction network using deep learning model. *Comput. Biol. Chem.* 80, 90–101. doi: 10.1016/j.compbiolchem.2019.03.016

Yuan, Q., Gao, J., Wu, D., Zhang, S., Mamitsuka, H., and Zhu, S. (2016). DrugE-rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics* 32, i18–i27. doi: 10.1093/bioinformatics/btw244

Zeng, X., Zhu, S., Hou, Y., Zhang, P., Li, L., Li, J., et al. (2020a). Network-based prediction of drug-target interactions using an arbitrary-order proximity embedded deep forest. *Bioinformatics* 36, 2805–2812. doi: 10.1093/bioinformatics/btaa010

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

Zhan, X., You, Z., Yu, C., Li, L., and Pan, J. (2020). Ensemble learning prediction of drug-target interactions using GIST descriptor extracted from PSSM-based evolutionary information. *Biomed. Res. Int.* 2020, 4516250–4516210. doi: 10.1155/2020/4516250

Zhang, Z., Cui, P., Wang, X., Pei, J., Yao, X., and Zhu, W. (2018). "Arbitrary-Order Proximity Preserved Network Embedding", In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Zhang, Y., Jiang, Z., Chen, C., Wei, Q., Gu, H., and Yu, B. (2022). DeepStack-DTIs: predicting drug-target interactions using LightGBM feature selection and deep-stacked ensemble classifier. *Interdiscip. Sci.* 14, 311–330. doi: 10.1007/s12539-021-00488-7

Zhou, Z.-H., and Feng, J. (2017). "Deep Forest: Towards An Alternative to Deep Neural Networks", in: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. 3553–3559.

Zhou, L., Li, Z., Yang, J., Tian, G., Liu, F., Wen, H., et al. (2019). Revealing drug-target interactions with computational models and algorithms. *Molecules* 24:1714. doi: 10.3390/molecules24091714

Zhou, L., Wang, Z., Tian, X., and Peng, L. (2021). LPI-deepGBDT: a multiple-layer deep framework based on gradient boosting decision trees for lncRNA–protein interaction identification. *BMC Bioinformatics* 22, 1–24. doi: 10.1186/s12859-021-04399-8

Zhou, K., Yin, Z., Peng, Y., and Zeng, Z. (2022). Methods for continuous blood pressure estimation using temporal convolutional neural networks and ensemble empirical mode decomposition. *Electronics* 11:1378. doi: 10.3390/electronics11091378

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 27 January 2023
DOI 10.3389/fmicb.2023.1092143

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Guohua Huang,
Shaoyang University,
China
Zhen Tang,
Shanghai Jiao Tong University,
China

*CORRESPONDENCE
ZhiXiang Yin
✉ zxyin66@163.com

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

RECEIVED 07 November 2022
ACCEPTED 11 January 2023
PUBLISHED 27 January 2023

CITATION
Jia X, Yin Z and Peng Y (2023) Gene differential co-expression analysis of male infertility patients based on statistical and machine learning methods.
Front. Microbiol. 14:1092143.
doi: 10.3389/fmicb.2023.1092143

Gene differential co-expression analysis of male infertility patients based on statistical and machine learning methods

Xuan Jia, ZhiXiang Yin* and Yu Peng

School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai, China

Male infertility has always been one of the important factors affecting the infertility of couples of gestational age. The reasons that affect male infertility includes living habits, hereditary factors, etc. Identifying the genetic causes of male infertility can help us understand the biology of male infertility, as well as the diagnosis of genetic testing and the determination of clinical treatment options. While current research has made significant progress in the genes that cause sperm defects in men, genetic studies of sperm content defects are still lacking. This article is based on a dataset of gene expression data on the X chromosome in patients with azoospermia, mild and severe oligospermia. Due to the difference in the degree of disease between patients and the possible difference in genetic causes, common classical clustering methods such as k-means, hierarchical clustering, etc. cannot effectively identify samples (realize simultaneous clustering of samples and features). In this paper, we use machine learning and various statistical methods such as hypergeometric distribution, Gibbs sampling, Fisher test, etc. and genes the interaction network for cluster analysis of gene expression data of male infertility patients has certain advantages compared with existing methods. The cluster results were identified by differential co-expression analysis of gene expression data in male infertility patients, and the model recognition clusters were analyzed by multiple gene enrichment methods, showing different degrees of enrichment in various enzyme activities, cancer, virus-related, ATP and ADP production, and other pathways. At the same time, as this paper is an unsupervised analysis of genetic factors of male infertility patients, we constructed a simulated data set, in which the clustering results have been determined, which can be used to measure the effect of discriminant model recognition. Through comparison, it finds that the proposed model has a better identification effect.

KEYWORDS

male infertility, hypergeometric distribution, Fisher test, Gibbs sampling, machine learning, gene interaction network, HPV

1. Introduction

For a long time, infertility has been a difficult problem for many couples of gestational age. With the increase of life pressure, infertility is increasing every year. About 15% of gestational age couples suffer from infertility symptoms of varying degrees, of which about 50% are caused by male infertility (Dada et al., 2003). About 7% of men in the general population suffer from different degrees of infertility. The causes of male infertility are related to many influencing factors, including different diseases, genetics, living habits and other factors that may cause or interact to cause male infertility. Although men with this disorder cannot pass on their genetic information naturally, genetic factors

can still contribute to male infertility. In approximately 15% of infertile men a genetic defect is most likely the underlying cause of the pathology (Tournaye et al., 2017; Krausz and Riera-Escamilla, 2018). For example, autosomal recessive or X-linked male infertility mutations transmitted by normal parents can cause infertility (Chillón et al., 1995; Yatsenko et al., 2015). Genetic causes have also been found to have an important role in severe male infertility, such as severe oligospermia (<5 million sperm cells per milliliter) or azoospermia (azoospermia in ejaculation; Lopes et al., 2013; Krausz and Riera-Escamilla, 2018). Identifying the genes responsible for male infertility is important for increasing our understanding of the biology of the disease and for genetic testing for diagnosis and clinical treatment. Genes such as NLRP3, BRD7 and others have been shown to affect male fertility (Aquila et al., 2004; Wang et al., 2016; Antonuccio et al., 2021). At the same time, with the rapid development of genetics, more than 3,000 genetic diseases have been discovered, of which about 250 are only found in men, and women have no or little disease. Because women have two X chromosomes, the pathogenic gene on one X chromosome can often be masked by the normal gene on the other X chromosome, so they do not show symptoms. Men, on the other hand, have only one X chromosome. If there is a disease-causing gene on it, there is no corresponding normal gene to cover up, resulting in the disease. In recent years, with the deepening of research, there are about 521 genes that cause male infertility in different forms (Xavier et al., 2021), many of which are related to the X chromosome, such as mouse androgen receptor gene mutation, through chain reaction mapping The X chromosome leads to infertility in mice (Lyon et al., 1970), and there is one more X chromosome in males, that is, the sex chromosome is XXY (Jacobs and Strong, 1959) and so on.

Many scholars have carried out various experimental methods to study the genetic causes of male infertility. Through RNA interference or knockout experiments, the gene cannot be expressed normally, and whether the target abnormality occurs in cells or individuals is observed, and whether the gene is related to the cause of the disease is detected. However, experimental methods are generally time-consuming, labor-intensive, and expensive, and experimental methods are generally designed in a targeted manner on the premise that the experimenter obtains genes that may have basic interference. Technological advances and methodological developments in genomics are critical for identifying genetic factors in male infertility.

In this paper, we use a data set covering all gene expression levels of the male X chromosome in the GEO database, the Gene Expression Omnibus (GEO), a public database that contains 659,203 gene sample data from 9,528 different platforms (Ron et al., 2002). And based on a variety of statistical methods and machine learning analysis of gene expression data of male infertility patients, to identify groups of interacting gene clusters that may contribute to male infertility of various phenotypes in various ways. Common hierarchical clustering, k-means and other clustering algorithms are clustering under the assumption that all samples have certain characteristics, and the cluster data of the identified clusters have the same characteristics in all samples. However, the expression of gene data is affected by different sampling individuals, different tissues of the same individual, etc., resulting in different expression of measured gene data in different samples, and common clustering algorithms cannot meet the identification of differential gene expression modules (implementation basis Partial samples of gene expression data to partition gene sample data). For the identification of differentially co-expressed modules, a biclustering algorithm can be used to screen functionally related genes, genes involved in the same pathway, and genes affected by the same drug or a pathological condition. The biclustering algorithm was first proposed in Hartigan (1972), is a two-dimensional data mining technique that allows simultaneous clustering of rows (representing genes) and columns (representing samples/conditions) in a gene expression matrix. Developments continued in the following decades, with (Cheng and Church, 2000; Lazzeroni and Owen, 2000; Bergmann et al., 2003; Kluger et al., 2003; Chiu et al., 2004; Prelić et al., 2006; Dhollander et al., 2007; Gu and Liu, 2008; Li et al., 2009; Hochreiter et al., 2010; Madeira et al., 2010; Medina et al., 2010; Chen et al., 2011; De Smet and Marchal, 2011; Zhao et al., 2011; Zhou et al., 2012; Goncalves and Madeira, 2014; Henriques and Madeira, 2016a,b; Alzahrani et al., 2017; Guo et al., 2021) being articles on different clustering algorithms. Among them, BCPlaid (Lazzeroni and Owen, 2000), QUBIC (Li et al., 2009), C&C (Cheng and Church, 2000), FABIA (Hochreiter et al., 2010) are the more popular biclustering algorithms. Genomics data analysis clustering using machine learning, deep learning, etc., for identifying cell subpopulations, genomic analysis, etc.(Jiang et al., 2020; Lazareva et al., 2020; Peng et al., 2020; Gerniers et al., 2021; Peng et al., 2021; Yi et al., 2021; Peng et al., 2022; Zhai et al., 2022). Analysis of bronchoalveolar immune cells in COVID-19 patients based on genetic data (Liao et al., 2020). By processing the GSE37948 data set (Krausz et al., 2012), which contains expression levels of gene data on the X chromosome in testicular tissue from patients with varying degrees of infertility, we identified 19 distinct double clusters, indicating the existence of multiple double clusters identified in this paper there are multiple enriched pathways and there are functional and organizational correlations between the enriched pathways. And the performance of the method is verified using a data set similar to the real gene expression level.

2. Materials and methods

2.1. Methods

Rank-rank hyper geometric overlap (RRHO; Plaisier et al., 2010) uses unsupervised learning to sort the gene expression profile data of two samples of different categories, and uses hyper geometric distribution to iteratively calculate the *p*-values of all combinations to find the optimal overlap gene combination. In this paper, the sample expression data of two different genes is brought into the RRHO method to find the optimal overlapping sample set, and the SNR value of the signal-to-noise ratio of the sample gene set is calculated to determine whether the clusters have differential expression. For a single gene in the sample set, the SNR value is defined as:

$$SNR(\mathrm{g},P') = \frac{\mu_{g,P'} - \mu_{g,\overline{P'}}}{\sigma_{\mathrm{g},P'} + \sigma_{\mathrm{g},\overline{P'}}}$$

$\mu_{g,P'}$, $\mu_{g,\overline{P'}}$ are the mean in the delimited sample set P' and the mean in the data outside the sample set, respectively. $\sigma_{\mathrm{g},P'}$, $\sigma_{\mathrm{g},P'}$ represent the standard deviation of the data in the corresponding set. The overall signal-to-noise ratio of the cluster is the average of the signal-to-noise ratios of individual genes in the sample set.

If the signal-to-noise ratio value of the identified sample and gene set is greater than the specified threshold, the set will be retained, and the corresponding genome is considered to have a relationship with the gene data. If one gene cannot form a relationship with other genes in the data,

it will be discarded in the subsequent processing, so as to realize the dimensionality reduction processing of the gene data. However, since the genes known to be associated with disease from Ghiassian et al. (2015) form a compact but not tightly connected subgraph on the PPI, this paper does not loop through all the genes in the data set, but adds a gene interaction network to the data processing. Using the String database, there is known and predicted gene-protein interaction networks in the database. In this paper, the genes involved in the data set are searched for the interaction network, and the isolated gene points are discarded. The genes existing in the gene network are combined in pairs, and the hierarchical clustering method is used for preliminary clustering to assist in determining the default set signal-to-noise ratio threshold. The set of gene samples constructed by preliminary clustering is calculated as the average of the signal-to-noise ratio values in all sets, and 1/2 of this mean is used as the threshold. When the signal-to-noise ratio of the gene sample set constructed by the RRHO method is used. If the ratio is greater than this threshold, the gene is retained and a new set of double clusters is obtained. Otherwise, in the gene network, the connected edges are discarded. Due to the large number of genes, a partial gene network is shown in Figure 1. Figure 2 briefly depicts the model's approach. The interrelation data of all genes are presented in Supplementary Table 1.

Since only gene pairs and their corresponding sample sets can be obtained after using the RRHO method, Gibbs sampling (Sheng et al., 2003) is used for the data processed in the first step to make assumptions about the distribution of gene sample data to merge gene clusters. The statistical assumptions for sampling are as following:

$$x_{ji} \mid \theta_{ic}, s_j \sim \text{Bernoulli}\left(x_{ji} \mid \theta_{is_j}\right)$$

$$\begin{aligned} s_j \mid m &\sim \text{Categorical}\left(s_j \mid m\right) \\ \theta_{ic} &\sim \text{Beta}\left(\alpha / 2\right) \\ m &\sim \text{Dirichlet}\left(\beta / K\right). \end{aligned}$$

i represents the gene, j represents the sample, if the association exists after step 1, x_{ji} is assigned 1 else it is 0. s_j represents which module the gene edge j belongs to, through the calculation of the edge transition probability in Gibbs sampling:

$$\begin{aligned} P\left(s_j = k \mid X, s_{-j}; \alpha, \beta\right) \propto & \\ & \times \prod_{i:x_{ji}=1} \left[\frac{\alpha/2 + \sum_{l:s_l=k, l\neq j} x_{li}}{\alpha + \left|\left\{l : s_l = k, l \neq j\right\}\right|} \right] \\ & \times \frac{\left|\left\{l : s_l = k, l \neq j\right\}\right| + \beta / K}{n - 1 + \beta} \left[\frac{\alpha/2 + \sum_{l:s_l=k, l\neq j} \left(1 - x_{li}\right)}{\alpha + \left|\left\{l : s_l = k, l \neq j\right\}\right|} \right] \end{aligned}$$

Among them, k is set to the number of clusters retained after the calculation and processing of the RRHO method. Finally, the statistical part of Gibbs sampling assumes that the data has a certain prior distribution involving parameter α and β, but because the genetic data lacks the corresponding statistical research foundation, the parameter α and β are set as hyperparameters. At the end of data processing, Fisher's exact test is used to process the calculated set data again, and the sample data in the two clusters are processed to calculate its value of *p*. The set threshold is used to determine whether there is a significant difference between the two sets, and the genes in the two sample sets without significant differences are merged, and the sample data of the corresponding gene is taken out and brought into the hierarchical clustering, and the number of clusters is 2. Since a gene is up-regulated in half of the samples, it will be differentially expressed in the remaining part, so, we limit samples in clusters to less than 55% of the total number of samples in the data set as a difference in the gene set. At the same time, in order to limit that the cluster is differentially expressed in the whole data, the SNR value of the newly formed cluster is required to be greater than the threshold value. Otherwise it will not be merged. All the identified clusters are merged cyclically until no new clusters are generated.

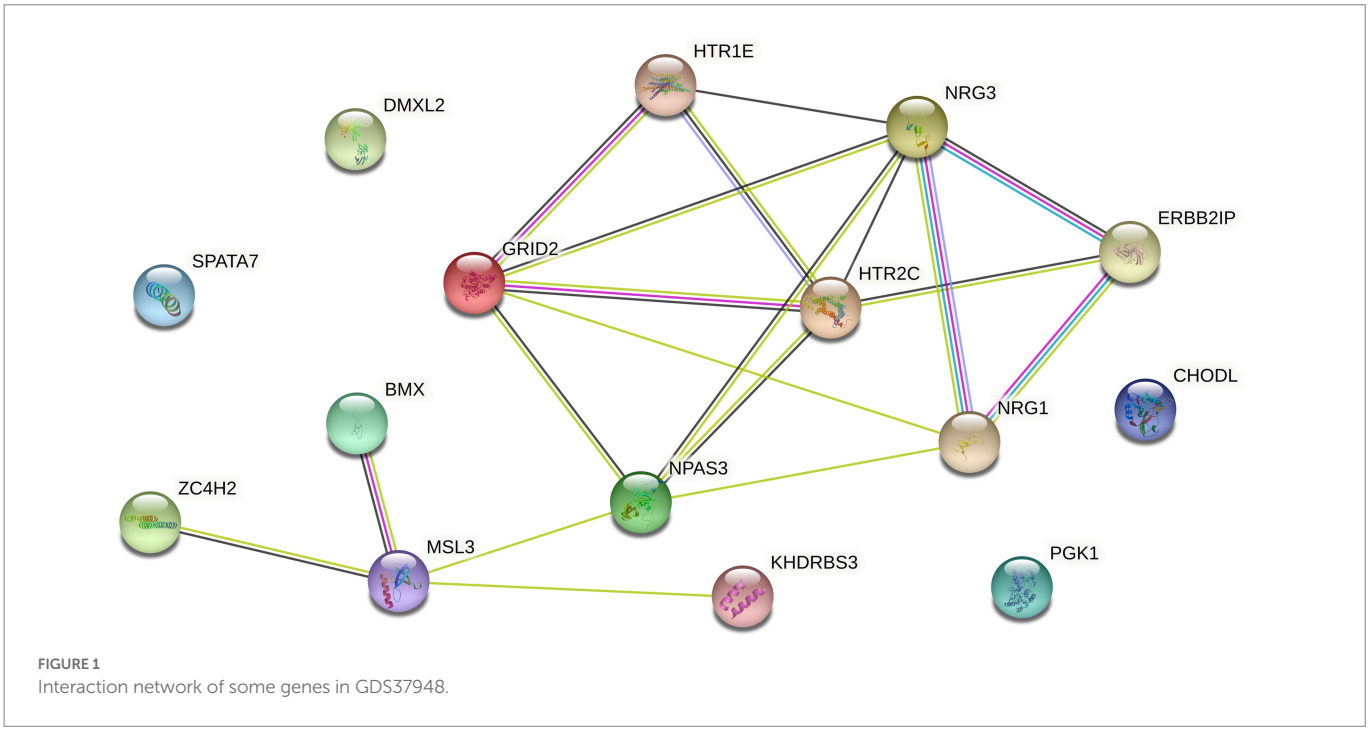


FIGURE 1
Interaction network of some genes in GDS37948.

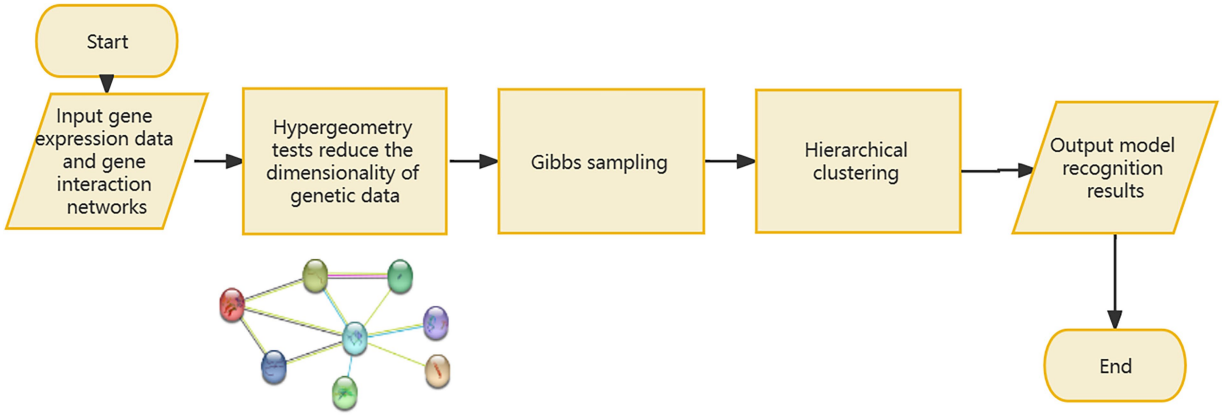


FIGURE 2
Introduction to the model process.

2.2. Datasets

2.2.1. Male infertility gene expression data

First, the corresponding gene expression data were obtained from the micro array gene expression database. In this paper, the GSE37948 (Krausz et al., 2012) gene expression data set was selected. This data set contains relevant gene expression data of 96 patients with different degrees of infertility, including 74 cases of azoospermia, 6 cases of mild oligozoospermia, and 16 cases of severe oligozoospermia. Excluding known causes of impairing spermatogenesis in patients, gene expression data identification was performed using testicular tissue from 47 men, and KNN nearest neighbor algorithm was used to impute missing values in gene expression profile data while normalizing data for each gene, to remove the effect of different units on the data. The GSE37948 data set contains 1855 genes and gene-identified expression data from 200 male sperm samples. The genes identified therein to cover the entire X chromosome. The related gene network based on the GSE37948 data set was extracted from the String database. Specific gene interaction data are shown in the Supplementary Table: Interrelation data among genes.

2.2.2. Synthetic datasets

Since the method in this paper belongs to unsupervised learning, there are no standard results for the study of male infertility-related genes, so we constructed simulation data similar in structure to GSE39748. The GSE37948 data set has a total of 1,855 genes and 200 samples, but the size of the double-cluster deletion is unknown. To this end, simulated data of 20 known differentially expressed modules were constructed with gene and sample dimensions of 2,000 and 200, respectively. Based on previous research (Prelić et al., 2006; Eren et al., 2013), we can generate simulation data according to the following rules: Genes and sample numbers are sampled from (100, 50, 20, 10, 5) and (100, 50, 20, 10) respectively, the data within the cluster is sampled from N (2, 1), and the rest of the data are sampled from N (0, 1) and allow the intersection of different clusters. Simulated data is used to determine hyperparameters and statistics are used to evaluate clustering results. Since the gene interaction network graph used in the gene data processing corresponds to the gene interaction graph with certain connectivity, we correspondingly construct the connected network graph according to the determined clustering data. Studies have shown that in the gene interaction network, genes related to disease can form compact linker maps (Ghiassian et al., 2015), so we use the method proposed in Bollobás et al. (2003) to construct the network diagram, which can construct a reasonable gene network connection map according to the clustering modules in the expression data.

3. Results

3.1. Experimental results of male infertility-related gene expression data

By processing the GSE37948 data set, which contains expression levels of gene data on the X chromosome in testicular tissue from patients with azoospermia, mild and severe oligozoospermia. We identified 19 distinct double clusters. There are multiple enriched pathways and there are functional and organizational correlations between the enriched pathways. The hypergeometric test involved in the RRHO method, in which the significance index is adjusted from the set (0.01, 0.05), and the parameter α and β/k involved in the statistical hypothesis in Gibbs sampling are adjusted from the set (5.0, 1.0, 0.5, 0.1) and (100, 1.0, 0.01), respectively. According to the recognition effect of the model on the simulated data set, the final parameters $p = 0.01$, $\alpha = 0.5$, and $\beta/k = 1.0$ were determined. The data processed based on the GSE39748 data is brought into the model to identify the gene sample module, and the results were analyzed using a variety of biometric indicators Includes: Disease (OMIN_DISEASE, UP_KW_DISEASE), Functional_Annotations (COG_ONTOLO, UP_KW_BIOLOGICAL_PROCESS, UP_KW_CELLOULAR_COMPONENT, UP_KW_MOLECULAR_FUNCTION, UP_KW_PTM, UP_SEQ_FEATURE), Protein_Domains (INTERPRO, PIR_SUPERFAMILY, SMART, UP_KW_DOMAIN), Gene_Ontology (GOTERBP, CC, MF), Interactins (UP_KW_LIGAND), Pathways (KEGG_PATHWAY, BBID,BIOCARTA), Protein_Domains (INTERPRO, PIR_SUPERFAMILY, SMART, UP_KW_DOMAIN).

Corresponding to the Enrichment analysis results with the cluster id of 1 in Table 1, there were four significantly enriched pathways after analysis by GO and KEGG, two of which were associated with proteins of the autism spectrum, which includes different phenotypic manifestations such as classic autism, Asperger's syndrome, childhood

TABLE 1 Clustering results identified in the statistical method proposed in this paper based on the GDS37948 male infertility data set.

ID	avgSNR	Number of samples	Number of samples
1	0.700870148	13	56
2	0.816555484	3	110
3	0.775713429	3	88
4	0.745638081	8	101
5	0.743384851	3	72
6	0.743381552	4	71
7	0.730139247	351	20
8	0.718222619	6	110
9	0.716803164	3	91
10	0.70627255	3	101
11	0.703721749	3	68
12	1.15234204	482	12
13	0.678448517	6	95
14	0.678084094	11	103
15	0.67773126	25	110
16	0.674885829	3	38
17	0.671869245	6	92
18	0.668664873	3	84
19	0.667155842	3	49

disintegration Sexual disorder, Rett's syndrome, and pervasive developmental disorder not otherwise specified. Also significantly enriched into axons, the site of neurotransmitter storage and release. And outside the cytoplasmic membrane, referring to gene products attached to the plasma membrane or cell wall.

Corresponding to the Enrichment analysis results with the cluster id of 2 in Table 1, enriched in chemical synaptic transmission, cell membrane, and plasma membrane pathways. Release of neurotransmitter molecules from presynaptic vesicles across chemical synapses followed by post synaptic activation of neurotransmitter receptors on target cells (neurons, muscles, or secretory cells), and the effect of this activation on synapses Post-membrane potential and ionic composition of the post synaptic cytoplasm. This process includes spontaneous and evoked release of neurotransmitters and all parts of synaptic vesicle exocytosis. Evoked transmission begins when the action potential reaches the presynaptic.

Corresponding to the Enrichment analysis results with the cluster id of 3 in Table 1, by SMART, INTERPRO, UP_KW_DOMAIN showed enrichment to the SH3 domain. The SH3 (src homology-3) domain is a small protein module containing approximately 50 amino acid residues. They are present in a variety of intracellular or membrane-associated proteins, for example, in a variety of proteins with enzymatic activity, in adaptor proteins such as fodrin and the yeast actin-binding protein ABP-1. The SH3 domain has a characteristic fold, which consists of five or six β-strands arranged in two tightly packed antiparallel β-sheets. The linker region may contain short helices. The surface of the SH3 domain bears a flat hydrophobic ligand-binding pocket consisting of three shallow grooves defined by conserved aromatic residues in which the ligands are arranged in an extended left-handed helix. Ligands bind with low affinity, but this can be enhanced by multiple interactions. The region bound by the SH3 domain is proline-rich in all cases and contains PXXP as a core conserved binding motif. The function of SH3 domains is unclear, but they may mediate many different processes, such as increasing the local concentration of proteins, changing their subcellular location and mediating the assembly of large multiprotein complexes.

Through enrichment analysis, we found that the gene sets of the identified clusters were enriched in a variety of enzyme activities, ADP and ATP related generation reactions, replication and translation of genetic material DNA and RNA, neurotransmitter transmission links and other pathways. Multiple clusters were enriched in RNA polymerase II forward and transcriptional regulatory pathways, protein tyrosine related enzyme pathways, neural synapses, neurotransmitter transmission links, ATP, ADP synthesis related links. There were two clusters of gene sets enriched to human papillomavirus infection pathway. One cluster was significantly enriched in calcium ion related pathways. Another cluster was significantly enriched in the inositol phosphate metabolism pathway. SH3 (src Homology-3) domains, proteoglycan cancer pathway, PDZ domain, Hippo signaling pathway, Tight junction pathway, PB1 domain and other pathways were also enriched in some clusters. Each cluster enriched in the above described pathways at the same time there are other enrichment pathways with different functions. There may be multiple gene interactions enriched in different pathways leading to differences in sperm motility.

In order to determine whether the data is significantly enriched, the *p*-values of the enrichment results are corrected using the Benjamini method and the Bonferroni method. The specific identified differentially expressed genes and the number of samples is shown in Table 1. Specific gene and sample data are included in the Supplementary Table: The result of identification. Table 2 is the cluster-related enrichment results, Figure 3 visualizes the correlation enrichment results, and the enrichment analysis results of all clusters are shown in Supplementary Data.

3.2. Simulation data experimental results

Since this paper belongs to unsupervised learning, there is no standard answer for the quantitative study of male sperm motility. At the same time, in order to better determine the value of hyper-parameters in the statistical method used in this paper, simulated data similar to gene expression profile datasets are constructed to be used in the method proposed in this paper. The clustering results in the simulated data have been determined and can be used to evaluate the model performance. Comparing the identification results of the simulated data set with the results of similar methods, and the results show that the model proposed in this paper may have higher accuracy in the analysis of genetic factors in the quantitative study of male sperm (Table 3).

To identify the differential expression module of the simulated data, we used the C&C (Cheng and Church, 2000) and BCPlaid (Lazzeroni and Owen, 2000) methods to cluster the data, and calculated the jaccard similarity coefficient of the results, which was often used to compare the similarity and difference between the limited sample sets, among which the jaccard coefficient. The higher the value, the higher the similarity between sets. The stable parameters were tuned best in each model. The specific results are shown in Supplementary Table 3, and the corresponding box plot is in Figure 4.

TABLE 2 Enrichment results of genes in a cluster identified by our method in the male infertility data set.

Category	Term	Genes	Bonferroni	Benjamini
GOTERM_CC_DIRECT	GO:0030424 ~ axon	CNTNAP2, CNTN5, IL1RAPL1, DMD, SCN1A	0.002330526	0.002333212
GOTERM_CC_DIRECT	GO:0009986 ~ cell surface	LGALS3, CNTNAP2, NLGN4X, IL1RAPL1, DMD	0.021009445	0.010615268
UP_KW_DISEASE	KW-1269 ~ Autism	CNTNAP2, NLGN4X, SCN1A	0.002854718	0.002858289
UP_KW_DISEASE	KW-1268 ~ Autism spectrum disorder	CNTNAP2, NLGN4X, SCN1A	0.014578999	0.007336422

Only the pathways and related parameters that were modified and significantly enriched by Bonferroni and Benjamini are listed in the table. The cluster is the id in Table 1: 1.

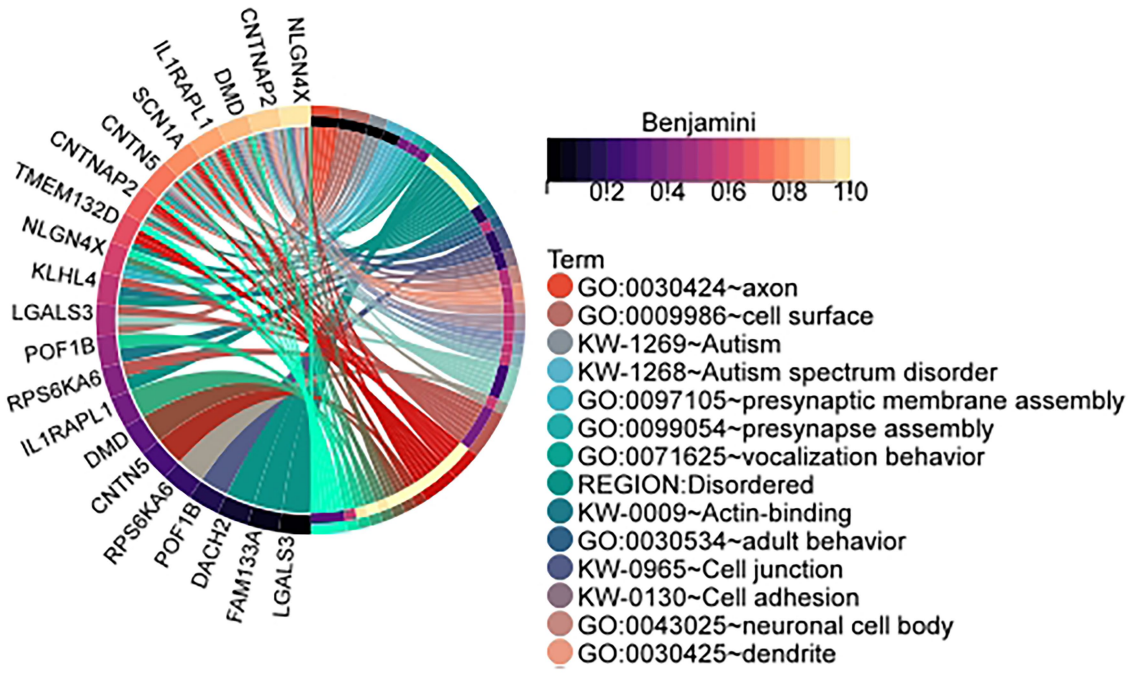


FIGURE 3
Enrichment circle plot of genes in clusters identified by our method in the male infertility data set. The cluster is the id in Table 1: 1. (Visualization of the relationship between genes and enrichment pathways).

4. Conclusion

Based on the analysis of the GSE37948 male infertility-related gene detection data set in the GEO database, this paper proposes a bicluster analysis method based on hypergeometric distribution, Gibbs sampling and machine learning, and establishes simulation data similar to the GSE37948 data set. The common bicluster analysis methods C&C (Cheng and Church, 2000) and BCPlaid (Lazzeroni and Owen, 2000) have compared the experimental results. The results show that the method proposed in this paper has a higher accuracy in the identification of biclusters on the established simulation data set.

Through enrichment analysis, we found that the gene sets of the identified clusters were enriched in a variety of enzyme activities, ADP and ATP related generation reactions, replication and translation of genetic material DNA and RNA, neurotransmitter transmission links and other pathways. Multiple clusters were enriched in RNA polymerase II forward and transcriptional regulatory pathways, protein tyrosine related enzyme pathways, neural synapses, neurotransmitter transmission links, ATP, ADP synthesis related links. There were two clusters of gene sets enriched to human papillomavirus infection pathway. One cluster was significantly enriched in the inositol phosphate metabolism pathway. Each cluster enriched in the above described pathways at the same time there are other enrichment pathways with different functions. There may be multiple gene interactions enriched in different pathways leading to differences in sperm motility.

Infertility is a complex pathological condition that presents with a wide range of heterogeneous prototypes, and identifying the genes that cause male infertility is important to increase our biological understanding and clinically relevant treatments. The genetic causes of male infertility are chromosomal abnormalities, gene mutations and other reasons, which may be present in autosomes or in sex

TABLE 3 The jaccard similarity coefficient between the clustering results identified by the three methods on different simulated datasets and the real clusters, where simulation data represents (the number of samples, the number of genes).

Simulation data	BCPlaid	C&C	COEXSML (this work)
(10, 5)	0.0000	0.0000	0.0346
(10, 10)	0.0000	0.0002	0.1089
(10, 20)	0.0000	0.0005	0.0552
(10, 50)	0.0003	0.0012	0.1150
(10, 100)	0.0002	0.0022	0.2023
(20, 5)	0.0000	0.0001	0.4509
(20, 10)	0.0000	0.0005	0.6126
(20, 20)	0.0000	0.0009	0.5373
(20, 50)	0.0004	0.0023	0.3382
(20, 100)	0.0012	0.0033	0.3195
(50, 5)	0.0000	0.0003	0.5112
(50, 10)	0.0000	0.0013	0.7917
(50, 20)	0.0020	0.0033	0.8291
(50, 50)	0.0000	0.0047	0.8715
(50, 100)	0.0024	0.0061	0.8097
(100,5)	0.0000	0.0004	0.5123
(100, 10)	0.0000	0.0042	0.6277
(100, 20)	0.0000	0.0038	0.6794
(100, 50)	0.0000	0.0074	0.6938
(100, 100)	0.0007	0.0214	0.5455

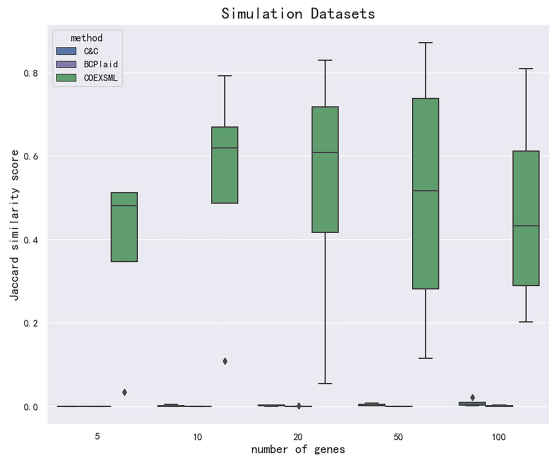


FIGURE 4
The data are divided according to the difference in the number of genes in the clusters in the simulated data set. The clustering effect is measured according to the jaccard similarity coefficient, and compared with other methods. COEXSML is the method proposed in this paper.

chromosomes, considering the particularity of male infertility, this article only considers the study of related genes on the X chromosome. With the development of genetic testing technology, the relevant data has increased significantly, and follow-up research can fully explore the information contained in the gene expression data of relevant patients from more aspects.

Data availability statement

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

Author contributions

XJ proposed the model and completed the manuscript writing. YP assisted in completing the model construction. YP and ZY reviewed and revised the manuscript. ZY provided financial support. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by the National Natural Science Foundation of China (No: 62072296).

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1092143/full#supplementary-material

References

Alzahrani, M., Kuwahara, H., Wang, W., and Gao, X. (2017). Gracob: a novel graph-based constant-column biclustering method for mining growth phenotype data. *Bioinformatics* 33, 2523–2531. doi: 10.1093/bioinformatics/btx199

Antonuccio, P., Micali, A. G., Romeo, C., Freni, J., Vermiglio, G., Puzzolo, D., et al. (2021). NLRP3 inflammasome: a new pharmacological target for reducing testicular damage associated with varicocele. *Int. J. Mol. Sci.* 22. doi: 10.3390/ijms22031319

Aquila, S., Sisci, D., Gentile, M., Middea, E., Catalano, S., Carpino, A., et al. (2004). Estrogen receptor (ER) alpha and ER beta are both expressed in human ejaculated spermatozoa: evidence of their direct interaction with phosphatidylinositol-3-OH kinase/Akt pathway. *J. Clin. Endocrinol. Metab.* 89, 1443–1451. doi: 10.1210/jc.2003-031681

Bergmann, S., Ihmels, J., and Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* 67:031902. doi: 10.1103/PhysRevE.67.031902

Bollobás, B., Borgs, C., and Chayes, J. (2003). "Directed scale-free graphs," in *Proceedings of the fourteenth annual ACM-SIAM symposium on discrete algorithms*. (Philadelphia, PA, USA). 132–139.

Chen, Y., Mao, F., Li, G., and Xu, Y. J. B. B. (2011). Genome-wide discovery of missing genes in biological pathways of prokaryotes. *BMC Bioinformatics* 12:S1. doi: 10.1186/1471-2105-12-S1-S1

Cheng, Y., and Church, G. M. (2000). "Biclustering of expression data," in *Proceedings of the eighth international conference on intelligent systems for molecular biology*. (AAAI Press). 93–103.

Chillón, M., Casals, T., Mercier, B., Bassas, L., Lissens, W., Silber, S., et al. (1995). Mutations in the cystic fibrosis gene in patients with congenital absence of the vas deferens. *N. Engl. J. Med.* 332, 1475–1480. doi: 10.1056/NEJM199506013322204

Chiu, H.S., Chuang, H.Y., Tsai, H.K., Huang, T.W., and Kao, C.Y. (2004). Discovering statistically significant clusters by using iterative genetic algorithms in gene expression data. In Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Scienes, METMBS, Las Vegas, Nevada, USA.

Dada, R., Gupta, N. P., and Kucheria, K. (2003). Molecular screening for Yq microdeletion in men with idiopathic oligozoospermia and azoospermia. *Proc. Anim. Sci.* 28, 163–168. doi: 10.1007/BF02706215

De Smet, R., and Marchal, K. (2011). An ensemble biclustering approach for querying gene expression compendia with experimental lists. *Bioinformatics* 27, 1948–1956. doi: 10.1093/bioinformatics/btr307

Dhollander, T., Sheng, Q., Lemmens, K., De Moor, B., Marchal, K., and Moreau, Y. (2007). Query-driven module discovery in microarray data. *Bioinformatics* 23, 2573–2580. doi: 10.1093/bioinformatics/btm387

Eren, K., Deveci, M., Kucuktunc, O., and Catalyurek, U. V. (2013). A comparative analysis of biclustering algorithms for gene expression data. *Brief. Bioinform.* 14, 279–292. doi: 10.1093/bib/bbs032

Gerniers, A., Bricard, O., and Dupont, P. (2021). MicroCellClust: mining rare and highly specific subpopulations from single-cell expression data. *Bioinformatics* 37, 3220–3227. doi: 10.1093/bioinformatics/btab239

Ghiassian, S. D., Menche, J., and Barabasi, A. L. (2015). A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* 11:e1004120. doi: 10.1371/journal.pcbi.1004120

Goncalves, J. P., and Madeira, S. C. (2014). LateBiclustering: efficient heuristic algorithm for time-lagged bicluster identification. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11, 801–813. doi: 10.1109/TCBB.2014.2312007

Gu, J., and Liu, J. S. (2008). Bayesian biclustering of gene expression data. *BMC Genomics* 9:S4. doi: 10.1186/1471-2164-9-S1-S4

Guo, F., Yin, Z., Zhou, K., and Li, J. (2021). PLncWX: a machine-learning algorithm for plant lncRNA identification based on WOA-XGBoost. *J. Chem.* 2021, 1–11. doi: 10.1155/2021/6256021

Hartigan, J. A. (1972). Direct clustering of a data matrix. *J. Am. Stat. Assoc.* 67, 123–129. doi: 10.1080/01621459.1972.10481214

Henriques, R., and Madeira, S. C. (2016a). BiC2PAM: constraint-guided biclustering for biological data analysis with domain knowledge. *Algorithms Mol. Biol.* 11:23. doi: 10.1186/s13015-016-0085-5

Henriques, R., and Madeira, S. C. (2016b). BicNET: flexible module discovery in large-scale biological networks using biclustering. *Algorithms Mol. Biol.* 11:14. doi: 10.1186/s13015-016-0074-8

Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., et al. (2010). FABIA: factor analysis for bicluster acquisition. *Bioinformatics* 26, 1520–1527. doi: 10.1093/bioinformatics/btq227

Jacobs, P. A., and Strong, J. A. (1959). A case of human intersexuality having a possible XXY sex-determining mechanism. *Nature* 183, 302–303. doi: 10.1038/183302a0

Jiang, J., Pan, W., Xu, Y., Ni, C., Xue, D., Chen, Z., et al. (2020). Tumour-infiltrating immune cell-based subtyping and signature gene analysis in breast cancer based on gene expression profiles. *J. Cancer* 11, 1568–1583. doi: 10.7150/jca.37637

Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M. (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *PCR Methods Appl.* 13, 703–716. doi: 10.1101/gr.648603

Krausz, C., Giachini, C., Lo Giacco, D., Daguin, F., Chianese, C., Ars, E., et al. (2012). High resolution X chromosome-specific array-CGH detects new CNVs in infertile males. *PLoS One* 7:e44887. doi: 10.1371/journal.pone.0044887

Krausz, C., and Riera-Escamilla, A. J. (2018). Genetics of male infertility. *Nat. Clin. Pract. Urol.* 15, 369–384. doi: 10.1038/s41585-018-0003-3

Lazareva, O., Canzar, S., Yuan, K., Baumbach, J., Blumenthal, D. B., Tieri, P., et al. (2020). BiCoN: network-constrained biclustering of patients and omics data. *Bioinformatics* 37, 2398–2404. doi: 10.1093/bioinformatics/btaa1076

Lazzeroni, L., and Owen, A. J. (2000). Plaid models for gene expression data. *Stat. Sin.* 12, 61–86.

Li, G., Ma, Q., Tang, H., Paterson, A. H., and Xu, Y. (2009). QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res.* 37:e101. doi: 10.1093/nar/gkp491

Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., et al. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* 26, 842–844. doi: 10.1038/s41591-020-0901-9

Lopes, A. M., Aston, K. I., Thompson, E. E., Carvalho, F., Gonçalves, J., Huang, N., et al. (2013). Human spermatogenic failure purges deleterious mutation load from the autosomes and both sex chromosomes, including the gene DMRT1. *Public Library Sci. Genet.* 9:e1003349. doi: 10.1371/journal.pgen.1003349

Lyon, M. F., Hawkes, S. G., and Nature, H. J. (1970). X-linked gene for testicular feminization in the mouse. *Nature* 227, 1217–1219. doi: 10.1038/2271217a0

Madeira, S. C., Teixeira, M. C., Sa-Correia, I., and Oliveira, A. L. (2010). Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 7, 153–165. doi: 10.1109/TCBB.2008.34

Medina, I., Carbonell, J., Pulido, L., Madeira, S. C., Goetz, S., Conesa, A., et al. (2010). Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res.* 38, W210–W213. doi: 10.1093/nar/gkq388

Peng, L., Tian, X., Tian, G., Xu, J., Huang, X., Weng, Y., et al. (2020). Single-cell RNA-seq clustering: datasets, models, and algorithms. *RNA Biol.* 17, 765–783. doi: 10.1080/15476286.2020.1728961

Peng, L., Wang, F., Wang, Z., Tan, J., Huang, L., Tian, X., et al. (2022). Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies. *Brief. Bioinform.* 23:bbac234. doi: 10.1093/bib/bbac234

Peng, L., Yuan, R., Shen, L., Gao, P., and Zhou, L. J. (2021). LPI-EnEDT: an ensemble framework with extra tree and decision tree classifiers for imbalanced lncRNA-protein interaction data classification. *BioData Min* 14, 50–22. doi: 10.1186/s13040-021-00277-4

Plaisier, S. B., Taschereau, R., Wong, J. A., and Graeber, T. G. (2010). Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Res.* 38:e169. doi: 10.1093/nar/gkq636

Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., et al. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* 22, 1122–1129. doi: 10.1093/bioinformatics/btl060

Ron, E., Michael, D., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 1, 207–210. doi: 10.1093/nar/30.1.207

Sheng, Q., Moreau, Y., and De Moor, B. (2003). Biclustering microarray data by Gibbs sampling. *Bioinformatics* 19:ii196-205. doi: 10.1093/bioinformatics/btg1078

Tournaye, H., Krausz, C., and Oates, R. D. (2017). Novel concepts in the aetiology of male reproductive impairment. *Lancet Diabetes Endocrinol.* 5, 544–553. doi: 10.1016/S2213-8587(16)30040-7

Wang, H., Zhao, R., Guo, C., Jiang, S., Yang, J., Xu, Y., et al. (2016). Knockout of BRD7 results in impaired spermatogenesis and male infertility. *Sci. Rep.* 6:21776. doi: 10.1038/srep21776

Xavier, M. J., Salas-Huetos, A., Oud, M. S., Aston, K. I., and Veltman, J. A. (2021). Disease gene discovery in male infertility: past, present and future. *Hum. Genet.* 140, 7–19. doi: 10.1007/s00439-020-02202-x

Yatsenko, A. N., Georgiadis, A. P., Röpke, A., Berman, A. J., Jaffe, T., Olszewska, M., et al. (2015). X-linked TEX11 mutations, meiotic arrest, and azoospermia in infertile men. *N. Engl. J. Med.* 372, 2097–2107. doi: 10.1056/NEJMoa1406192

Yi, H., Huang, L., Mishne, G., and Chi, E. C. (2021). COBRAC: a fast implementation of convex biclustering with compression. *Bioinformatics* 37, 3667–3669. doi: 10.1093/bioinformatics/btab248

Zhai, Z., Lei, Y. L., Wang, R., and Xie, Y. (2022). Supervised capacity preserving mapping: a clustering guided visualization method for scRNA-seq data. *Bioinformatics* 38, 2496–2503. doi: 10.1093/bioinformatics/btac131

Zhao, H., Cloots, L., Bulcke, T. V. D., Wu, Y., and Marchal, K. J. B. B. (2011). Query-based biclustering of gene expression data using probabilistic relational models. *Bioinformatics* 12:S37. doi: 10.1186/1471-2105-12-S1-S37

Zhou, F., Ma, Q., Li, G., and Xu, Y. (2012). QServer: a biclustering server for prediction and assessment of co-expressed gene clusters. *PLoS One* 7:e32660. doi: 10.1371/journal.pone.0032660

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 03 February 2023
DOI 10.3389/fmicb.2023.1116592

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Xiao Wang,
Qingdao University, China
Liu Fuxiang,
China Three Gorges University, China

*CORRESPONDENCE
Hongping Guo
guohongping@hbnu.edu.cn

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

RECEIVED 05 December 2022
ACCEPTED 06 January 2023
PUBLISHED 03 February 2023

CITATION
Guo H, Cao W, Zhu Y, Li T and Hu B (2023) A genome-wide cross-cancer meta-analysis highlights the shared genetic links of five solid cancers. *Front. Microbiol.* 14:1116592. doi: 10.3389/fmicb.2023.1116592

A genome-wide cross-cancer meta-analysis highlights the shared genetic links of five solid cancers

Hongping Guo[1]*, Wenhao Cao[2], Yiran Zhu[1], Tong Li[1] and Boheng Hu[1]

[1]School of Mathematics and Statistics, Hubei Normal University, Huangshi, China, [2]Division of Biostatistics, University of Minnesota, Minneapolis, MN, United States

Breast, ovarian, prostate, lung, and head/neck cancers are five solid cancers with complex interrelationships. However, the shared genetic factors of the five cancers were often revealed either by the combination of individual genome-wide association study (GWAS) approach or by the fixed-effect model-based meta-analysis approach with practically impossible assumptions. Here, we presented a random-effect model-based cross-cancer meta-analysis framework for identifying the genetic variants jointly influencing the five solid cancers. A comprehensive genetic correlation analysis (genome-wide, partitioned, and local) approach was performed by using GWAS summary statistics of the five cancers, and we observed three cancer pairs with significant genetic correlation: breast–ovarian cancer ($r_g = 0.221$, $p = 0.0003$), breast–lung cancer ($r_g = 0.234$, $p = 7.6 \times 10^{-6}$), and lung–head/neck cancer ($r_g = 0.652$, $p = 0.010$). Furthermore, a random-effect model-based cross-trait meta-analysis was conducted for each significant cancer pair, and we found 27 shared genetic loci between breast and ovarian cancers, 18 loci between breast and lung cancers, and three loci between lung and head/neck cancers. Functional analysis indicates that the shared genes are enriched in *human T-cell leukemia virus 1 infection (HTLV-1)* and *antigen processing and presentation (APP)* pathways. Our study investigates the shared genetic links across five solid cancers and will help to reveal their potential molecular mechanisms.

KEYWORDS

solid cancers, summary statistics, shared genetic loci, meta-analysis, random effect model

1. Introduction

Cancer has become one of the most fatal diseases and it poses a serious threat to human life and health. There have been ~18.1 million new cancer cases and 9.6 million cancer deaths each year (Bray et al., 2018). According to the prediction of the National Cancer Institute, the number of new cancer cases per year is expected to rise to 29.5 million, and the amount of cancer-related deaths will go up to 16.4 million by 2040. The high incidence of cancer has not only brought an enormous health burden to individuals but also caused heavy economic losses to countless families. Numerous pieces of evidence indicated widespread genetic pleiotropy and shared genetic basis among different cancers (Rashkin et al., 2020). As a few representative elements of solid cancer, breast, ovarian, prostate, lung, and head/neck cancers showed substantial heritability (ranging from 9 to 57%) in previous twin and family studies (Polderman et al., 2015; Mucci et al., 2016; Yu et al., 2017). Moreover, Jiang et al. (2019) quantified the pairwise genetic correlations of six solid cancers and found significant correlations between breast and ovarian cancers, breast and lung cancers, breast and colorectal cancers, and lung and head/neck cancers. The aforementioned conclusions demonstrate indirectly that these solid cancers may share inherited genetic mechanisms, which play important roles in cancer etiology. We would like to understand the shared genetic loci influencing the five solid cancers.

Genome-wide association studies (GWASs) have identified a number of susceptibility loci associated with each of the five solid cancers, ranging from dozens to hundreds (Buniello et al., 2019), but few of them overlap in at least two of these cancers. This indicates that rare pleiotropic loci are detected by cancer-specific GWAS. Identifying the shared genetic loci between diseases can help to reveal the underlying mechanisms driving disease etiology (Guo et al., 2020). There are mainly two strategies available to identify the shared loci in the previous literature. One strategy is based on the combination of GWASs and other scan analyses. For example, Ghoussaini et al. found pleiotropic loci located at 8q24, associated with breast, prostate, and other specific cancers by using this approach (Ghoussaini et al., 2008). Another strategy is based on a cross-cancer meta-analysis. For example, Kar et al. identified seven new loci shared by at least two of the three hormone-related cancers (breast, ovarian, and prostate); Fehringer et al. (2016) detected a novel pleiotropic locus 1q22 associated with both breast and lung cancers by performing a cross-cancer genome-wide analysis of breast, ovary, prostate, lung, and colorectal cancers. However, the pleiotropic loci identified by the above studies are still not sufficient, and this may due to the fact that the cross-cancer meta-analyses in the existing studies are based on the fix-effect model. The fix-effect model meta-analysis causes the loss of statistical power because it assumes the same real effect for each genetic variant in different studies, which is practically impossible and will inevitably yield inaccurate conclusions.

Random-effect model-based cross-trait meta-analysis methods can effectively account for the heterogeneous effect of each genetic variant by adding an additional variance term, addressing the shortcomings of fix-effect model-based meta-analysis. Here, we use the summary statistics of five solid cancers (breast, ovarian, prostate, lung, and head/neck) from the largest-to-date cancer-specific GWAS consortia, which include a total of 241,479 cases and 226,810 controls. We then estimate the genetic correlation between different cancer pairs. Furthermore, we conducted a cross-cancer meta-analysis to detect shared genetic loci between the cancer pairs using the current state-of-the-art random-effect model-based approach PLEIO (Pleiotropic Locus Exploration and Interpretation using Optimal test) (Lee et al., 2021), which enables us to properly account for the correlation of traits and the heterogeneity of variants. Finally, we perform functional analyses of pleiotropic variants to uncover the underlying biological mechanisms shared across the five solid cancers.

2. Materials and methods

2.1. Data and contributing consortia

We used the most recent GWAS summary-level data from the Breast Cancer Association Consortium (BCAC) for breast cancer (122,977 cases and 105,974 controls) (Michailidou et al., 2017), the Ovarian Cancer Association Consortium (OCAC) for ovarian cancer (25,509 cases and 40,941 controls) (Phelan et al., 2017), the Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) consortium for prostate cancer (79,148 cases and 61,106 controls) (Schumacher et al., 2018), the International Lung Cancer Consortium (ILCCO) for lung cancer (11,348 cases and 15,861 controls) (Wang et al., 2014), and the Oncoarray oral cavity and oropharyngeal cancer consortium for head/neck cancer (2,497 cases and 2,928 controls) (Lesseur et al., 2016).

2.2. Genome-wide genetic correlations

To measure genome-wide genetic correlations for each cancer pair, we used the linkage disequilibrium (LD) score regression (LDSC) method (Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2015). We applied pre-computed LD scores derived from ∼1.2 million imputed variants from European populations that did not include the HLA region in the HapMap3 reference panel. LDSC controls for population structure using GWAS summary statistics without individual-level data.

2.3. Partitioned genetic correlations

We evaluated the partitioned genetic correlation across the five solid cancers within functional categories by using partitioned LDSC (ReproGen Consortium et al., 2015). We chose 11 functional categories as previously recommended (Zhu et al., 2019), including the DNase I digital genomic footprinting (DGF) region, DNase I hypersensitivity sites (DHSs), fetal DHS, intron, super-enhancer, transcription factor-binding sites (TFBS), transcribed region, and the histone markers H3K9ac, H3K4me1, H3K4me3, and H3K27ac from the Roadmap Epigenomics Project (Bernstein et al., 2010). Re-computed LD scores for variants classified in each particular annotation were used for estimating the cross-cancer genetic correlation within that functional group.

2.4. Local genetic correlations

We estimated local genetic correlations between each pair of cancers in 1,703 pre-specified LD-independent regions using ρ-HESS (Shi et al., 2017). The goal of this method was to detect small contiguous regions of the genome in which the genetic associations of two traits are locally concordant, and to measure the local genetic correlation and p-values ($p_{\rho-HESS}$) between pairs of traits at local regions. Cancer pairs were considered to have genetic correlation at the local region if $p_{\rho-HESS}$ passed the multiple testing correction ($p_{\rho-HESS} < 0.05/1703$).

2.5. Cross-cancer meta-analysis

For the cancer pairs with significant genome-wide genetic correlation, we conducted a pairwise cross-cancer meta-analysis by using PLEIO (Lee et al., 2021). The approach is based on a random-effect model, which can not only model genetic correlations across pairs of traits but can also correct for environmental correlations. It can seamlessly test multiple traits with various types by standardizing the effect sizes. Moreover, it maps pleiotropic loci through a variance component test and calculates statistical significance through an important sampling method. It overcomes the drawback of fixed-effect model methods such as ASSET (association analysis based on subsets) (Bhattacharjee et al., 2012). We conducted the cross-cancer

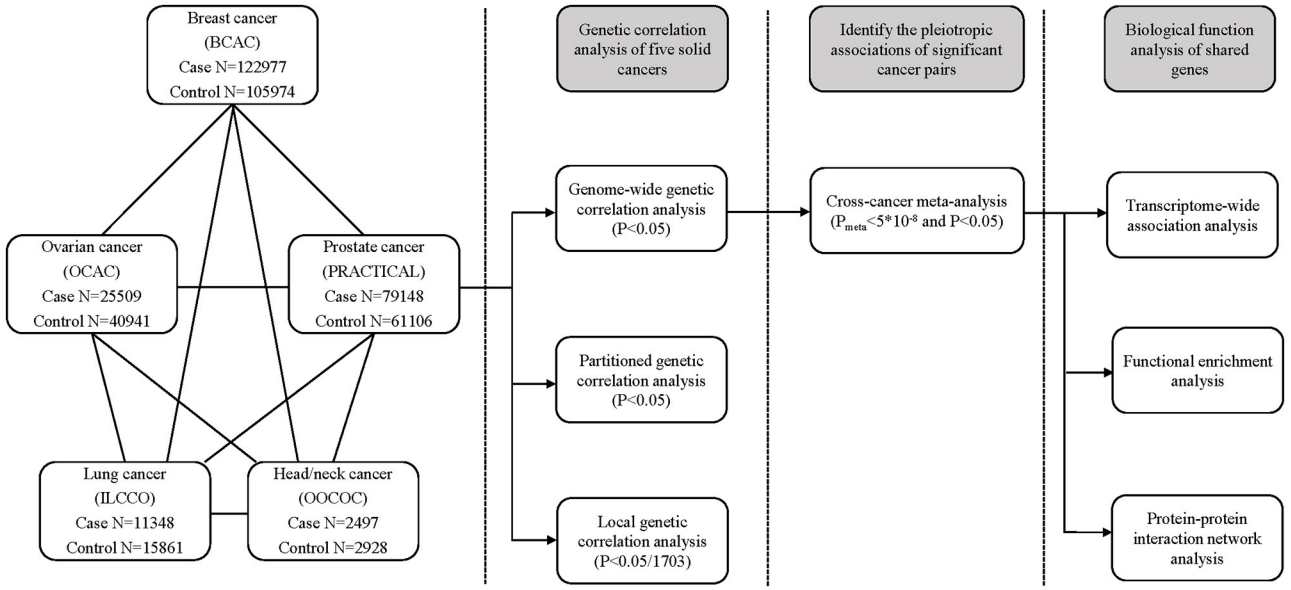


FIGURE 1
Schematic overview of the present study.

meta-analysis on an Intel Xeon E5-2695 computer with the CPU operating at 2.10 GHz. This wastes ∼10 min for each pair of cancers.

To separate the independent loci from the significant loci ($p < 5 \times 10^{-8}$), we used the clumping function in PLINK software (Purcell et al., 2007). SNPs with $p < 1 \times 10^{-5}$, an LD statistic $r^2 > 0.05$, and a distance from the peak < 1,000 kb were assigned to the clump of that peak. Moreover, we set the NCBI human genome build 37 as the reference gene list.

2.6. Transcriptome-wide association studies

We performed TWAS to identify gene–tissue pairs for each of the five solid cancers and used FUSION software based on the pre-computed 48 GTEx (version 7) tissue expression reference weights (Gusev et al., 2016). LD-reference data were derived from European descendants from the 1,000 Genomes Project. For each cancer, we conducted 48 TWASs, one tissue-cancer pair at a time. The false discovery rate (FDR) Benjamin–Hochberg procedure correction was used, and a result with an FDR < 0.05 was considered to be significant.

2.7. Replication analysis in the UK Biobank cohort

To validate our findings, we further conducted genome-wide genetic correlation analysis and cross-cancer meta-analysis of the five solid cancer GWAS datasets with the UK Biobank cohort from the IEU GWAS database project (Matthew et al., 2021): breast cancer (ID: ieu-b-4810), ovarian cancer (ID: ieu-b-4963), prostate cancer (ID: ieu-b-4809), lung cancer (ID: ieu-b-4954), and head/neck cancer (ID: ieu-b-4912). We applied the 1,000 Genomes Project variants (Phase 3) as the reference panel. The cross-cancer meta-analysis between each pair of replication datasets was implemented using the R software RE2C (Lee et al., 2017), which is another classical random-effect model-based method that tests heterogeneous effect size between individual summary statistics.

2.8. Pathway enrichment analysis

To gain biology insights from the shared risk genes, we performed Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis using the Enrichr web server (Kuleshov et al., 2016), which is a comprehensive resource for curated gene sets and a search engine that accumulates biological knowledge for further biological discoveries. The significant criterion is that the adjusted *p*-value is <0.05.

2.9. Protein–protein interaction network analysis

We used STRING v10 (Szklarczyk et al., 2015) to analyze the PPI network. The basic assumption is that if two proteins are functionally associated, they may contribute to a common biological purpose. The interaction scores were derived from different sources, including experimentally determined interaction, database annotated information, and automated text mining knowledge.

A schematic overview of the present study is shown in Figure 1, that is, we estimated genome-wide, partitioned, and local genetic correlations of the five solid cancers. For the cancer pairs with

significant genome-wide genetic correlation, we performed a cross-cancer meta-analysis to identify shared genetic loci. Finally, we conducted TWAS, pathway enrichment analysis, and PPI network analysis of the shared risk genes.

3. Results

3.1. Three cancer pairs have significant genetic correlations

Among pairs of solid cancers, we found three pairs with positive genetic correlations at a significant threshold of $p = 0.05$: breast and ovarian cancers ($r_g = 0.221$, $p = 0.0003$), breast and lung cancers ($r_g = 0.234$, $p = 7.6 \times 10^{-6}$), and lung and head/neck cancers ($r_g = 0.652$, $p = 0.010$). The remaining pairs do not show significant genetic correlations (Table 1).

TABLE 1 Genome-wide genetic correlation between five solid cancers.

Cancer type[a]	Breast	Ovarian	Prostate	Lung	Head/neck
Breast	1	0.221	0.077	0.234	−0.065
Ovarian	0.0003	1	0.026	0.139	−0.072
Prostate	0.087	0.672	1	0.069	0.160
Lung	7.6×10^{-6}	0.164	0.272	1	0.652
Head/neck	0.528	0.761	0.070	0.010	1

[a]The upper off-diagonal shows the genetic correlation estimates of the LD score regression (r_g ranges from −1 to 1), and the lower off-diagonal shows the corresponding p-values.

3.2. Most of the three cancer pairs have significant functional partitioned genetic correlations

In the partitioned genetic correlation analysis, we observed significant genetic correlation in all 11 functional categories for the breast–lung cancer pair, with only two exceptions: Intron and SuperEnhance for the lung–head/neck cancer pair. As to the breast–ovarian cancer pair, there is no significant signal in H3K27ac, H3K4me3, H3K9ac, and SuperEnhance. The partitioned genetic correlations range from 0.033 to 0.546 (Figure 2; Supplementary Table S1).

3.3. Two cancer pairs have four genomic regions with significant local genetic correlations

We conducted ρ-HESS to investigate whether specific regions had a genetic correlation between each pair of the five solid cancers. The results show that the breast–ovarian cancer pair has a strong local genetic correlation in the 2q33 region (chromosome 2: 201576284-202818637, $p = 8.83 \times 10^{-6}$) (Figure 3A). In addition, three regions, including the 9p21 region (chromosome 9: 20463534-22206559, $p = 6.71 \times 10^{-6}$), 10q26 region (chromosome 10: 123231465-123900545, $p = 4.26 \times 10^{-7}$), and 11q13 region (chromosome 11: 68005825-69516130, $p = 4.90 \times 10^{-6}$), are found to have strong local genetic correlations in the breast–prostate cancer pair (Figure 3B). We did not observe any significant local genetic correlations for the other cancer pairs.

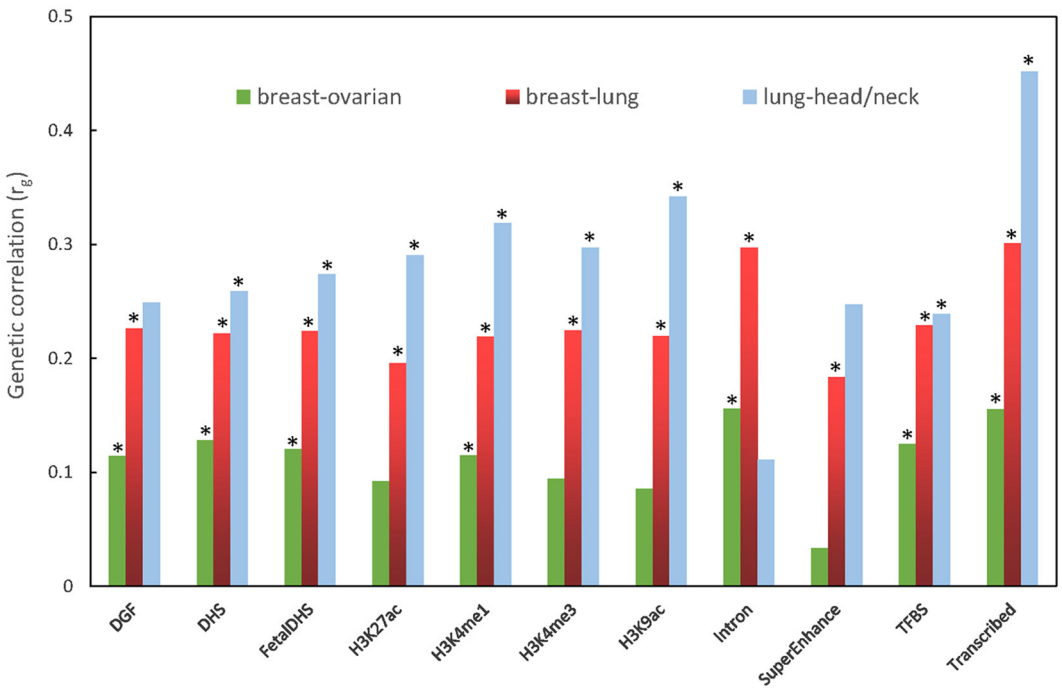


FIGURE 2
Partitioned genetic correlation between breast and ovarian cancers, breast and lung cancers, and lung and head/neck cancers. The vertical axis represents the genetic correlation r_g, and the horizontal axis represents 11 functional categories. The asterisk represents significance ($p < 0.05$).

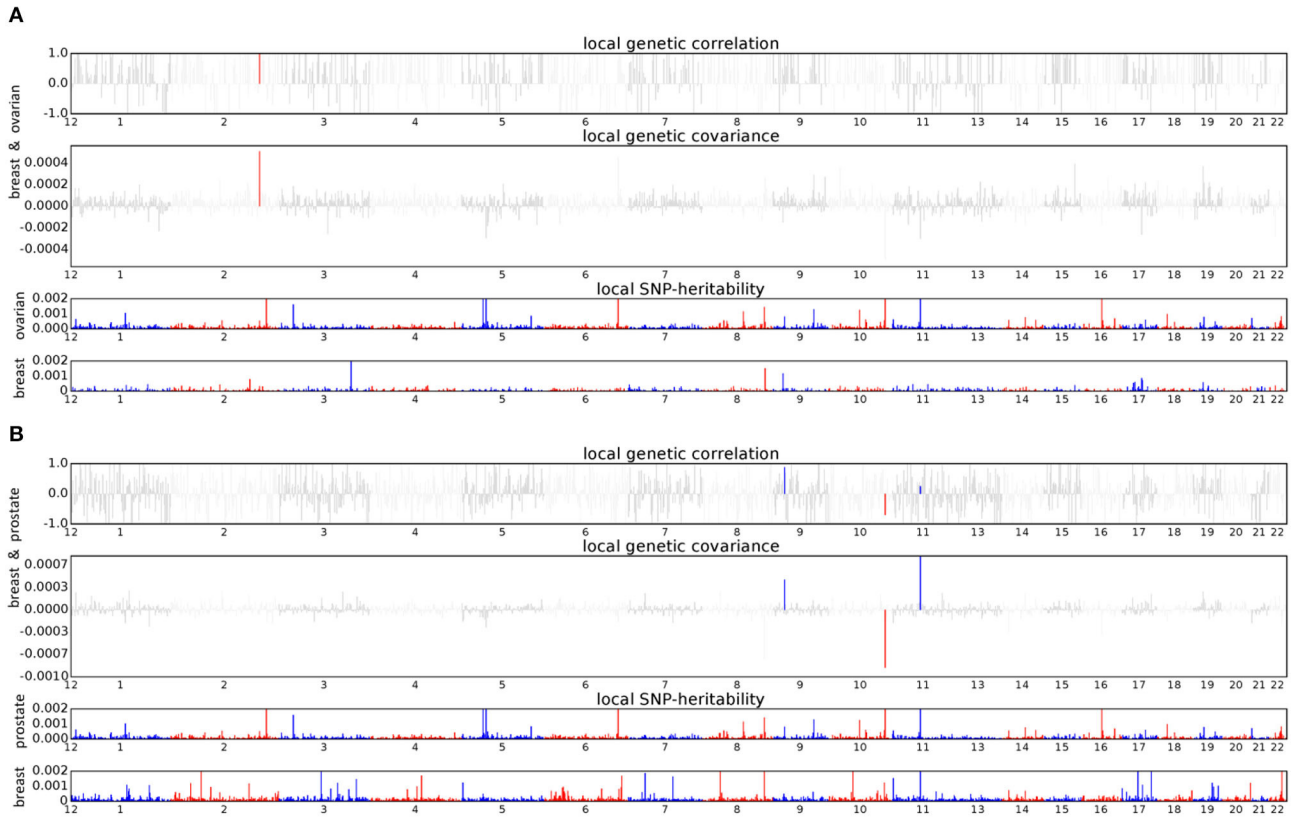


FIGURE 3
Local genetic correlation and local SNP heritability between cancer pair. **(A)** Breast and ovarian cancers; **(B)** Breast and prostate cancers. For each subfigure, the top part represents local genetic correlation, the middle part represents local genetic covariance, blue or red highlights indicate significant local genetic correlation and covariance after multiple testing corrections, and the bottom part represents local SNP heritability for each trait.

3.4. Pleiotropic loci were identified for the three cancer pairs by cross-cancer meta-analysis

3.4.1. Breast and ovarian cancer

In the cross-cancer meta-analysis, we identified 27 independent loci with a significant association between breast and ovarian cancers ($p_{meta} < 5 \times 10^{-8}$ and single-trait $p < 0.05$, Table 2). The strongest pleiotropic signal is mapped to *FGFR2* in the region 10q26.13 (rs1219648, $p_{meta} = 4.16 \times 10^{-254}$), a gene that has been altered in a number of patients with malignant solid tumors according to the AACR Project GENIE (The AACR Project GENIE Consortium et al., 2017). This SNP showed a pleiotropic association between breast and ovarian cancers according to a previous cross-cancer analysis (Kar et al., 2016). The second strongest signal is observed for chromosome 9q31.2 (rs630965, $p_{meta} = 1.01 \times 10^{-63}$). Patients with deletions on 9q31.2 may have delayed puberty (Iivonen et al., 2021). The third strongest signal observed on *BNC2* (rs3814113, $p_{meta} = 2.16 \times 10^{-43}$) is a putative tumor suppressor gene in high-grade serous ovarian carcinoma, which impacted cell survival after oxidative stress (Cesaratto et al., 2016). Notably, four loci (rs7098100, rs4277389, rs4808616, and rs10069690) are not only significant after the meta-analysis but also reach a significant level in their original single-trait GWAS.

3.4.2. Breast and lung cancers

For the breast–lung cancer pair, we detected 18 pleiotropic loci in the cross-cancer meta-analysis (Table 3). The most significant pleiotropic association is in the region 5q11.2 (rs16886181, $p_{meta} = 4.57 \times 10^{-122}$), and the mapped gene *MAP3K1* regulates apoptosis, survival, migration, differentiation, and other functions, which suggests that it may be a target for cancer treatment (Pham et al., 2013). Moreover, we also found dense signals in the *HIST1H* gene family.

3.4.3. Lung and head/neck cancers

A total of three loci were identified after conducting a meta-analysis of lung and head/neck cancers (Table 4). The first (rs380286, $p_{meta} = 2.72 \times 10^{-12}$) is mapped on *CLPTM1L* and *MIR4457*, genes encoding the catalytic subunit of human telomerase reverse transcriptase (McKay et al., 2017). The second (rs3117575, $p_{meta} = 8.06 \times 10^{-12}$) is in close proximity to *ABHD16A* and many other genes. *ABHD16A* is an emerging enzyme, mainly involved in lipid metabolism and intracellular signaling, leading to the metastasis of cancer (Xu et al., 2018). The third (rs2736100, $p_{meta} = 1.09 \times 10^{-9}$) is mapped on *TERT*, a gene that plays a central role in modulating telomerase activity in tumors (Colebatch et al., 2019).

TABLE 2 Cross-trait meta-analysis result between breast and ovarian cancers ($p_{meta} < 5 \times 10^{-8}$; single-trait $p < 0.05$).

SNP	Genome position	Allele	Breast cancer		Ovarian cancer		Meta	Genes within clumping region
			Beta	*p*	Beta	*p*	*p*	
rs1219648	chr10:123274062-123438122	A/G	0.2338	1.00×10^{-200}	−0.0266	0.0480	4.16×10^{-254}	*FGFR2*
rs630965	chr9:110759922-111073103	C/T	0.0992	3.21×10^{-54}	0.0301	0.0269	1.01×10^{-63}	CHCHD4P2[#]
rs3814113	chr9:16846323-16915021	T/C	0.0135	0.0410	−0.1780	9.40×10^{-36}	2.16×10^{-43}	*BNC2*
rs244353	chr17:52975892-53256579	G/A	−0.0754	1.14×10^{-28}	−0.0295	0.0399	1.40×10^{-31}	*COX11, STXBP4, TOM1L1*
rs6826366	chr4:175822759-175914966	G/A	−0.103	5.20×10^{-26}	−0.0426	0.0380	2.74×10^{-28}	*ADAM29*
rs7098100	chr10:21782842-22288132	G/A	0.0572	1.47×10^{-18}	0.0852	6.14×10^{-10}	4.41×10^{-27}	*CASC10, DNAJC1, MIR1915, MLLT10, SKIDA1*
rs4277389	chr17:43513441-44865603	A/G	−0.0484	2.01×10^{-10}	0.1151	1.20×10^{-12}	1.20×10^{-23}	*ARL17, CRHR1, KANSL1, LRRC37A, MAPT,MGC57346, MIR4315, NSF, PLEKHM1, SPPL2C, STH, WNT3*
rs4808616	chr19:17354825-17403033	C/A	0.0379	1.97×10^{-8}	0.1194	8.11×10^{-17}	1.94×10^{-23}	*ABHD8, ANKLE1, BABAM1, NR2F6, USHBP1*
rs10069690	chr5:1279790-1279790	C/T	0 .0599	7.79×10^{-17}	0.0830	3.42×10^{-8}	5.28×10^{-23}	*TERT*
rs2290202	chr15:91489705-91561182	G/T	−0.0728	1.87×10^{-15}	−0.0985	4.38×10^{-7}	4.20×10^{-20}	*PRC1, RCCD1, UNC45A, VPS33B*
rs851980	chr6:152008780-152070928	T/C	0.0619	1.13×10^{-18}	0.0400	0.0083	9.44×10^{-20}	*ESR1*
rs3769823	chr2:202119789-202271347	A/G	−0.0554	1.43×10^{-16}	−0.0289	0.0448	1.33×10^{-16}	*ALS2CR12, CASP8, TRAK2*
rs1474961	chr22:28324866-29318724	C/T	0.0667	1.74×10^{-10}	−0.1091	1.80×10^{-6}	2.02×10^{-15}	*CCDC117, CHEK2, HSCB, MIR5739, TTC28, XBP1, ZNRF3*
rs7017073	chr8:129143680-129218127	T/C	0.0572	2.32×10^{-14}	0.0359	0.0227	3.95×10^{-14}	*MIR1208*
rs35958868	chr17:29164023-29247715	G/A	−0.0426	1.37×10^{-9}	−0.0747	5.21×10^{-7}	5.44×10^{-13}	*ATAD5, TEFM*
rs10498635	chr14:93086918-93111120	C/T	−0.0571	3.46×10^{-12}	−0.0748	0.0109	9.26×10^{-13}	*RIN3*
rs381551	chr6:13638243-13722523	G/A	−0.0447	6.45×10^{-13}	−0.0297	0.0250	2.37×10^{-12}	*RANBP9*
rs12233670	chr4:38765720-38894380	C/T	0.0509	2.20×10^{-12}	0.0370	0.0178	8.05×10^{-12}	*FAM114A1, MIR574, TLR1, TLR6, TLR10*
rs2277509	chr14:91749595-91749595	C/A	0.0473	2.32×10^{-12}	0.0296	0.0381	1.53×10^{-11}	*CCDC88C*
rs2916074	chr19:19358672-19650096	G/A	0.0444	7.15×10^{-12}	0.0357	0.0097	2.04×10^{-11}	*CILP2, GATAD2A, HAPLN4, MAU2, NCAN, NDUFA13, SUGP1, TM6SF2, TSSK6, YJEFN3*
rs495828	chr9:136153875-136326248	G/T	0.0377	5.99×10^{-7}	0.0860	9.25×10^{-8}	7.21×10^{-11}	*ADAMTS13, C9orf96, CACFD1, MED22, REXO4, RPL7A, SNORD24, SURF*
rs720475	chr7:144074929-144074929	G/A	−0.0488	1.20×10^{-11}	−0.0308	0.0409	8.55×10^{-11}	*ARHGEF5*
rs2822991	chr21:16343812-16413682	T/C	0.0533	2.44×10^{-10}	0.0447	0.0094	5.50×10^{-10}	*NRIP1*
rs1550623	chr2:174207470-174212894	G/A	0.0531	5.39×10^{-10}	0.0360	0.0472	2.80×10^{-9}	*CDCA7*[#]
rs4743687	chr9:106856452-106898410	C/T	0.0322	2.29×10^{-7}	0.0545	4.15×10^{-5}	4.93×10^{-9}	*SMC2*
rs9878602	chr3:71517643-71535338	T/G	−0.0337	5.21×10^{-8}	0.0297	0.0243	3.72×10^{-8}	*FOXP1*
rs2941478	chr8:76474058-76476737	A/C	−0.0433	3.70×10^{-8}	−0.0430	0.0101	4.85×10^{-8}	*HNF4G*

[#]The nearest gene to this locus. SNP, single nucleotide polymorphisms; chr, chromosome; Allele, the character before the slash is the effect allele, and the character after the slash is the reference allele.

3.5. Overlapped gene–tissue pairs shared by cancer pairs in TWAS

To assess the association of gene expression in specific tissue between each pair of the five solid cancers, we performed TWAS. A total of 1,669 gene–tissue pairs are significantly associated with breast cancer after Benjamini–Hochberg correction (Supplementary Table S2), in addition to 418 gene–tissue pairs with ovarian cancer (Supplementary Table S3), 1,116 gene–tissue pairs with prostate cancer (Supplementary Table S4), 155 gene–tissue pairs

TABLE 3 Cross-trait meta-analysis result between breast and lung cancers ($p_{meta} < 5 \times 10^{-8}$; single-trait $p < 0.05$).

SNP	Genome position	Allele	Breast cancer		Lung cancer		Meta	Genes within clumping region
			Beta	*p*	Beta	*p*	*p*	
rs16886181	chr5:55983856-56306286	T/C	0.1730	8.89×10^{-98}	−0.0670	0.0078	4.57×10^{-122}	*MAP3K1, MIER3, SETD9*
rs2736108	chr5:1287194-1355058	C/T	−0.0622	3.88×10^{-19}	0.0988	6.49×10^{-5}	4.65×10^{-24}	*CLPTM1L, MIR4457, TERT*
rs7097066	chr10:80883083-80891631	G/A	0.0765	6.18×10^{-20}	−0.0571	0.0228	7.47×10^{-22}	*ZMIZ1*
rs3217992	chr9:21953137-22072719	C/T	−0.0581	1.18×10^{-19}	−0.0512	0.0227	1.78×10^{-21}	*C9orf53, CDKN2*
rs13214023	chr6:27413924-28366151	G/A	−0.0710	1.01×10^{-9}	0.1398	1.73×10^{-5}	8.48×10^{-13}	*HIST1H family, LINC01012, LOC100131289, NKAPL, OR2B, PGBD1, TOB2P1, ZKSCAN family*
rs10498635	chr14:93086918-93111120	C/T	−0.0571	3.46×10^{-12}	0.0513	0.0292	5.24×10^{-12}	*RIN3*
rs4971059	chr1:155148781-155666961	G/A	0.0424	4.83×10^{-11}	0.0549	0.0041	4.36×10^{-11}	*ASH1L, CLK2, DAP3, FAM189B, FDPS, GBA, GBAP1, HCN3, MIR92B, MIR555, MSTO1, MSTO2P, MTX1, MUC1, PKLR, POU5F1P4, RUSC1, SCAMP3, THBS3, TRIM46, YY1AP1*
rs13207082	chr6:26309908-27251379	A/T	−0.0710	2.10×10^{-9}	0.1225	0.0002	5.13×10^{-11}	*ABT1, BTN1A1, BTN2A, BTN3A, GUSBP2, HCG11, HIST1H, HMGN4, LINC00240, LOC285819, LOC100270746, MIR3143, PRSS16, ZNF322*
rs3117574	chr6:31081838-32064726	G/A	−0.0233	0.0286	0.1839	2.18×10^{-10}	4.27×10^{-10}	*ABHD16A, AIF1, APOM, ATP6V1G2, BAG6, C2, C4A, C4B, C6orf25, C6orf47, C6orf48, CCHCR1, CDSN, CFB, CLIC1, CSNK2B, CYP21A, DDAH2, DDX39B, DXO, EHMT2, GPANK1, HCG26, HCG27, HCP5, HLA-B, HLA-C, HSPA1 family, LSM2, LST1, LTA, LTB, LY6G family, MCCD1, MICA, MICB, MIR1236, MIR4646, MIR6832, MIR6891, MSH5, NCR3, NELFE, NEU1, NFKBIL1, POU5F1, PRRC2A, PSORS1C, SAPCD1, SKIV2L, SLC44A4, SNORA38, SNORD family, STK19, TCF19, TNF, TNXA, TNXB, VARS, VWA7, ZBTB12*
rs1550623	chr2:174207470-174212894	G/A	0.0531	5.39×10^{-10}	0.0655	0.0090	7.07×10^{-10}	*CDCA7*[#]
rs4930103	chr11:2018168-2024683	G/A	0.0382	6.60×10^{-10}	0.0389	0.0318	1.95×10^{-9}	*H19*
rs4635969	chr5:1308552-1308552	G/A	−0.0173	0.0276	−0.1444	5.33×10^{-10}	2.28×10^{-9}	*MIR4457*[#]
rs13212534	chr6:25874423-25983010	G/A	−0.0647	1.72×10^{-7}	0.1241	0.0005	5.70×10^{-9}	*SLC17A2, SLC17A3, TRIM38*
rs1707302	chr1:46600917-46603348	A/G	0.0364	2.95×10^{-8}	0.0625	0.0016	7.40×10^{-9}	*PIK3R3*[#]
rs13718	chr5:132384689-132444509	A/G	−0.0437	9.38×10^{-9}	−0.0560	0.0092	9.17×10^{-9}	*HSPA4*
rs224121	chr10:64447352-64588680	A/C	0.0396	7.38×10^{-8}	−0.0614	0.0041	1.72×10^{-8}	*ADO, EGR2*
rs2524005	chr6:29899677-29899677	G/A	−0.0297	0.0003	0.1080	2.12×10^{-6}	4.17×10^{-8}	*HLA-K*[#]
rs4808616	chr19:17403033-17403033	C/A	0.0379	1.97×10^{-8}	0.0414	0.0380	4.43×10^{-8}	*ABHD8*

[#]The nearest gene to the locus, SNP, single nucleotide polymorphisms; chr, chromosome; Allele, the character before the slash is the effect allele, and the character after the slash is the reference allele.

TABLE 4 Cross-trait meta-analysis result between the lung and head/neck cancers ($p_{meta} < 5 \times 10^{-8}$; single-trait $p < 0.05$).

SNP	Genome position	Allele	Lung cancer		Head/neck cancer		Meta	Genes within clumping region
			Beta	*p*	Beta	*p*	*p*	
rs380286	chr5:1299213-1355058	G/A	−0.1286	3.39×10^{-12}	0.0890	0.0332	2.72×10^{-12}	*CLPTM1L, MIR4457*
rs3117575	chr6:31094703-32059867	T/C	0.1839	2.37×10^{-10}	0.2990	0.0024	8.06×10^{-12}	*ABHD16A, AIF1, APOM, ATP6V1G2, BAG6, C2, C4A, C4B, C4B_2, C6orf25, C6orf47, C6orf48, CCHCR1, CFB, CLIC1, CSNK2B, CYP21A1P, CYP21A2, DDAH2, DDX39B, DXO, EHMT2, GPANK1, HCG26, HCG27, HCP5, HLA-B, HLA-C, HSPA1A, HSPA1B, HSPA1L, LOC102060414, LSM2, LST1, LTA, LTB, LY6G5B, LY6G5C, LY6G6C, LY6G6D, LY6G6E, LY6G6F, MCCD1, MICB, MIR1236, MIR4646, MIR6832, MIR6891, MSH5, MSH5-SAPCD1, NCR3, NELFE, NEU1, NFKBIL1, POU5F1, PRRC2A, PSORS1C1, PSORS1C2, PSORS1C3, SAPCD1, SKIV2L, SLC44A4, SNORA38, SNORD48, SNORD52, SNORD84, SNORD117, STK19, TNF, TNXA, TNXB, VARS, VWA7, ZBTB12*
rs2736100	chr5:1286516-1286516	C/A	−0.1062	3.97×10^{-9}	−0.0970	0.0210	1.09×10^{-9}	*TERT*

SNP, single nucleotide polymorphisms; chr, chromosome; Allele, the character before the slash is the effect allele, and the character after the slash is the reference allele.

with lung cancer (Supplementary Table S5), and 15 gene–tissue pairs with head/neck (Supplementary Table S6). Among them, 306 gene–tissue pairs are overlapped for the breast–ovarian cancer pair, and the tissues involved are scattered; however, a number of genes are almost concentrated in the clumping region of rs4277389 on chromosome 17, such as *CRHR1*, *LRRC37A*, and *MAPT* (Supplementary Table S7). Moreover, 23 gene–tissue pairs are overlapped for the breast–lung cancer pair, and most of the gene signals are observed in the 1q22 region, especially gene *GBAP1*, which is simultaneously significant in eight tissues (adipose, artery, breast, fibroblast cell, sigmoid colon, transverse colon, esophagus, and vagina) (Supplementary Table S7). In addition, one gene–tissue pair (*CFB*-pituitary) is overlapped for the lung–head/neck cancer pair (Supplementary Table S7).

3.6. Results of replication analysis in the UK Biobank cohort

In the replication analysis, we confirmed the significance of the genetic correlation between the breast and ovarian cancer pair ($r_g = 0.175$, $p = 0.0061$), the breast and lung cancer pair ($r_g = 0.125$, $p = 0.0018$), and the lung and head/neck cancer pair ($r_g = 0.506$, $p = 0.0005$) in the UK Biobank. Then, we used cross-cancer meta-analysis (RE2C) to identify the shared genes between each of the three cancer pairs. For the breast–ovarian cancer pair, nine loci showed genome-wide significance. Of these, genes *FGFR2*, *BNC2*, *ADAM29*, *ESR1*, *ATAD5*, and *TEFM* were replicated when compared with their specific consortium results (Supplementary Table S8). Moreover, six loci demonstrated significance in the breast–lung cancer pair. Some genes were found to be replicated, such as *MAP3K1* (rs12653202, $p_{meta} = 4.34 \times 10^{-23}$), *HIST1H family* (rs13214023, $p_{meta} = 2.83 \times 10^{-14}$), *ASH1L* (rs4971059, $p_{meta} = 5.47 \times 10^{-9}$), and *ZMIZ1* (rs7904249, $p_{meta} = 1.22 \times 10^{-8}$) (Supplementary Table S9). In addition, we identified two loci shared in the lung–head/neck cancer pair, but neither was replicated (Supplementary Table S10).

3.7. Results of biological analysis and pathway enrichment analysis

We observed shared genes enriched in *human T-cell leukemia virus 1 infection (HTLV-1)* and *antigen processing* and *presentation (APP)* pathways. *HTLV-1* was the first retrovirus discovered to cause adult T-cell leukemia (*ATL*), a highly aggressive blood cancer (Matsuoka and Jeang, 2011). The *APP* pathway is a key element for an efficient response to immune checkpoint inhibitor therapy, which can be exploited to enhance tumor immunogenicity and to increase the efficacy of immunotherapy. The use of immune checkpoint inhibitors has already shown significant clinical advances in a wide range of patients with cancer (D'Amico et al., 2022).

3.8. Results of protein–protein interaction network analysis

In total, we found 849 pairs of interaction in the PPI network (Supplementary Table S11). A total of 44 gene pairs have combined scores >0.95, in which the *ESR1-NRIP1* pair has the highest score of 0.999. *HIST1H* family genes around the 6p22.1 region show strong interactions with high scores. We observed 26 genes with degrees >20, most of which are *HIST1H* family genes, in addition to *ESR*, *HSPA4*, *TNF*, and *EHMT2* genes. *HIST1H* gene set expression was reported to be positively correlated with large tumor size, high grade, metastasis, and poor survival in patients with breast

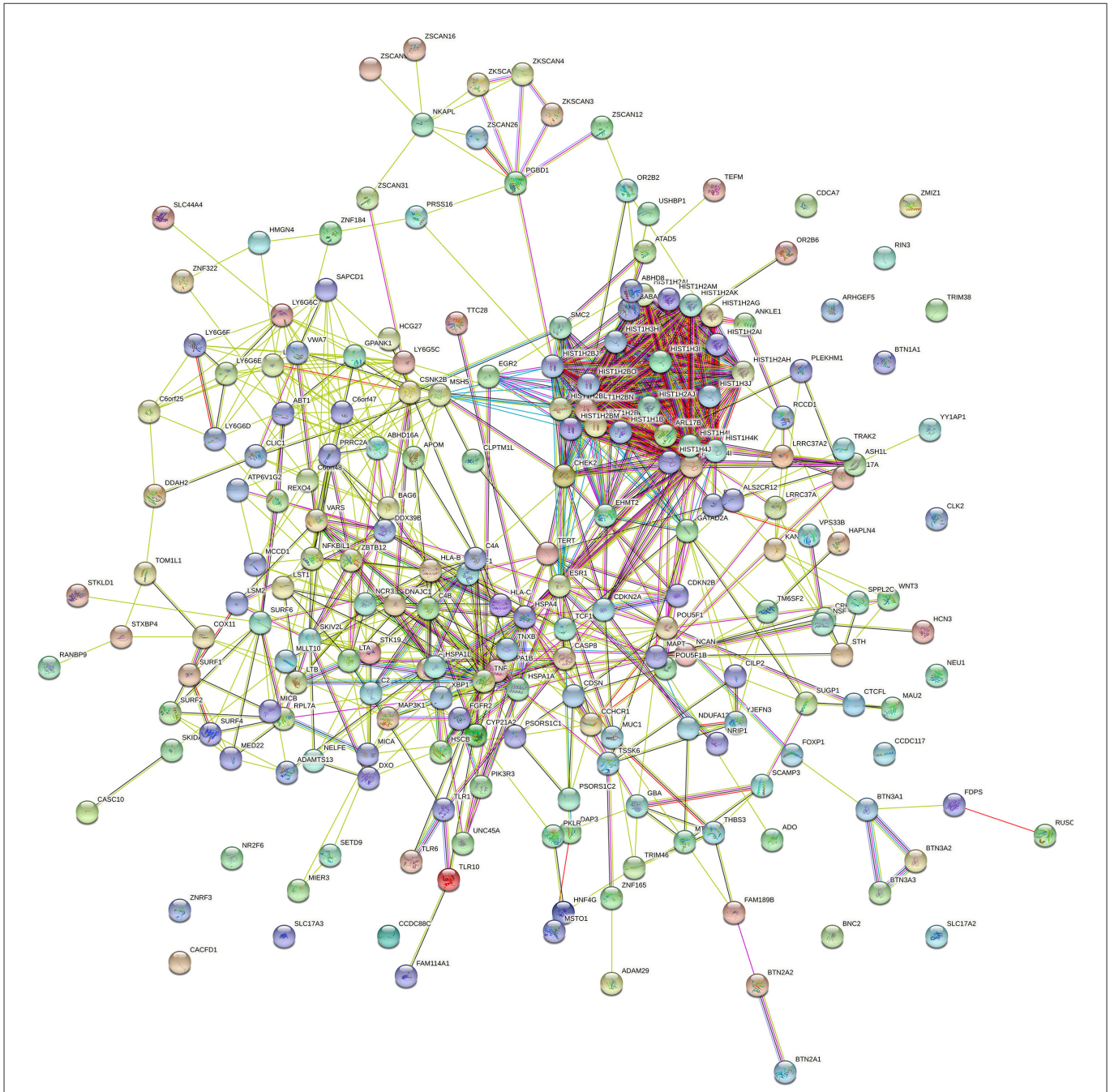


FIGURE 4
Protein–protein interaction network of share genes.

cancer (Liao et al., 2021), which were used as prognostic factors for survival prediction among patients with cervical cancer (Li et al., 2017). The PPI network for shared risk genes is shown in Figure 4.

4. Discussion

In the present study, we conducted a comprehensive analysis measuring the genetic correlation of five solid cancers, leveraging summary statistics from the current largest GWAS cancer consortia. We found significant positive genome-wide genetic correlations in three cancer pairs: breast–ovarian cancer, breast–lung cancer, and lung–head/neck cancer. Although the correlation in the prostate–head/neck cancer pair was up to 0.139, it failed to reach a significant level.

In partitioned genetic correlation, we detected positive genetic correlation and statistical significance in most function regions of the genome for the three cancer pairs, which showed significance in LDSC. Among them, the transcribed region had the strongest magnitude and significance. Most of the susceptibility variants detected by GWAS are located in non-coding regions and affect most cancers by affecting gene expression (Sud et al., 2017). Histone markers, including H3K27ac, H3K4me1, H3K4me1, and H3K9ac, are

important modifications that are associated with the dysregulation of many genes that play important roles in cancer development and progression (Kurdistani, 2007). Transcribed regions have diverse transcripts that impact cancer initiation and progression through several mechanisms of action (Gibert et al., 2022).

In the analysis of local genetic correlation, we identified a novel pleiotropic region (11q13) that showed a significant local genetic correlation between breast and prostate cancers. Although the 2q33 region was previously reported as a shared region for breast–ovarian and breast–prostate cancers (Jiang et al., 2019), we only observed the pleiotropic signal in the breast–ovarian cancer pair. In addition, the 9p21 and 10q26 regions we identified were indicated to share breast and prostate cancers (Jiang et al., 2019). However, we did not find any significant local correlation between the breast–lung cancer pair and the lung–head/neck cancer pair, which showed genome-wide statistical significance.

There are some common findings in the aforementioned three kinds of genetic correlation analyses. The three cancer pairs (breast–ovarian, breast–lung, and lung–head/neck), which were significant in genome-wide genetic association analysis, also showed strong significance in most functional categories in the partitioned genetic correlation analysis (Figure 2). In addition, the breast–ovarian cancer pair also showed strong significance in the 2q33 region in the local genetic correlation analysis (Figure 3A).

In the cross-cancer meta-analysis, we discovered 27 shared loci between breast and ovarian cancers, 18 shared loci between breast and lung cancers, and three shared loci between lung and head/neck cancers. Except for four of the shared loci that showed a significant association in trait-specific GWAS of two cancers, the others were newly discovered. In contrast, a previous study, which used the fixed-effect model-based approach ASSET, only identified one novel pleiotropic association at 1q22 involved in breast and lung cancers (Kar et al., 2016). This comparison demonstrated the high statistical power of the cross-cancer meta-analysis *via* the PLEIO test, which is based on a random-effect model.

In the TWAS analysis, we explored the significant gene–tissue pair in the five solid cancers by integrating GWAS summary statistics and GTEx tissue expression data. We identified 1,669 gene–tissue pairs associated with breast cancer at the transcriptome-wide level, in addition to 418 with ovarian cancer, 1,116 with prostate cancer, 155 with lung cancer, and 15 with head/neck cancer. Furthermore, we noticed that 306 gene–tissue pairs overlapped in the breast–ovarian cancer pair, 23 pairs overlapped in the breast–lung cancer pair, and one pair overlapped in the lung–head/neck cancer pair. These overlaps may implicate specific common regulations for biological function.

In the replication analysis, we found some shared genes in two independent cohorts, such as *FGFR2* for the breast–ovarian cancer pair and *MAP3K1* for the breast–lung cancer pair. Since there are more cases (tens of thousands) in specialized cohorts (such as BCAC for breast cancer) than those in the UK Biobank cohort (nearly 1,000), the small number of cases could affect the genetic correlation estimation; this may be the reason only a fraction of pleiotropic genes were found in UK Biobank replications.

The post-GWAS analyses enabled us to provide biological insights into the shared genes. We found that the shared genes were enriched in *HTLV-1* and *APP* pathways *via* pathway enrichment analysis. In the PPI network analysis, we observed obvious aggregations around HIST1H family genes, which were proved to be used as prognostic factors for survival prediction among patients with cancer (Li et al., 2017).

There are some advantages of the present study. On the one hand, we conducted a cross-cancer meta-analysis using two large-scale cohorts for each cancer separately, which facilitated the detection of novel associations. On the other hand, we performed association analyses under two kinds of mainstream random-effect model-based methods, which confirmed some of the discoveries. We also point out the limitations of this study. First, the UK Biobank cohort cancers we used in our replication analysis are not independent because there may be some shared cases and substantial shared controls among these five solid cancers. Moreover, the identified pleiotropic loci can be divided into causal and non-causal, and further experiments are required to distinguish the causal loci and to study their biological function. Finally, our study focuses on identifying shared genetic factors across five solid cancers, and their shared environmental factors require further investigation.

5. Conclusion

Identifying the shared genetic loci across five solid cancers plays an important role in the etiology and pathogenesis of each cancer. Our study finds several significant genetic correlations in specific cancer pairs, and their corresponding pleiotropic variants are detected by a cross-cancer meta-analysis. We observe shared genes enriched in the *human T-cell leukemia virus 1 infection (HTLV-1)* and *antigen processing and presentation (APP)* pathways. These shared genes and pathways may help to provide clues for future drug development.

Data availability statement

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

Author contributions

HG: conceptualization, methodology, and software. HG, WC, TL, and YZ: writing the original draft preparation. HG, WC, YZ, and BH: writing, reviewing, and editing. All authors have read and agreed to the published version of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the Natural Science Foundation of Hubei Province (Grant No. 2022CFB942) and the Talent Introduction Project of Hubei Normal University in 2021 (Grant No. HS2021RC013).

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1116592/full#supplementary-material

References

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., et al. (2010). The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* 28, 1045–1048. doi: 10.1038/nbt1010-1045

Bhattacharjee, S., Rajaraman, P., Jacobs, K., Wheeler, W., Melin, B., Hartge, P., et al. (2012). A subset based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.* 90, 821–835. doi: 10.1016/j.ajhg.2012.03.015

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Buniello, A., MacArthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012. doi: 10.1093/nar/gky1120

Cesaratto, L., Grisard, E., Coan, M., Zandona, L., De Mattia, E., Poletto, E., et al. (2016). BNC2 is a putative tumor suppressor gene in high-grade serous ovarian carcinoma and impacts cell survival after oxidative stress. *Cell Death Dis.* 7, e2374–e2374. doi: 10.1038/cddis.2016.278

Colebatch, A. J., Dobrovic, A., and Cooper, W. A. (2019). TERT gene: its function and dysregulation in cancer. *J. Clin. Pathol.* 72, 281–284. doi: 10.1136/jclinpath-2018-205653

D'Amico, S., Tempora, P., Melaiu, O., Lucarini, V., Cifaldi, L., Locatelli, F., et al. (2022). Targeting the antigen processing and presentation pathway to overcome resistance to immune checkpoint therapy. *Front. Immunol.* 13, 948297. doi: 10.3389/fimmu.2022.948297

Fehringer, G., Kraft, P., Pharoah, P. D., Eeles, R. A., Chatterjee, N., Schumacher, F. R., et al. (2016). Cross-cancer genome-wide analysis of lung, ovary, breast, prostate, and colorectal cancer reveals novel pleiotropic associations. *Cancer Res.* 76, 5103–5114. doi: 10.1158/0008-5472.CAN-15-2980

Ghoussaini, M., Song, H., Koessler, T., Al Olama, A. A., Kote-Jarai, Z., Driver, K. E., et al. (2008). Multiple loci with different cancer specificities within the 8q24 gene desert. *J. Natl. Cancer Inst.* 100, 962–966. doi: 10.1093/jnci/djn190

Gibert, M. K., Sarkar, A., Chagari, B., Roig-Laboy, C., Saha, S., Bednarek, S., et al. (2022). Transcribed ultraconserved regions in cancer. *Cells* 11, 1684. doi: 10.3390/cells11101684

Guo, H., An, J., and Yu, Z. (2020). Identifying shared risk genes for asthma, hay fever, and eczema by multi-trait and multiomic association analyses. *Front. Genet.* 11, 270. doi: 10.3389/fgene.2020.00270

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48, 245–252. doi: 10.1038/ng.3506

Iivonen, A.-P., Karkinen, J., Yellapragada, V., Sidoroff, V., Almusa, H., Vaaralahti, K., et al. (2021). Kallmann syndrome in a patient with Weiss–Kruszka syndrome and a de novo deletion in 9q31.2. *Eur. J. Endocrinol.* 185, 57–66. doi: 10.1530/EJE-20-1387

Jiang, X., Finucane, H. K., Schumacher, F. R., Schmit, S. L., Tyrer, J. P., Han, Y., et al. (2019). Shared heritability and functional enrichment across six solid cancers. *Nat. Commun.* 10, 431. doi: 10.1038/s41467-019-12095-8

Kar, S. P., Beesley, J., Amin Al Olama, A., Michailidou, K., Tyrer, J., Kote-Jarai, Z., et al. (2016). Genome-wide meta-analyses of breast, ovarian, and prostate cancer association studies identify multiple new susceptibility loci shared by at least two cancer types. *Cancer Discov.* 6, 1052–1067. doi: 10.1158/2159-8290.CD-15-1227

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

Kurdistani, S. K. (2007). Histone modifications as markers of cancer prognosis: a cellular view. *Br. J. Cancer* 97, 1–5. doi: 10.1038/sj.bjc.6603844

Lee, C. H., Eskin, E., and Han, B. (2017). Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. *Bioinformatics* 14, i379–i388. doi: 10.1093/bioinformatics/btx242

Lee, C. H., Shi, H., Pasaniuc, B., Eskin, E., and Han, B. (2021). PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics. *Am. J. Hum. Genet.* 108, 36–48. doi: 10.1016/j.ajhg.2020.11.017

Lesseur, C., Diergaarde, B., Olshan, A. F., Wunsch-Filho, V., Ness, A. R., Liu, G., et al. (2016). Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. *Nat. Genet.* 48, 1544–1550. doi: 10.1038/ng.3685

Li, X., Tian, R., Gao, H., Yang, Y., Williams, B. R. G., Gantier, M. P., et al. (2017). Identification of a histone family gene signature for predicting the prognosis of cervical cancer patients. *Sci. Rep.* 7, 16495. doi: 10.1038/s41598-017-16472-5

Liao, R., Chen, X., Cao, Q., Wang, Y., Miao, Z., Lei, X., et al. (2021). HIST1H1B promotes basal-like breast cancer progression by modulating CSF2 expression. *Front. Oncol.* 11, 780094. doi: 10.3389/fonc.2021.780094

Matsuoka, M., and Jeang, K.-T. (2011). Human T-cell leukemia virus type 1 (HTLV-1) and leukemic transformation: viral infectivity, Tax, HBZ and therapy. *Oncogene* 30, 1379–1389. doi: 10.1038/onc.2010.537

Matthew, S. L., Shea, J. A., Ben, E., Tom, R. G., Gibran, H., and Edoardo, M. (2021). The variant call format provides efficient and robust storage of GWAS summary statistics. *Genome Biol.* 22, 32. doi: 10.1186/s13059-020-02248-0

McKay, J. D., Hung, R. J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D. C., et al. (2017). Large scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.* 49, 1126–1132. doi: 10.1038/ng.3892

Michailidou, K., Lindstrom, S., Dennis, J., Beesley, J., Hui, S., Kar, S., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature* 551, 92–94. doi: 10.1038/nature24284

Mucci, L. A., Hjelmborg, J. B., Harris, J. R., Czene, K., Havelick, D. J., Scheike, T., et al. (2016). Familial risk and heritability of cancer among twins in nordic countries. *J. Am. Med. Assoc.* 315, 68. doi: 10.1001/jama.2015.17703

Pham, T. T., Angus, S. P., and Johnson, G. L. (2013). MAP3K1: genomic alterations in cancer and function in promoting cell survival or apoptosis. *Genes Cancer* 4, 419–426. doi: 10.1177/1947601913513950

Phelan, C. M., Kuchenbaecker, K. B., Tyrer, J. P., Kar, S. P., Lawrenson, K., Winham, S. J., et al. (2017). Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat. Genet.* 49, 680–691. doi: 10.1038/ng.3826

Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., et al. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* 47, 702–709. doi: 10.1038/ng.3285

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Rashkin, S. R., Graff, R. E., Kachuri, L., Thai, K. K., Alexeeff, S. E., Blatchins, M. A., et al. (2020). Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat. Commun.* 11, 4423. doi: 10.1038/s41467-020-18246-6

ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, The RACI Consortium, Finucane, H. K., Bulik-Sullivan, B., Gusev, A., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47, 1228–1235. doi: 10.1038/ng.3404

Schizophrenia Working Group of the Psychiatric Genomics Consortium, Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., et al. (2015). LD score

regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47, 291–295. doi: 10.1038/ng.3211

Schumacher, F. R., Al Olama, A. A., Berndt, S. I., Benlloch, S., Ahmed, M., Saunders, E. J., et al. (2018). Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* 50, 928–936. doi: 10.1038/s41588-018-0142-8

Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* 101, 737–751. doi: 10.1016/j.ajhg.2017.09.022

Sud, A., Kinnersley, B., and Houlston, R. S. (2017). Genome-wide association studies of cancer: current insights and future perspectives. *Nat. Rev. Cancer* 17, 692–704. doi: 10.1038/nrc.2017.82

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

The AACR Project GENIE Consortium, The AACR Project GENIE Consortium, Andre, F., Arnedos, M., Baras, A. S., Baselga, J., et al. (2017). AACR Project GENIE: powering precision medicine through an international consortium. *Cancer Discov.* 7, 818–831. doi: 10.1158/2159-8290.CD-17-0151

Wang, Y., McKay, J. D., Rafnar, T., Wang, Z., Timofeeva, M. N., Broderick, P., et al. (2014). Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat. Genet.* 46, 736–741. doi: 10.1038/ng.3002

Xu, J., Gu, W., Ji, K., Xu, Z., Zhu, H., and Zheng, W. (2018). Sequence analysis and structure prediction of ABHD16A and the roles of the ABHD family members in human disease. *Open Biol.* 8, 180017. doi: 10.1098/rsob.180017

Yu, H., Frank, C., Sundquist, J., Hemminki, A., and Hemminki, K. (2017). Common cancers share familial susceptibility: implications for cancer genetics and counselling. *J. Med. Genet.* 54, 248–253. doi: 10.1136/jmedgenet-2016-103932

Zhu, Z., Lin, Y., Li, X., Driver, J. A., and Liang, L. (2019). Shared genetic architecture between metabolic traits and Alzheimer's disease: a large-scale genome-wide cross-trait analysis. *Hum. Genet.* 138, 271–285. doi: 10.1007/s00439-019-01988-9

frontiers | Frontiers in Microbiology

TYPE Methods
PUBLISHED 22 February 2023
DOI 10.3389/fmicb.2023.1117027

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Baoshan Ma,
Dalian Maritime University, China
Lei Xu,
Shenzhen Polytechnic, China

*CORRESPONDENCE
Guohua Huang
✉ guohuahhn@163.com

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

RECEIVED 06 December 2022
ACCEPTED 17 January 2023
PUBLISHED 22 February 2023

CITATION
Qi Y, Zheng P and Huang G (2023) DeepLBCEPred: A Bi-LSTM and multi-scale CNN-based deep learning method for predicting linear B-cell epitopes. *Front. Microbiol.* 14:1117027. doi: 10.3389/fmicb.2023.1117027

DeepLBCEPred: A Bi-LSTM and multi-scale CNN-based deep learning method for predicting linear B-cell epitopes

Yue Qi, Peijie Zheng and Guohua Huang*

School of Information Engineering, Shaoyang University, Shaoyang, Hunan, China

The epitope is the site where antigens and antibodies interact and is vital to understanding the immune system. Experimental identification of linear B-cell epitopes (BCEs) is expensive, is labor-consuming, and has a low throughput. Although a few computational methods have been proposed to address this challenge, there is still a long way to go for practical applications. We proposed a deep learning method called DeepLBCEPred for predicting linear BCEs, which consists of bi-directional long short-term memory (Bi-LSTM), feed-forward attention, and multi-scale convolutional neural networks (CNNs). We extensively tested the performance of DeepLBCEPred through cross-validation and independent tests on training and two testing datasets. The empirical results showed that the DeepLBCEPred obtained state-of-the-art performance. We also investigated the contribution of different deep learning elements to recognize linear BCEs. In addition, we have developed a user-friendly web application for linear BCEs prediction, which is freely available for all scientific researchers at: http://www.biolscience.cn/DeepLBCEPred/.

KEYWORDS

epitope, B-cell, CNN, LSTM, protein sequence

1. Introduction

B cells are a class of leukocytes that are subtypes of lymphocytes in the immune system (Murphy and Weaver, 2012). B cells respond to foreign antigens by producing B-cell receptors that bind to the antigen (Murphy and Weaver, 2012). The sites where an antigen binds to an antibody are called epitopes (also known as antigenic determinants), which are specific pieces of the antigen. According to the structure and interaction with antibodies, epitopes can be grouped into conformational and linear epitopes (Huang and Honda, 2006). Conformational epitopes consist of discontinuous amino acid residues, and linear epitopes comprise contiguous amino acid residues. Identification of B-cell epitopes (BCEs) is not only essential for understanding the mechanisms of antigen–antibody interactions but also for vaccine design and therapeutic antibody development (Sharon et al., 2014; Shirai et al., 2014).

In contrast to labor-intensive and costly experimental methods, computational identification is cheap and high-throughput (Peng et al., 2022; Shen et al., 2022; Tian et al., 2022). Over the past decades, no less than 10 computational methods for predicting BCEs have been created (El-Manzalawy et al., 2008a, 2017; Ansari and Raghava, 2010; El-Manzalawy and Honavar, 2010; Jespersen et al., 2017; Ras-Carmona et al., 2021; Sharma et al., 2021; Alghamdi et al., 2022). The sequence is the simplest manifestation of protein but is pivotal for structure and function formation, and thus, the sequence compositions were frequently employed as a factor to identify BCEs (Chen et al., 2007; Singh et al., 2013). The sequence composition included but was not limited to the

physico-chemical profile (Ansari and Raghava, 2010), amino acid pair propensities (Chen et al., 2007; Singh et al., 2013), the composition–transition–distribution (CTD) profile (El-Manzalawy et al., 2008b), the tri-peptide similarity and propensity score (Yao et al., 2012), and subsequence kernel (El-Manzalawy et al., 2008a). The sequence composition might not represent all characteristics of the BCEs because it lacks position-related or order-related information. Other representations such as evolutionary features (Hasan et al., 2020) and structural features (Zhang et al., 2011) were explored as a determinant for identifying BCEs. There are three key factors responsible for the accuracy of identifying BCEs: the number and quality of BCEs served as training samples, representations, and learning algorithms. Jespersen et al. (2017) used the BCEs derived from crystal structures as the training set to improve prediction accuracy. Informative representations for BCEs are highly desirable but are too difficult to achieve in practice. Exploring new representations or combining various existing representations are two inevitable selections. Hasan et al. (2020) employed a non-parametric Wilcoxon rank-sum test to explore informative representations, while Chen et al. (2007) proposed a new amino acid pair antigenicity scale to represent BCEs. New representations are not always more informative than existing representations, and searching for an optimal combination of representations is both time-consuming and not always efficient. The learning algorithm is another factor to consider when developing methods for BCEs recognition, which plays equivalent roles with representations. The effectiveness of the learning algorithm might be associated with representations, that is, algorithms are representation-specific. It is ideal to search for an optimal scheme between algorithms and representations to enhance predictive performance. For example, Manavalan et al. (2018) explored six machine learning algorithms as well as appropriate representations and proposed an ensemble learning algorithm for linear BCEs recognition. Recently, deep learning is emerging as the next-generation artificial intelligence, exhibiting powerful learning ability. Deep learning has made a great breakthrough in areas such as image recognition (Krizhevsky et al., 2017) and mastering Go game as well as protein structure prediction (Silver et al., 2017; Cramer, 2021; Du et al., 2021; Jumper et al., 2021). To the best of our knowledge, there are more than three deep learning-based methods for predicting BCEs (Liu et al., 2020; Collatz et al., 2021; Xu and Zhao, 2022). Liu et al. demonstrated remarkable superiority of deep learning over traditional machine learning methods by cross-validation. Collatz et al. (2021) proposed a bi-directional long short-term memory (Bi-LSTM)-based deep learning method (called EpiDope) to identify linear BCEs. The EpiDope showed better performance in empirical experiments. Inspired by this, we improved EpiDope by adding a multi-scale convolutional neural networks (CNNs) to promote representation.

2. Dataset

We utilized the same benchmark datasets as BCEPS (Ras-Carmona et al., 2021) to evaluate and compare our proposed method with state-of-the-art methods. These datasets were initially extracted from the Immune Epitope Database (IEDB) (Vita et al., 2015, 2019), a repository of experimentally validated B- and T-cell epitopes (Vita et al., 2010). Ras-Carmona et al. (2021) constructed a nonredundant dataset $BCETD_{555}$ as the training set, which includes 555 sequences of BCEs and 555 sequences without BCEs. The BCEs in $BCETD_{555}$ consisted of linearized conformational B-cell epitopes (Ras-Carmona et al., 2021), obtained from the tertiary structure of the antigen–antibody complexes (Ras-Carmona et al., 2021). Ras-Carmona et al. (2021) used CD-HIT (Li and Godzik, 2006) to reduce sequence redundancy by deleting epitope sequences with more than 80% homology. Two independent testing sets were downloaded directly from https://www.mdpi.com/article/10.3390/cells10102744/s1 (Ras-Carmona et al., 2021): one set is the $ILED_{2195}$ dataset containing 2,195 sequences of linear BCEs and 2,195 sequences of non-BCEs and another set is the $IDED_{1246}$ dataset containing 1,246 sequences of BCEs and 1,246 sequences of non-BCEs. The $ILED_{2195}$ dataset and the $IDED_{1246}$ dataset were retrieved from the experimental B-cell epitope sequences retrieved from the IEDB database (Vita et al., 2015, 2019). All non-BCE sequences were extracted randomly from the same antigens as the BCEs.

3. Method

Figure 1 showed the schematic diagram of the proposed method DeepLBCEPred, which mainly consists of input, quantitative coding, embedding, feature extraction, and classification. Inputs are protein primary sequences that comprise 20 amino acid characters. For any sequences of less than a given length, we added the corresponding number of special characters 'X' at the end of it. Inputs were 21-character text sequences. The character sequence must be converted into an integer sequence by quantization coding using a conversion table (Table 1) so that the integer sequence can be embedded in a continuous vector using an embedding layer. Feature extraction includes two paralleling parts, one consisting mainly of the Bi-LSTM (Schuster and Paliwal, 1997) layer followed by a feed-forward attention layer (Raffel and Ellis, 2015) and another comprising multi-scale CNNs. Bi-LSTM (Schuster and Paliwal, 1997) was intended to extract the contextual semantics of the sequences, while the feed-forward attention (Raffel and Ellis, 2015) was intended to promote the semantic representation of protein sequences. CNNs at different scales reflect the representation of protein sequences at different scales. We used three different scale CNNs for extracting multi-scale features of sequences. The classification includes three fully connected layers, where the first has 64 neurons, the second has nine neurons, and the third has one neuron, which represents the probabilities of predicting inputs as BCEs.

3.1. Bi-LSTM

Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) is a specific type of recurrent neural network (RNN). Long short-term memory is capable of learning semantic relationships between long-distance words (Hochreiter and Schmidhuber, 1997). LSTM acts as a conveyor belt since it runs directly along the entire chain with only a few linear interactions (Hochreiter and Schmidhuber, 1997). At the heart of the LSTM is the cell state, which allows information to flow selectively by gate mechanisms (Hochreiter and Schmidhuber, 1997). There are three common gates: forget gate, input gate, and output gate. The forget gate is to determine how much information flows into the next cell state. The forget gate uses a sigmoid function to map the hidden

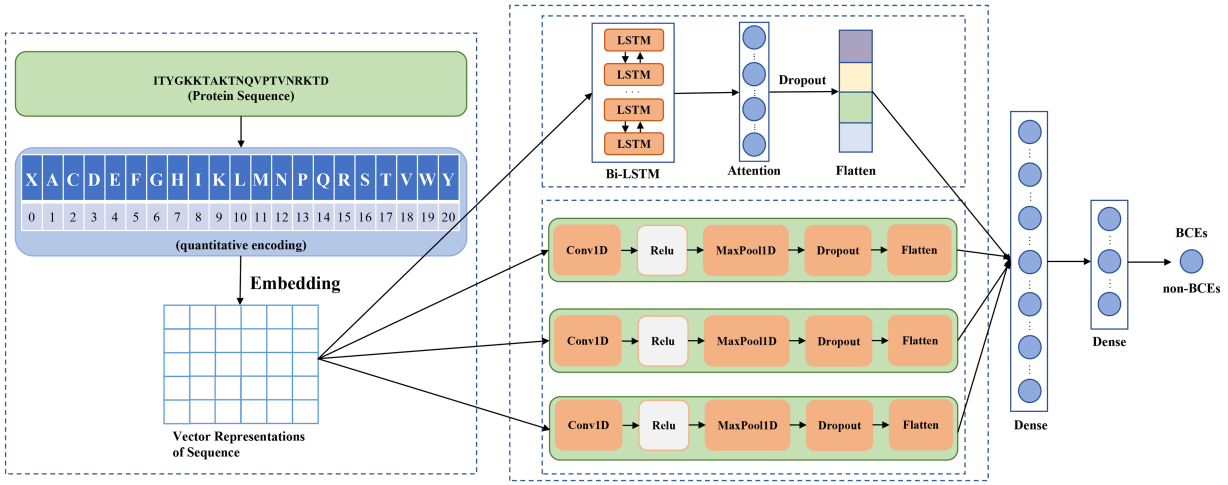


FIGURE 1
Schematic diagram of DeepLBCEPred.

TABLE 1 Conversion between amino acid and integer.

X	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

state and input variables into a number between 0 and 1. While 1 represents all information to pass completely, 0 indicates that no information is passing through. The question of how much information is added to the state cell is determined jointly by the input gate and the candidate cell state. The hidden state is updated jointly by the cell state and the output gate. To capture bidirectional dependency between words, we used Bi-LSTM (Schuster and Paliwal, 1997) to refine the semantics.

3.2. Feed-forward attention

Attention mechanisms have received increasing attention from the deep learning community due to better interpretability. Over the past 5 years, many attention mechanisms have been proposed to facilitate the interpretation of representations, such as well-known self-attention (Vaswani et al., 2017), feed-forward attention (Raffel and Ellis, 2015), external attention (Guo et al., 2022), and double attention (Chen et al., 2018). The attention mechanism is a scheme for assigning weights to different parts. Here, we employed feed-forward attention (Raffel and Ellis, 2015) for improving semantic representation. The attention weight was computed by

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T}\exp(e_k)} \tag{1}$$

where $e_t = a(h_t)$. h_t denoted the hidden state at the time step t in the Bi-LSTM and a was the learnable parameter. The output was computed by

$$c = \sum_{t=1}^{T}\alpha_t h_t \tag{2}$$

3.3. Multi-scale CNNs

CNNs are one of the most popular machine learning algorithms and thus have extensively been applied for image recognition. CNNs are mainly comprised of two elements: a convolutional layer and a pooling layer. At the heart of the CNNs is convolutional operation, which is to multiply the convolutional kernel by the receptive field in an element-wise manner and then sum them up. The convolution operation is accompanied by the activation function that produces a non-linear transformation. The activation function is associated with the efficiency and effectiveness of CNNs to a certain extent, and thus, selecting the appropriate activation function is critical to promote the performance of CNN. The commonly used activation function includes sigmoid, tanh, and rectified linear unit (ReLu). The convolutional kernel slides along the input to convolve with the receptive field to generate different feature maps. The convolutional kernel is shared by all the receptive fields in the same input and is the learnable parameter. The size of the convolutional kernel determines the different-scale characterization of the input. The larger size convolutional kernel reflects the global information, and the smaller size convolutional kernel discovers the local structure. To capture multi-scale characterization, we used multi-scale CNNs. The pooling layer is a sub-sampling operation, which reduces the dimensionality of the representation and thus speeds up the calculation. The pooling includes max, average, overlapping, and spatial

pyramid pooling (Wang et al., 2012; He et al., 2015; Khan et al., 2020). The dropout layer is used to randomly drop out some connections with a given probability to reduce computation and avoid overfitting (Hinton et al., 2012).

3.4. Fully connected layer

The fully connected layer is similar to the hidden layer in the multilayer perceptron where each neuron is linked to all the neurons in the previous layer. The outputs of the attention layer and the CNNs are of more than one dimension and, therefore, must be converted into one dimension to link to the fully connected layer. We used the flattened layer to bridge the fully connected layers and the non-fully connected layers. The flattened layers do not have any learnable parameters, and its actual task is to transform the shape of the data. We used three fully-connected layers. The first fully connected layer contains 64 neurons, the second contains 9 neurons, and the third contains only 1 neuron, which represents the probabilities of identifying inputs as BCEs.

4. Metrics

This is a binary classification question. The commonly used evaluation indices, namely, sensitivity (Sn), specificity (Sp), accuracy (ACC), and Matthews correlation coefficient (MCC), were employed to assess performance. Sn, Sp, ACC, and MCC were defined as follows:

$$Sn = \frac{TP}{TP + FN} \tag{3}$$

$$Sp = \frac{TN}{TN + FP} \tag{4}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FN) \times (TP + FP) \times (TN + FP)}} \tag{6}$$

where TP stands for the number of correctly predicted BCEs, TN stands for the number of correctly predicted non-BCEs, FP stands for the number of the non-BCEs, which were in reality non-BCEs but were erroneously predicted as BCEs, and FN stands for the number of the BCEs, which were in reality BCEs but were erroneously predicted as non-BCEs. Sn, Sp, and ACC lie between 0 and 1. The more the value is, the better performance there is. MCC considers not only TP and TN but also FP and FN and thus is generally viewed as a better measure for imbalanced datasets. MCC ranges from −1 to 1. An MCC of 1 implies perfect prediction, 0 implies random prediction, and −1 implies inverse prediction.

5. Results

Protein sequences of BCEs are of variable length, which is not favorable for subsequent sequence embedding. Therefore, we had to standardize the length of all BCEs sequences. The maximum length of BCEs sequences is 25, the average length is 16, and the minimum length is 11. We used 20% of the training BCEs in the training set to validate the effect of sequence length on the predictive performance. As listed in Table 2, the maximum length reached the best performance, followed by the average length and then the minimum length. Therefore, we uniformed all the sequences into a fixed length of 25.

Different scales reflect different scale characterization of the sequences. In this study, we used multi-scale CNNs. The combination of multi-scale CNNs is an optimal issue. To date, there is no scientific theory on how to effectively combine CNNs of different scales. In most cases, it relies on experience, especially experimental performances, to make choice. We investigated the effects of different scale combinations on the proposed method. The size of each scale ranged from 7 to 15 with a step size of 2. We used holdout to examine the performance. In the holdout, 80% was used to train the DeepLBCEPred and the remaining 20% was used to test the trained DeepLBCEPred, and the performance is presented in Table 3. When three scales of CNNs were set to 11, 13, and 15, respectively, the DeepLBCEPred reached the best ACC and the best MCC. Therefore, we set three scales to 11, 13, and 15, respectively.

6. Discussion

6.1. Comparison with existing models

As mentioned previously, many computational methods, including BepiPred (Larsen et al., 2006; Jespersen et al., 2017), LBtope (Singh et al., 2013), IBCE-EL (Manavalan et al., 2018), LBCEPred (Alghamdi et al., 2022), and BCEPS (Ras-Carmona et al., 2021), have been developed for BCEs prediction over the recent decades. We extensively compared the DeepLBCEPred with those methods by conducting 10-fold cross-validation on the $BCETD_{555}$ and independent tests on both $ILED_{2195}$ and $IDED_{1246}$. The 10-fold cross-validation divides $BCETD_{555}$ into 10 parts in equivalent or approximately equivalent size, with one part used to test the trained DeepLBCEPred by the other nine parts. The process is repeated 10 times. When this process is over, each sample is used only one time for testing the model and nine times for training the model. The independent test is to use $ILED_{2195}$ or $IDED_{1246}$ to test the DeepLBCEPred trained by $BCETD_{555}$. Table 4 lists their performance comparisons in 10-fold cross-validation. Compared to BCEPS, DeepLBCEPred increased ACC by 0.02, Sn by 0.05, and MCC by 0.03.

We compared DeepLBCEPred with five state-of-the-art algorithms by independent tests: BepiPred (Larsen et al., 2006; Jespersen et al., 2017), LBtope (Singh et al., 2013), LBCEPred (Alghamdi et al., 2022), IBCE-EL (Manavalan et al., 2018), and BCEPS (Ras-Carmona et al.,

TABLE 2 Performance over the various sequence length.

Sequence length	Sn	Sp	ACC	MCC
11(minimum)	0.64	0.78	0.70	0.42
16(average)	0.74	0.73	0.73	0.47
25(Maximum)	0.80	0.74	0.77	0.54

TABLE 3 Performance of different scale combinations.

Scale 1	Scale 2	Scale 3	Sn	Sp	ACC	MCC
7	9	11	0.79	0.58	0.69	0.38
7	9	13	0.61	0.84	0.72	0.46
7	9	15	0.86	0.55	0.72	0.43
7	11	13	0.70	0.81	0.75	0.50
7	11	15	0.75	0.68	0.72	0.43
7	13	15	0.63	0.80	0.71	0.43
9	11	13	0.72	0.81	0.76	0.53
9	11	15	0.71	0.70	0.70	0.40
9	13	15	0.78	0.73	0.76	0.51
11	13	15	0.80	0.74	0.77	0.54

TABLE 4 Ten-fold cross-validation results of DeepLBCEPred.

Ten-fold cross-validation	Sn	Sp	ACC	MCC
1	0.82	0.71	0.77	0.54
2	0.75	0.73	0.74	0.48
3	0.73	0.79	0.76	0.51
4	0.85	0.70	0.77	0.56
5	0.69	0.82	0.76	0.52
6	0.88	0.62	0.75	0.51
7	0.77	0.82	0.79	0.59
8	0.75	0.80	0.77	0.55
9	0.70	0.82	0.76	0.52
10	0.86	0.73	0.79	0.59
Ten-fold cross-validation (Mean)	0.78	0.75	0.77	0.54
BCEPS (Ras-Carmona et al., 2021)	0.73	0.78	0.75	0.51

2021). The LBCEPred is a newly developed method for predicting linear BCEs (Alghamdi et al., 2022). We uploaded two independent datasets to the LBCEPred webserver which are available at http://lbcepred.pythonanywhere.com/pred for prediction. All the predictive performances are listed in Tables 5 and 6. The DeepLBCEPred obtained a distinct superiority in ACC as well as MCC over BepiPred (Larsen et al., 2006; Jespersen et al., 2017), LBtope (Singh et al., 2013), LBCEPred (Alghamdi et al., 2022), and IBCE-EL (Manavalan et al., 2018). On the $ILED_{2195}$ independent dataset, the DeepLBCEPred exceeded the IBCE-EL by 0.16 of ACC as well as 0.33 of MCC, the LBtope by 0.17 of ACC as well as 0.35 of MCC, the BepiPred by 0.31 of ACC as well as 0.63 of MCC, and the LBCEPred by 0.15 of ACC as well as 0.31 of MCC. On the $IDED_{1246}$ independent dataset, the DeepLBCEPred exceeded the IBCE-EL by 0.14 of ACC as well as 0.26 of MCC, the LBtope by 0.10 of ACC as well as 0.21 of MCC, the BepiPred by 0.19 of ACC as well as 0.39 of MCC, and the LBCEPred by 0.15 of ACC as well as 0.29 of MCC. Compared with the BCEPS (Ras-Carmona et al., 2021), the DeepLBCEPred still has a slight advantage in ACC as well as MCC. The

TABLE 5 Comparison with existing models on the $ILED_{2195}$ independent dataset.

Model	Sn	Sp	ACC	MCC
IBCE-EL (Manavalan et al., 2018)	0.64	0.33	0.48	−0.04
LBtope (Singh et al., 2013)	0.36	0.58	0.47	−0.06
BepiPred (Jespersen et al., 2017)	0.24	0.43	0.33	−0.34
LBCEPred (Alghamdi et al., 2022)	0.74	0.24	0.49	−0.02
BCEPS (Ras-Carmona et al., 2021)	0.50	0.71	0.60	0.21
DeepLBCEPred	0.56	0.73	0.64	0.29

TABLE 6 Comparison with existing models on the $IDED_{1246}$ independent dataset.

Model	Sn	Sp	ACC	MCC
IBCE-EL (Manavalan et al., 2018)	0.86	0.20	0.53	0.09
LBtope (Singh et al., 2013)	0.40	0.74	0.57	0.14
BepiPred (Jespersen et al., 2017)	0.42	0.52	0.48	−0.04
LBCEPred (Alghamdi et al., 2022)	0.79	0.26	0.52	0.06
BCEPS (Ras-Carmona et al., 2021)	0.63	0.71	0.67	0.34
DeepLBCEPred	0.60	0.75	0.67	0.35

DeepLBCEPred increased ACC by 0.04 and MCC by 0.08 over the $ILED_{2195}$, and MCC by 0.01 over the $IDED_{1246}$.

6.2. Ablation experiments

Over the past decades, many basic structural units such as CNN, LSTM (Hochreiter and Schmidhuber, 1997), and self-attention (Vaswani et al., 2017) have been developed for deeper neural networks. Different units play different roles in characterizing studied objects. For instance, the CNN does well in refining local structure and Bi-LSTM (Schuster and Paliwal, 1997) in capturing long-distance dependency between words, while the self-attention emphasizes the key relationship of words. We investigated the contribution of a single individual to predicting BCEs by removing the corresponding part from the DeepLBCEPred. For the investigation, we performed independent tests after, respectively, removing (a) Bi-LSTM; (b) scale 1 in multi-scale CNNs; (c) scale 1 and scale 2 in multi-scale CNNs; (d) multi-scale CNNs; and (e) attention mechanism. As shown in Tables 7 and 8, the removal of these parts leads the performance to decrease. Deleting Bi-LSTM causes Sp to significantly reduce.

6.3. t-distributed stochastic neighbor embedding (t-SNE) visualization

We investigated the discriminative power of the representation captured by different layers in the DeepLBCEPred. We used the t-SNE (Van der Maaten and Hinton, 2008) to plot a scattering diagram of the first two components in the $ILED_{2195}$ dataset. The initial embedding was highly indistinguishable. The representations output by multi-scale CNNs and Bi-LSTM were significantly distinguishable. The feed-forward attention improved representations to a tiny extent. The overall combined representations promoted discriminative ability, demonstrating the ability to distinguish between BCEs and non-BCEs from a representational perspective (Figure 2).

6.4. Deep learning community due to better interpretability web server

To help researchers use DeepLBCEPred more easily, we have exploited a user-friendly web server, which is available at: http://www.biolscience.cn/DeepLBCEPred/. As shown in Figure 3, after the user writes a sequence in the text box or uploads a sequence file and clicks "Submit," the page will display the final prediction result. It is worth noting that only the sequence in FASTA format is allowed, and the input sequence must consist of the characters in "ACDEFGHIKLMNPQRSTVWY." Otherwise, it will prompt Format Error. To clear the contents of the text box, click "Clear." Click "Example" to see a sample. The dataset used in this study can be downloaded from the bottom left corner of the page.

TABLE 7 Comparison of five ablation experiments on the $ILED_{2195}$ independent dataset.

Ablation experiments	Sn	Sp	ACC	MCC
delete Bi-LSTM	0.69	0.53	0.61	0.22
delete scale 1	0.56	0.70	0.63	0.26
delete scale 1_2	0.53	0.68	0.60	0.21
delete Multi-scale CNN	0.45	0.71	0.58	0.17
delete Attention mechanism	0.55	0.66	0.60	0.21
DeepLBCEPred	0.56	0.73	0.64	0.29

TABLE 8 Comparison of five ablation experiments on the $IDED_{1246}$ independent dataset.

Ablation experiments	Sn	Sp	ACC	MCC
delete Bi-LSTM	0.79	0.55	0.67	0.35
delete scale 1	0.62	0.70	0.66	0.31
delete scale 1_2	0.66	0.70	0.68	0.36
delete Multi-scale CNN	0.61	0.73	0.67	0.35
delete Attention mechanism	0.68	0.66	0.67	0.35
DeepLBCEPred	0.60	0.75	0.67	0.35

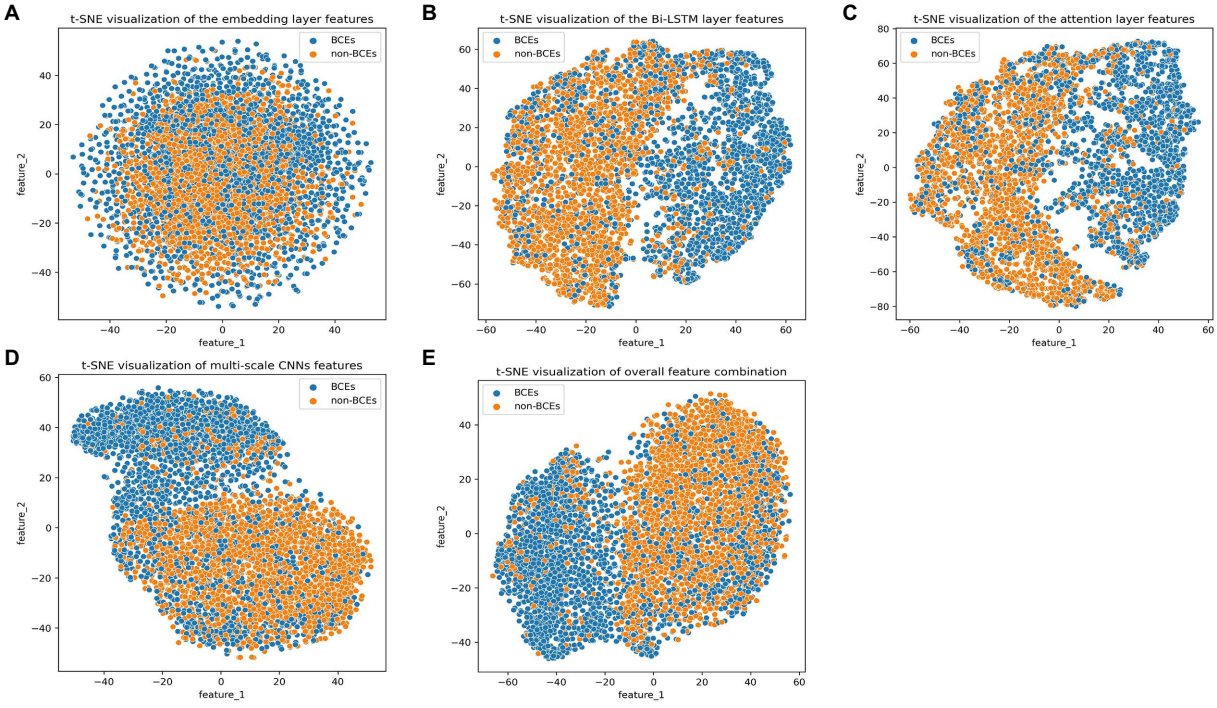


FIGURE 2
t-SNE visualization of outputs of **(A)** the embedding layer, **(B)** the Bi-LSTM layer, **(C)** the attention layer, **(D)** the multi-scale CNNs, and **(E)** overall combination.

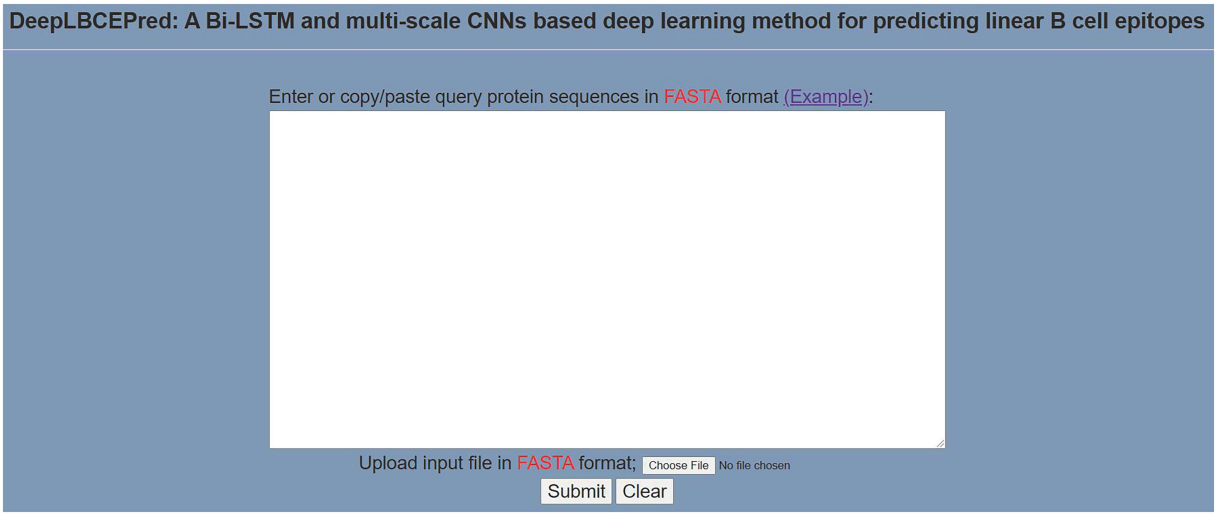


FIGURE 3
Prediction page of the web server.

7. Conclusion

B-cell epitopes play critical roles in antigen–antibody interactions and vaccine design. Identification of BCEs is a key foundation for understanding BCEs functions. In the article, we developed a deep learning-based method DeepLBCEPred to predict linear BCEs. The DeepLBCEPred is an end-to-end method that takes protein sequence as input and directly outputs decisions about BCEs. On the benchmark datasets, DeepLBCEPred reached state-of-the-art performance and was implemented as a user-friendly web server for ease of use.

Data availability statement

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

Author contributions

YQ conducted experiments, analysis, and wrote the original manuscript. PZ conducted experiments and developed the software. GH conceived the methodology, supervised the project and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work is supported by Hunan Province Natural Science Foundation of China (2022JJ50177), by Scientific Research Fund of Hunan Provincial Education Department (21A0466), and the Shaoyang University Innovation Foundation for Postgraduate (CX2021SY037).

Conflict of interest

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

Publisher's note

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

References

Alghamdi, W., Attique, M., Alzahrani, E., Ullah, M. Z., and Khan, Y. D. (2022). LBCEPred: a machine learning model to predict linear B-cell epitopes. *Brief. Bioinform.* 23:bbac035. doi: 10.1093/bib/bbac035

Ansari, H. R., and Raghava, G. P. S. (2010). Identification of conformational B-cell epitopes in an antigen from its primary sequence. *Immunome Res.* 6, 6–9. doi: 10.1186/1745-7580-6-6

Chen, Y., Kalantidis, Y., Li, J., Yan, S., and Feng, J. (2018). "A^2-nets: double attention networks" in *Advances in Neural Information Processing Systems*. eds. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Chen, J., Liu, H., Yang, J., and Chou, K.-C. (2007). Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids* 33, 423–428. doi: 10.1007/s00726-006-0485-9

Collatz, M., Mock, F., Barth, E., Hölzer, M., Sachse, K., and Marz, M. (2021). EpiDope: a deep neural network for linear B-cell epitope prediction. *Bioinformatics* 37, 448–455. doi: 10.1093/bioinformatics/btaa773

Cramer, P. (2021). AlphaFold2 and the future of structural biology. *Nat. Struct. Mol. Biol.* 28, 704–705. doi: 10.1038/s41594-021-00650-1

Du, Z., Su, H., Wang, W., Ye, L., Wei, H., Peng, Z., et al. (2021). The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.* 16, 5634–5651. doi: 10.1038/s41596-021-00628-9

El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2008a). Predicting linear B-cell epitopes using string kernels. *J Mol Recognit.* 21, 243–255. doi: 10.1002/jmr.893

El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2008b). Predicting flexible length linear B-cell epitopes. *Comput. Syst. Bioinformatics (World Scientific)* 7, 121–132. doi: 10.1142/9781848162648_0011

El-Manzalawy, Y., Dobbs, D., and Honavar, V. G. (2017). In silico prediction of linear B-cell epitopes on proteins. *Methods Mol. Biol.* 1484, 255–264. doi: 10.1007/978-1-4939-6406-2_17

El-Manzalawy, Y., and Honavar, V. (2010). Recent advances in B-cell epitope prediction methods. *Immunome Res.* 6, S2–S9. doi: 10.1186/1745-7580-6-S2-S2

Guo, M.-H., Liu, Z.-N., Mu, T.-J., and Hu, S.-M. (2022). Beyond self-attention: external attention using two linear layers for visual tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* 14, 1–13. doi: 10.1109/TPAMI.2022.3211006

Hasan, M. M., Khatun, M. S., and Kurata, H. (2020). iLBE for computational identification of linear B-cell epitopes by integrating sequence and evolutionary features. *Genom. Proteom. Bioinform.* 18, 593–600. doi: 10.1016/j.gpb.2019.04.004

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 1904–1916. doi: 10.1109/TPAMI.2015.2389824

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv. arXiv:1207.0580 [Epub ahead of preprint]. doi: 10.48550/arXiv.1207.0580

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Huang, J., and Honda, W. (2006). CED: a conformational epitope database. *BMC Immunol.* 7, 1–8. doi: 10.1186/1471-2172-7-7

Jespersen, M. C., Peters, B., Nielsen, M., and Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.* 45, W24–W29. doi: 10.1093/nar/gkx346

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi: 10.1038/s41586-021-03819-2

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* 53, 5455–5516. doi: 10.1007/s10462-020-09825-6

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Larsen, J. E. P., Lund, O., and Nielsen, M. (2006). Improved method for predicting linear B-cell epitopes. *Immunome Res.* 2, 1–7. doi: 10.1186/1745-7580-2-2

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

Liu, T., Shi, K., and Li, W. (2020). Deep learning methods improve linear B-cell epitope prediction. *BioData Mining.* 13, 1–13. doi: 10.1186/s13040-020-00211-0

Manavalan, B., Govindaraj, R. G., Shin, T. H., Kim, M. O., and Lee, G. (2018). iBCE-EL: a new ensemble learning framework for improved linear B-cell epitope prediction. *Front. Immunol.* 9:1695. doi: 10.3389/fimmu.2018.01695

Murphy, K., and Weaver, C. (2012). "The induced responses of innate immunity" in *Janeway's Immunobiology. 8th ed* eds. J. Scobie, E. Lawrence, J. Moldovan, G. Lucas, B. Goatly and M. Toledo (New York, NY: Garland Science), 75–125.

Peng, L., Wang, F., Wang, Z., Tan, J., Huang, L., Tian, X., et al. (2022). Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies. *Brief. Bioinform.* 23:bbac234. doi: 10.1093/bib/bbac234

Raffel, C., and Ellis, D. P. (2015). Feed-forward networks with attention can solve some long-term memory problems. arXiv preprint arXiv:1512.08756 [Epub ahead of preprint]. doi: 10.48550/arXiv.1512.08756

Ras-Carmona, A., Pelaez-Prestel, H. F., Lafuente, E. M., and Reche, P. A. (2021). BCEPS: a web server to predict linear B cell epitopes with enhanced immunogenicity and cross-reactivity. *Cells* 10:2744. doi: 10.3390/cells10102744

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681. doi: 10.1109/78.650093

Sharma, S., Vashisht, S., Gaur, S. N., Lavasa, S., and Arora, N. (2021). Identification of B cell epitopes of per a 5 allergen using bioinformatic approach. *Immunobiology* 226:152146. doi: 10.1016/j.imbio.2021.152146

Sharon, J., Rynkiewicz, M. J., Lu, Z., and Yang, C. Y. (2014). Discovery of protective B-cell epitopes for development of antimicrobial vaccines and antibody therapeutics. *Immunology* 142, 1–23. doi: 10.1111/imm.12213

Shen, L., Liu, F., Huang, L., Liu, G., Zhou, L., and Peng, L. (2022). VDA-RWLRLS: an anti-SARS-CoV-2 drug prioritizing framework combining an unbalanced bi-random walk and Laplacian regularized least squares. *Comput. Biol. Med.* 140:105119. doi: 10.1016/j.compbiomed.2021.105119

Shirai, H., Prades, C., Vita, R., Marcatili, P., Popovic, B., Xu, J., et al. (2014). Antibody informatics for drug discovery. *Biochim Biophys Acta* 1844, 2002–2015. doi: 10.1016/j.bbapap.2014.07.006

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of go without human knowledge. *Nature* 550, 354–359. doi: 10.1038/nature24270

Singh, H., Ansari, H. R., and Raghava, G. P. S. (2013). Improved method for linear B-cell epitope prediction using antigen's primary sequence. *PLoS One* 8:e62216. doi: 10.1371/journal.pone.0062216

Tian, G., Wang, Z., Wang, C., Chen, J., Liu, G., Xu, H., et al. (2022). A deep ensemble learning-based automated detection of COVID-19 using lung CT images and vision transformer and ConvNeXt. *Front. Microbiol.* 13:1024104. doi: 10.3389/fmicb.2022.1024104

Van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need" in *Advances in Neural Information Processing Systems* eds. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., et al. (2019). The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.* 47, D339–D343. doi: 10.1093/nar/gky1006

Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., et al. (2015). The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* 43, D405–D412. doi: 10.1093/nar/gku938

Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., et al. (2010). The immune epitope database 2.0. *Nucleic Acids Res.* 38, D854–D862. doi: 10.1093/nar/gkp1004

Wang, T., Wu, D. J., Coates, A., and Ng, A. Y. (2012). "End-to-end text recognition with convolutional neural networks," in *Proceedings of the 21st International Conference on Pattern Recognition (IEEE)*, pp. 3304–3308.

Xu, H., and Zhao, Z. (2022). NetBCE: an interpretable deep neural network for accurate prediction of linear B-cell epitopes. bioRxiv [Epub ahead of preprint]. doi: 10.1101/2022.05.23.493092

Yao, B., Zhang, L., Liang, S., and Zhang, C. (2012). SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. *PLoS One* 7:e45152. doi: 10.1371/journal.pone.0045152

Zhang, W., Xiong, Y., Zhao, M., Zou, H., Ye, X., and Liu, J. (2011). Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature. *BMC Bioinform.* 12, 1–10. doi: 10.1186/1471-2105-12-341

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 17 March 2023
DOI 10.3389/fmicb.2023.1138674

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology,
China

REVIEWED BY
Nizhuan Wang,
ShanghaiTech University,
China
Qi Dai,
Zhejiang Sci-Tech University,
China

*CORRESPONDENCE
Tao Huang
✉ tohuangtao@126.com
Yu-Dong Cai
✉ cai_yud@126.com

†These authors have contributed equally to this work

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

RECEIVED 06 January 2023
ACCEPTED 01 March 2023
PUBLISHED 17 March 2023

CITATION
Li J, Ren J, Liao H, Guo W, Feng K, Huang T and Cai Y-D (2023) Identification of dynamic gene expression profiles during sequential vaccination with ChAdOx1/BNT162b2 using machine learning methods.
Front. Microbiol. 14:1138674.
doi: 10.3389/fmicb.2023.1138674

Identification of dynamic gene expression profiles during sequential vaccination with ChAdOx1/BNT162b2 using machine learning methods

Jing Li[1†], JingXin Ren[2†], HuiPing Liao[3†], Wei Guo[4], KaiYan Feng[5], Tao Huang[6,7*] and Yu-Dong Cai[2*]

[1]School of Computer Science, Baicheng Normal University, Baicheng, Jilin, China, [2]School of Life Sciences, Shanghai University, Shanghai, China, [3]Changping Laboratory, Beijing, China, [4]Key Laboratory of Stem Cell Biology, Shanghai Jiao Tong University School of Medicine (SJTUSM) and Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences (CAS), Shanghai, China, [5]Department of Computer Science, Guangdong AIB Polytechnic College, Guangzhou, China, [6]CAS Key Laboratory of Computational Biology, Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Science, Shanghai, China, [7]CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

To date, COVID-19 remains a serious global public health problem. Vaccination against SARS-CoV-2 has been adopted by many countries as an effective coping strategy. The strength of the body's immune response in the face of viral infection correlates with the number of vaccinations and the duration of vaccination. In this study, we aimed to identify specific genes that may trigger and control the immune response to COVID-19 under different vaccination scenarios. A machine learning-based approach was designed to analyze the blood transcriptomes of 161 individuals who were classified into six groups according to the dose and timing of inoculations, including I-D0, I-D2-4, I-D7 (day 0, days 2–4, and day 7 after the first dose of ChAdOx1, respectively) and II-D0, II-D1-4, II-D7-10 (day 0, days 1–4, and days 7–10 after the second dose of BNT162b2, respectively). Each sample was represented by the expression levels of 26,364 genes. The first dose was ChAdOx1, whereas the second dose was mainly BNT162b2 (Only four individuals received a second dose of ChAdOx1). The groups were deemed as labels and genes were considered as features. Several machine learning algorithms were employed to analyze such classification problem. In detail, five feature ranking algorithms (Lasso, LightGBM, MCFS, mRMR, and PFI) were first applied to evaluate the importance of each gene feature, resulting in five feature lists. Then, the lists were put into incremental feature selection method with four classification algorithms to extract essential genes, classification rules and build optimal classifiers. The essential genes, namely, *NRF2*, *RPRD1B*, *NEU3*, *SMC5*, and *TPX2*, have been previously associated with immune response. This study also summarized expression rules that describe different vaccination scenarios to help determine the molecular mechanism of vaccine-induced antiviral immunity.

KEYWORDS

SARS-CoV-2, vaccination, immune response, machine learning, blood transcriptome

1. Introduction

Coronavirus disease-19 (COVID-19) is a pandemic infectious disease that is currently affecting many people in approximately 200 countries around the world. It is caused by acute respiratory syndrome coronavirus-2 (SARS-CoV-2), a highly pathogenic coronavirus that belongs to the subfamily Coronaviridae. The SARS-CoV-2 genome contains a variety of structural and nonstructural proteins. The rapid rate at which the virus mutates and spreads has created enormous challenges for prevention and control efforts. Currently, vaccination against SARS-CoV-2 is accepted as an effective strategy against COVID-19 (Folegatti et al., 2020; Amano et al., 2022), with two or more doses giving better protection than one dose alone. The risk of death from COVID-19 varies widely in different countries and may be related to factors such as vaccination rate and number of vaccinations (Masic et al., 2020).

When the body receives the first dose of the COVID-19 vaccine (basic immunization injection), it recognizes viral-specific antigens and produces antibodies and memory cells against SARS-CoV-2. However, the amount of antibodies produced by the primary immune response is much lower than the level required to resist viral invasion. Early clinical trials showed that with just one dose (initial exposure), the body's resistance to SARS-CoV-2 is very low at about 50%. Therefore, a second vaccine dose and a booster shot have been recommended after a period of time (3–4 weeks). When exposed to the same antigen twice, the memory cells that have been generated in the human body respond rapidly, producing sufficient antibodies and a strong secondary immune response. Therefore, two doses of vaccination are more effective for protection. The ChAdOx1 nCoV-19 (AZD1222) vaccine is constructed from a replication-defective simian adenovirus vector encoding the spike (S) protein of SARS-CoV-2. Clinical trials have shown that the ChAdOx1 vaccine is 74% protective against symptomatic COVID-19 (Cross et al., 2003). Meanwhile, BNT162b2, also known as the Pfizer-BioNTech COVID-19 vaccine, is a messenger RNA (mRNA) vaccine that has been approved by the US FDA for the prevention of COVID-19 caused by the SARS-CoV-2 Beta coronavirus. A heterologous ChAdOx1-S-nCoV-19 and BNT162b2 vaccination combination provides better protection against severe SARS-CoV-2 infection in a real-world observational study ($n = 13{,}121$). Studies have shown that T-cell responses following ChAdOx1 vaccination were higher than those elicited by BNT162b2. Meanwhile, T-cell responses elicited by BNT162b2 booster doses were enhanced in different vaccination strategies. Both homologous and heterologous vaccinations were able to induce progressively increased frequencies of CD4 and CD8 T cells. However, the heterologous combination elicited stronger CD4 T-cell responses; CD8 T-cell responses were also progressively stronger after the booster dose (Pozzetto et al., 2021). The tolerability and safety profile of BNT162b2 at 30 μg administered as a 2-dose regimen are favorable. In participants who received only one ChAdOx1 dose, antibodies against the SARS-CoV-2 spike protein peaked at day 28 (median 157 ELISA units [EU]); on day 56, the median was 119 EU. Among participants who received the booster dose, the median antibody at day 56 was 639 EU (Folegatti et al., 2020). Studies have demonstrated the efficacy of a two-dose regimen of the BNT162b2 vaccine (Mizrahi et al., 2021).

An increasing number of studies have confirmed that high-throughput sequencing data information can provide important guidance for revealing the pathogenic mechanism of diseases and tackling various medical problems (Dai et al., 2018; Kong et al., 2020; Yang et al., 2020, 2022). Our team has long been working on using machine learning analysis methods to screen for disease-related signatures and explain their pathogenic mechanisms. We divided the data on 161 people vaccinated against COVID-19 into six groups according to the injection and vaccination time, aiming to further explore changes in blood gene expression after different doses, especially the molecular characteristics of antiviral immunity. A variety of algorithms were used to analyze gene expression information on vaccines from different vaccinations. The algorithms included feature ranking algorithms, such as least absolute shrinkage and selection operator (Lasso) (Tibshirani, 2011), light gradient-boosting machine (LightGBM) (Ke et al., 2017), Monte Carlo feature selection (MCFS) (Draminski et al., 2007), max-relevance and min-redundancy (mRMR) (Peng et al., 2005), and permutation feature importance (PFI) (Fisher et al., 2019), as well as classification algorithms, such as decision tree (DT) (Safavian and Landgrebe, 1991), random forest (RF) (Breiman, 2001), K-nearest neighbor (KNN) (Cover and Hart, 1967), and support vector machine (SVM) (Cortes and Vapnik, 1995). Based on feature ranking algorithms, gene feature lists were obtained, which were subjected to incremental feature selection (IFS) method (Liu and Setiono, 1998), incorporating four classification algorithms, for extracting essential genes, classification rules, and build optimal classifiers. This study revealed that blood gene expression changed after the initial immunization and booster vaccination. A number of important genes (e.g., *NRF2*, *RPRD1B*, *NEU3*, *SMC5*, and *TPX2*) may be closely related to the antiviral immunity induced by vaccines. These findings are helpful for understanding the importance of vaccination and boosting injections by revealing the effects of different injections on the expression of immune-related molecules in the host and by providing a reference for viral immune intervention strategies for COVID-19.

2. Materials and methods

The workflow of the machine learning framework is shown in Figure 1. The samples were grouped according to the number of inoculations and inoculation time. The genes were subsequently ranked using five methods and further processed by IFS method with four classification algorithms. By observing the performance of the classifiers, a number of key genes and summarized quantitative classification rules were identified. Last, the key genes were functionally enriched to determine the biological processes involved in their action. The methods used are described in detail in this section.

2.1. Data

Blood transcriptome data from 161 individuals were obtained from the GEO database under the registration number GSE201533 (Lee et al., 2022a). We divided the vaccinees into two groups: I for the first COVID-19 vaccination dose and II for the second dose. For the first group, three subsets were included: I-D0, I-D2-4, and I-D7, meaning day 0, days 2–4, and day 7 after the first dose of ChAdOx1, respectively. There were also three subsets in the second group, say II-D0, II-D1-4, II-D7-10, meaning the day 0, days 1–4, and days 7–10 after the second dose of BNT162b2, respectively. Four of the vaccinees received a second

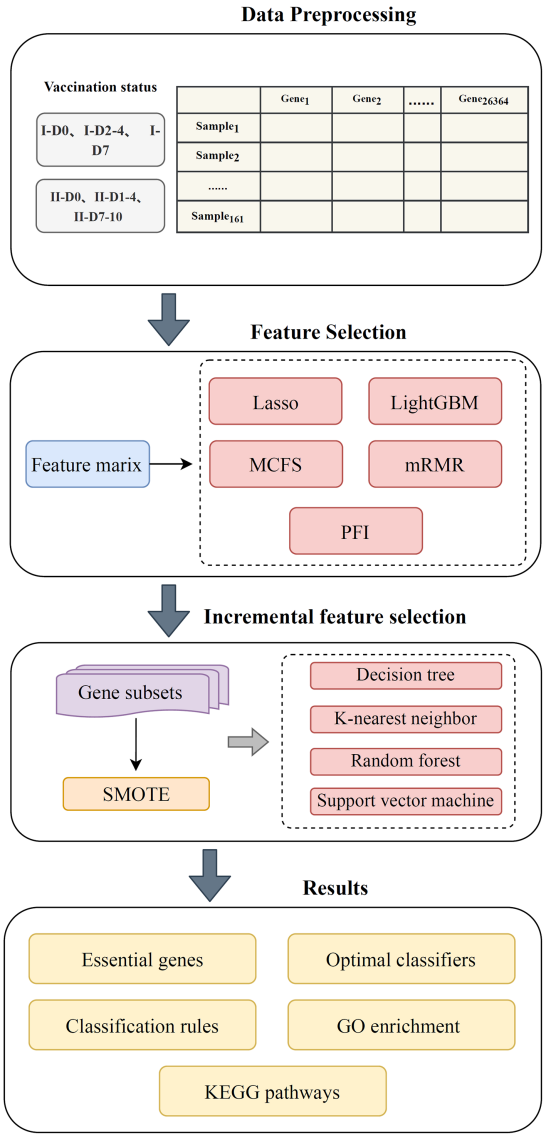


FIGURE 1
Flow chart of the entire analysis process. The blood transcriptome data of 161 vaccinees with different COVID-19 vaccination was investigated. Each vaccinee was represented by 26,364 gene expression levels. Five feature ranking algorithms (Lasso, LightGBM, MCFS, mRMR, and PFI) were used to rank gene features according to their importance. Subsequently, these lists were fed into incremental feature selection method, which contained four classification algorithms, to extract essential genes, classification rules, and build optimal classifiers.

dose of ChAdOx1. Table 1 shows the number of samples in each subset. Each sample was represented by 26,364 gene expression levels, which were deemed as features in this study. The six subsets (I-D0, I-D2-4, I-D7, II-D0, II-D1-4, and II-D7-10) were termed as labels. The current study was conducted by deeply investigating such classification problem.

2.2. Feature ranking algorithms

Lots of features were used to represent each sample. Evidently, some were important and others were useless. It was necessary to extract important features. To date, several feature analysis methods have been proposed, which can evaluate the importance of features. The selection of such method is a challenge problem as each method has its own merits and defects. Generally, one method can only output a part of essential features. Thus, it was beneficial to employ multiple methods, thereby providing a more complete picture on essential features. Here, five algorithms, namely, Lasso (Tibshirani, 2011), LightGBM (Ke et al., 2017), MCFS (Dramiński et al., 2007), mRMR (Peng et al., 2005), and PFI (Fisher et al., 2019), were employed to rank genes according to their importance. These algorithms have been frequently applied to solve many life science problems (Zhao et al., 2018; Ren et al., 2022; Li et al., 2022a,b,c; Huang et al., 2023a,b).

2.2.1. Least absolute shrinkage and selection operator

Based on the nonnegative garrote proposed by Breiman (1995), Robert Tibshirani first proposed the Lasso algorithm in 1996 (Tibshirani, 2011). The algorithm proposes a first-order penalty function containing regularized formulas, where each feature is regarded as an independent variable in the function. The coefficients of the features are then obtained by solving the optimization function. The absolute value of a coefficient indicates the degree of correlation of each feature to the target dependent variable. To achieve data compression and reduce overfitting, the algorithm regularizes the coefficients of some variables while setting some to zero to eliminate the features that tend to contribute less to the follow-up prediction. Accordingly, the algorithm can rank features according to the absolute values of their coefficients. In present study, the Lasso program in Scikit-learn (Pedregosa et al., 2011) was adopted, which was executed using default parameters.

2.2.2. Light gradient-boosting machine

LightGBM (Ke et al., 2017) is based on the gradient-boosting decision tree framework and introduces gradient one-sided sampling, exclusive feature bundling, histogram algorithm, and leaf-wise growth strategy. It enables data slicing, bundling, and dimensionality reduction and ultimately reduces computational cost while improving prediction accuracy. The importance of each feature is determined by the number of trees that the feature participates in building: the higher the participation, the higher the importance. Thus, features can be ranked in a list with decreasing order of this number. The current study used the LightGBM program obtained from.[1] For convenience, it was performed using default parameters.

2.2.3. Monte Carlo feature selection

Monte Carlo feature selection was originally developed by Dramiński et al. (2007). The algorithm selects some features randomly and repeatedly to obtain p feature subsets. Each feature subset is then divided into a training set and a test set t times, and t trees are constructed. Thus, $p \times t$ trees are obtained. The importance of features can be evaluated by their contributions to building these trees and is defined as the relative importance (RI) score, which is calculated as follows:

1 https://lightgbm.readthedocs.io/en/latest/

TABLE 1 Sample sizes of six vaccination status.

Index	Vaccination status		Sample size
1	I-D0	(Day 0 after the first dose)	37
2	I-D2-4	(Day2 2–4 after the first dose)	36
3	I-D7	(Day 7 after the first dose)	37
4	II-D0	(Day 0 after the second dose)	17
5	II-D1-4	(Days 1–4 after the second dose)	18
6	II-D7-10	(Days 7–10 after the second dose)	16

$$RI_g = \sum_{\tau=1}^{p\times t} \left(\omega A_{CC}\right)^u \sum_{ng(\tau)} IG\left(ng\left(\tau\right)\right)\left(\frac{no.in\ ng\left(\tau\right)}{no.in\ \tau}\right)^v, \quad (1)$$

where ωA_{CC} is the weighted precision of the tree τ under consideration, $ng\left(\tau\right)$ is a node of the tree whose information gain is denoted as $IG\left(ng\left(\tau\right)\right)$, and $no.in\ ng\left(\tau\right)\left(no.in\ \tau\right)$ denotes the sample size of $ng\left(\tau\right)\left(\tau\right)$. u and v are two positive numbers weighting the ωA_{CC} and the ratio $no.in\ ng\left(\tau\right)/no.in\ \tau$, respectively. To execute MCFS, we downloaded its program from.[2] Default parameters were used.

2.2.4. Max-relevance and Min-redundancy

The mRMR method was proposed by Peng et al. (2005) in 2005. It screens features based on their correlation with the target variable and the redundancy between features. The correlation and redundancy can be calculated from the mutual information between features or target variables. The tradeoff of correlation and redundancy is used to evaluate the importance of features. At each round, one feature with the maximum correlation to target variables and minimum redundancy to features in the current list is selected and appended to the current list. Here, we used the mRMR program sourced from.[3] It was executed with default parameters.

2.2.5. Permutation feature importance

The PFI for RFs was first introduced in 2001 by Breiman (2001) and was later extended to any fitted estimator for features by Fisher et al. (2019). The idea is relatively simple. If a feature is important, the prediction error will further increase after the feature's values are shuffled. If a feature is not important, shuffling its values does not increase the prediction error. The PFI program used in this study was retrieved from scikit-learn (Pedregosa et al., 2011), which was executed with default parameters.

Above five algorithms were applied to the blood transcriptome data one by one. Each algorithm produced one feature list. For easy descriptions, the generated lists were called Lasso, LightGBM, MCFS, mRMR and PFI feature lists.

2.3. Incremental feature selection

When the feature list contains an excessive number of features, it is not suitable for direct use in building prediction models. In this study, the IFS (Liu and Setiono, 1998) method was used to extract the best subset of features. From the feature list, a series of feature subsets can be constructed. Each subset includes 10 more features than the previous subset in the order of the list. These feature subsets were then fed to one classification algorithm to build the classifier. The performance of these classifiers was evaluated by 10-fold cross-validation. Lastly, the best classifier can be obtained, which was termed as the optimal classifier. The feature subset for constructing this classifier was called the optimal feature subset.

2.4. Synthetic minority oversampling technique

According to Table 1, some classes (e.g., I-D0) contained much more samples than other classes (e.g., II-D7-10). The dataset was imbalanced. The results of the classifier would have preferences for the majority class when the number of samples from different categories differs significantly. This study used synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) to balance the dataset. For each class with a small number of samples, a sample is random chosen. Then its k nearest neighbors in the same class are identified by Euclidean distance. A neighbor is randomly selected. A new sample is then randomly generated by linearly interpolating the randomly chosen sample and the selected nearest neighbor. New samples are continuously generated until such class contains samples as many as those in the largest class. The SMOTE package reported in[4] was used in this study. Default settings were adopted.

2.5. Classification algorithms for building classifiers

Four classification algorithms were used in the IFS approach. Key genes were then screened based on the performance of the constructed classifiers.

2 http://www.ipipan.eu/staff/m.draminski/mcfs.html

3 http://home.penglab.com/proj/mRMR/

4 https://github.com/scikit-learn-contrib/imbalanced-learn

2.5.1. Decision tree

The DT algorithm (Safavian and Landgrebe, 1991) constructs a tree-like structure in which instances are judged in each internal node of the tree. Starting from the root node, all samples are assigned to different classes through continuous judgments. Each tree branch contains clues to the classification of instances and thus provides interpretable classification rules that underlie the understanding of biological mechanisms. In this study, we used the CART classification tree algorithm with node ranking using the Gini coefficient.

2.5.2. Random forest

In the RF algorithm for classification, a judgment is completed by constructing DTs based on different training sets and then combining their results to make predictions (Breiman, 2001; Wang et al., 2021; Ran et al., 2022; Tang and Chen, 2022; Wu and Chen, 2023). The training set with the same number of samples in the input dataset is repeatedly sampled to generate numerous new training sets. Each new training set is then used to build a new DT, and an ensemble of DTs is constructed. Given a new instance, each DT makes a prediction. Predictions taken from all DTs are combined to reach a final decision.

2.5.3. K-nearest neighbor

In KNN (Cover and Hart, 1967), new samples are predicted by comparing each with samples with known labels (training samples) and determining the k-nearest neighbors. Subsequently, the class of a new sample is determined by voting according to the classes of the k-nearest neighbors. In this study, the distance was defined as the Minkowski distance.

2.5.4. Support vector machine

The SVM algorithm (Cortes and Vapnik, 1995; Wang and Chen, 2022; Wang and Chen, 2023) utilizes a kernel function that maps the attributes of the instances, i.e., the feature vectors, into a higher-dimensional space and attempts to find a separating hyperplane. This hyperplane partitions the instances by class and ensures that the margin between the two categories is maximum. This method is generally to have good generalization.

We adopted public packages in scikit-learn (Pedregosa et al., 2011) to implement above four classification algorithms. All packages were performed using default parameters.

2.6. Performance evaluation

In the multi-class classification problem, weighted F1 is an important measurement to evaluate the performance of the classifier. It is obtained by calculating and integrating the F1-measure values of different classes based on the proportion of the samples in each class. It is known that F1-measure is an integrated measurement combining precision and recall, which can be computed by

$$\mathrm{Pr}ecision_i = \frac{TP_i}{TP_i + FP_i}, \tag{2}$$

$$\mathrm{Re}call_i = \frac{TP_i}{TP_i + FN_i}, \tag{3}$$

$$F1-measure_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}, \tag{4}$$

where i represents the index of class, TP represents true positive, FP represents false positive, and FN represents false negative. Then, weighted F1 can be calculated by

$$Weighted\ F1 = \sum_{i=1}^{L} w_i \times F1-measure_i, \tag{5}$$

where L represents the number of classes and w_i represents the proportion of samples in the i-th class to overall samples. Here, weighted F1 was selected as the major measurement.

In addition, overall accuracy (ACC) and Matthew correlation coefficient (MCC) (Matthews, 1975) are also widely used to assess the quality of classifiers. ACC is defined as the proportion of correctly predicted samples to all samples. MCC is a balanced measurement, which is more objective than ACC when the dataset is imbalanced. For the calculation of MCC, two matrices X and Y must be constructed first, which store the one-hot representation of true and predicted class of each sample. Then, MCC can be computed by

$$MCC = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{cov}(X,X)\mathrm{cov}(Y,Y)}} \tag{6}$$

where $\mathrm{cov}(X,Y)$ denotes the correlation coefficient of X and Y.

2.7. Functional enrichment analysis

Using the IFS method, we can obtain the best subset of features under different rankings. To clarify the biological processes behind genes in these subsets, thereby uncovering their relationship with antiviral immunity, this study used gene ontology (GO) enrichment analysis to discover the role of the genes and applied Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis to identify the underlying pathways. ClusterProfiler package (Wu et al., 2021) in R was used to perform GO and KEGG enrichment analyses.

3. Results

3.1. Results of feature ranking

To evaluate the importance of features from multiple aspects. Five feature ranking algorithms were employed, which were applied to the blood transcriptome data one by one. As a result, five feature lists, named Lasso, LightGBM, MCFS, mRMR and PFI feature lists, were obtained, which are provided in Supplementary Table S1. Table 2 shows the top 10 genes in each list. It can be observed that top genes in different lists were very different, meaning that the

TABLE 2 The top 10 features in five feature lists.

Index	Lasso feature list	LightGBM feature list	MCFS feature list	mRMR feature list	PFI feature list
1	CENPF	RPRD1B	NEU3	FAM98B	SLC16A14
2	NDUFB9	ITM2C	C2	TSSK4	THRAP3
3	BRCA2	HSP90B1	SMC5	CSF1R	STAC3
4	LOC102031319	TK1	ZFC3H1	TOP1	ATF5
5	SSBP1	LPAR3	GLS2	NEU3	RAD51
6	PDP1	CENPF	NFE2L2	UBE2H	CDC45
7	LINC01089	TPX2	C1QC	ATP6V1E1	GABPB1
8	C2orf16	ITGAE	SDC1	SRPRB	CTNNBL1
9	ID2	SPATA24	CAV1	ZNF672	ARHGAP42
10	LINC00630	GTSE1	SNORA2B	CUL3	PSME2

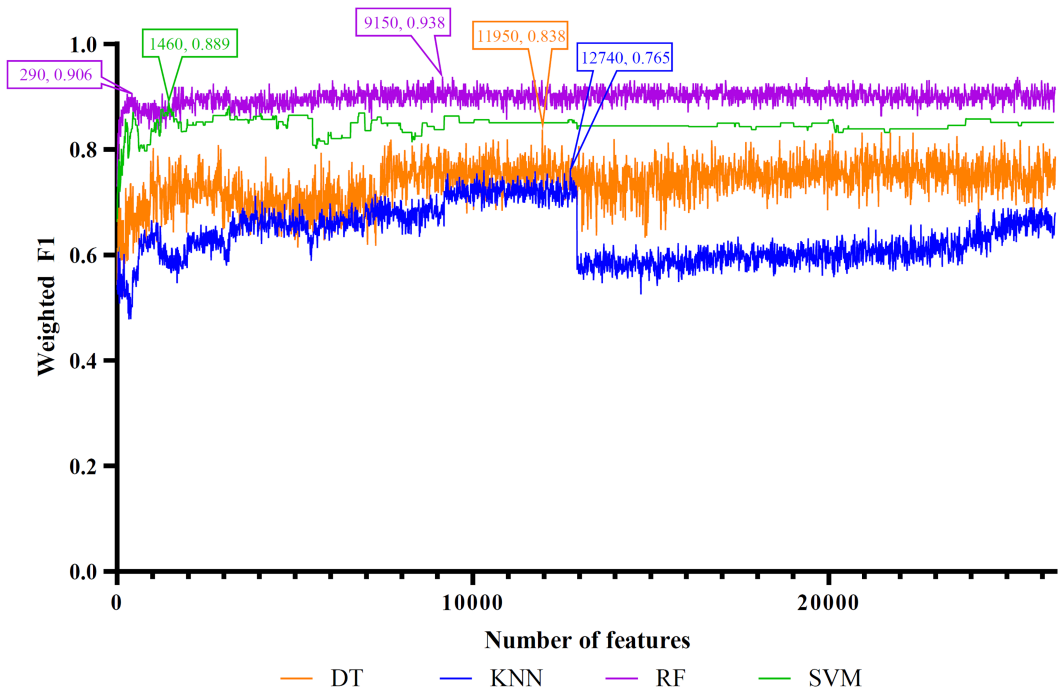


FIGURE 2
IFS curves of four classification algorithms on Lasso feature list. DT, KNN, RF, and SVM yielded the highest weighted F1 values of 0.838, 0.765, 0.938, and 0.889 when top 11,950, 12,740, 9,150, and 1,460 features were adopted, respectively. RF can yield quite high performance (weighted F1=0.906) when top 290 features were used.

importance of one feature was quite different under the evaluation of different methods. Usage of different methods can provide more opportunities to discover more essential features.

3.2. Results of incremental feature selection

Five feature lists were subjected to the IFS method one by one. From each feature list, a series of feature subsets with step ten were constructed. On each subset, one classifier was built for each of four classification algorithms (DT, KNN, RF, and SVM). When constructing the classifiers, the dataset was processed by SMOTE to tackle the imbalanced problem. All classifiers were evaluated by 10-fold cross-validation. The evaluation results were counted as weighted F1, ACC, and MCC, which are provided in Supplementary Table S2. Weighted F1was selected as the major measurement. Thus, several IFS curves were plotted for different classification algorithms and feature lists, as shown in Figures 2–6, in which weighted F1 was set as Y-axis and number of features was defined as *X*-axis.

For the Lasso feature list, the IFS curves of four classification algorithms are illustrated in Figure 2. It can be observed that when top 11,950, 12,740, 9,150 and 1,460 features were adopted,

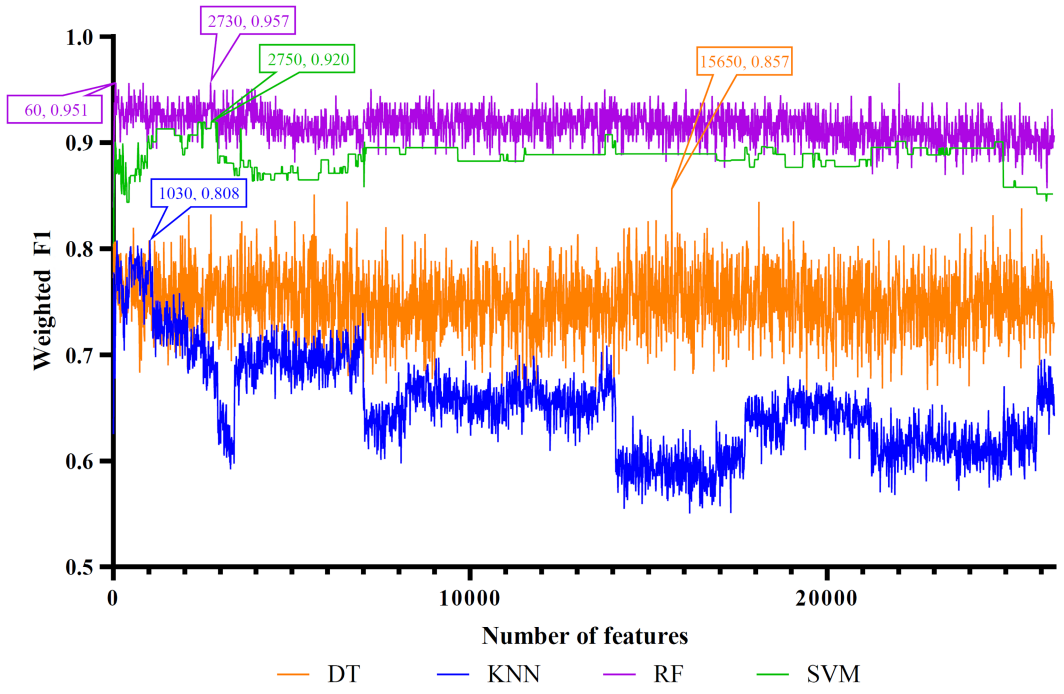


FIGURE 3
IFS curves of four classification algorithms on LightGBM feature list. DT, KNN, RF, and SVM yielded the highest weighted F1 values of 0.857, 0.808, 0.957, and 0.920 when top 15,650, 1,030, 2,730, and 2,750 features were adopted, respectively. RF can yield quite high performance (weighted F1=0.951) when top 60 features were used.

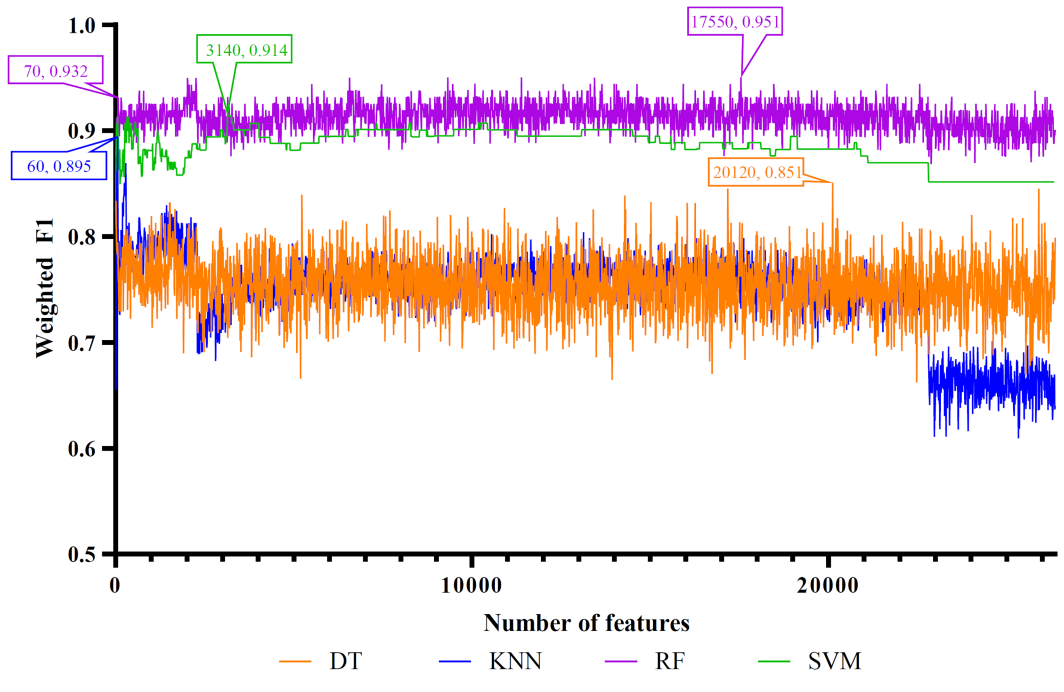


FIGURE 4
IFS curves of four classification algorithms on MCFS feature list. DT, KNN, RF and SVM yielded the highest weighted F1 values of 0.851, 0.895, 0.951, and 0.914 when top 20,120, 60, 17,550, and 3,140 features were adopted, respectively. RF can yield quite high performance (weighted F1=0.932) when top 70 features were used.

four algorithms yielded the highest weighted F1 values of 0.838, 0.765, 0.938, and 0.889, respectively. Thus, the optimal DT, KNN, RF, and SVM classifiers can be built using these features. The ACC and MCC values of these classifiers are listed in Table 3. Evidently, the optimal RF classifier was best among these optimal classifiers.

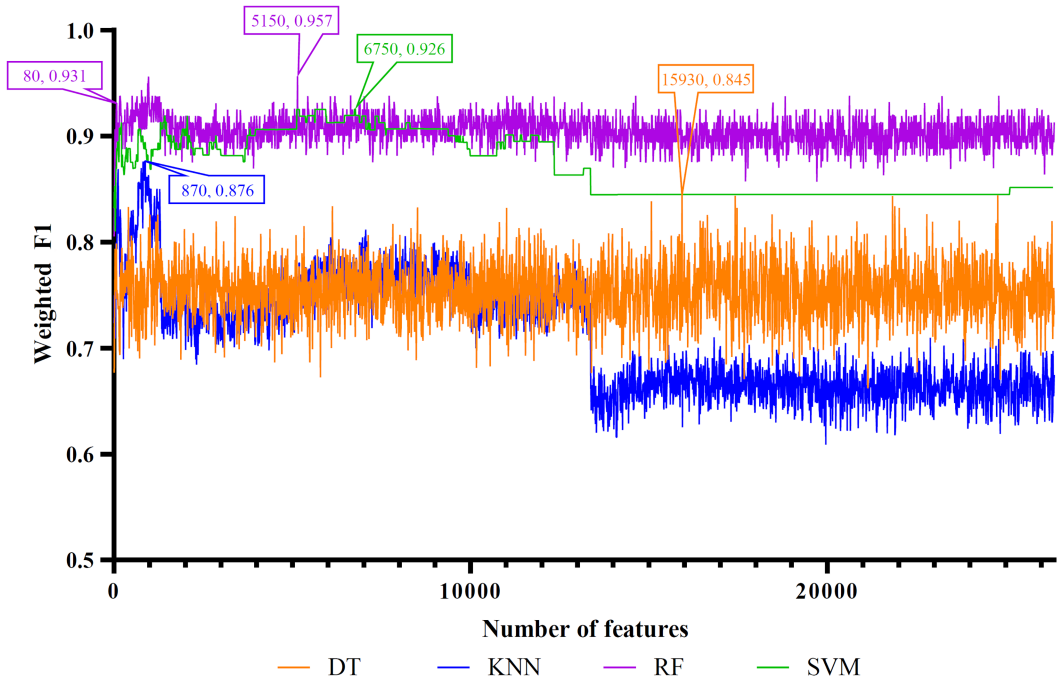


FIGURE 5
IFS curves of four classification algorithms on mRMR feature list. DT, KNN, RF, and SVM yielded the highest weighted F1 values of 0.845, 0.876, 0.957, and 0.926 when top 15,930, 870, 5,150, and 6,750 features were adopted, respectively. RF can yield quite high performance (weighted F1=0.931) when top 80 features were used.

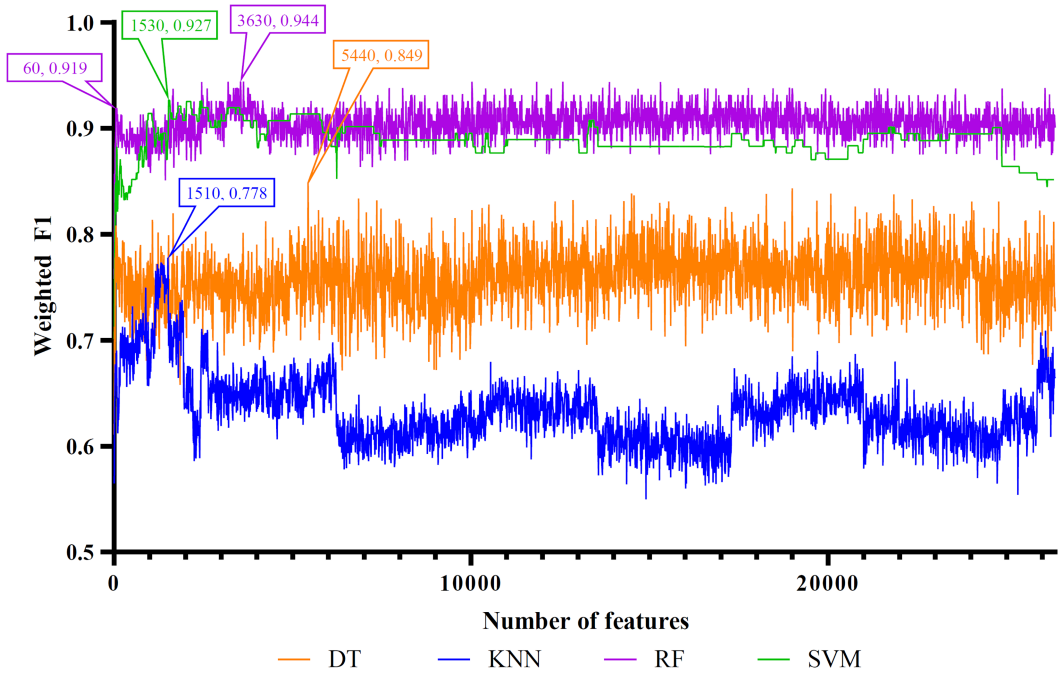


FIGURE 6
IFS curves of four classification algorithms on PFI feature list. DT, KNN, RF, and SVM yielded the highest weighted F1 values of 0.849, 0.778, 0.944, and 0.927 when top 5,440, 1,510, 3,630, and 1,530 features were adopted, respectively. RF can yield quite high performance (weighted F1=0.919) when top 60 features were used.

For the LightGBM feature list, Figure 3 shows the IFS curves of four classification algorithms. The optimal DT/KNN/RF/SVM classifier can be built using top 15,650/1030/2730/2750 features in this list. Their ACC, MCC, and weighted F1 values are listed in Table 3. Clearly, RF still provided the best performance as the optimal RF classifier yielded the highest weighted F1 of 0.957.

TABLE 3 Performance of the optimal classifiers based on different classification algorithms and feature lists.

Feature list	Classification algorithm	Number of features	Weighted F1	MCC	ACC
Lasso feature list	Decision tree	11,950	0.838	0.801	0.839
	K-nearest neighbor	12,740	0.765	0.722	0.770
	Random forest	9,150	0.938	0.924	0.938
	Support vector machine	1,460	0.889	0.863	0.888
LightGBM feature list	Decision tree	15,650	0.857	0.825	0.857
	K-nearest neighbor	1,030	0.808	0.764	0.807
	Random forest	2,730	0.957	0.947	0.957
	Support vector machine	2,750	0.920	0.901	0.919
MCFS feature list	Decision tree	20,120	0.851	0.817	0.851
	K-nearest neighbor	60	0.895	0.870	0.894
	Random forest	17,550	0.951	0.939	0.950
	Support vector machine	3,140	0.914	0.894	0.913
mRMR feature list	Decision tree	15,930	0.845	0.809	0.845
	K-nearest neighbor	870	0.876	0.847	0.876
	Random forest	5,150	0.957	0.947	0.957
	Support vector machine	6,750	0.926	0.908	0.925
PFI feature list	Decision tree	5,440	0.849	0.817	0.851
	K-nearest neighbor	1,510	0.778	0.734	0.783
	Random forest	3,630	0.944	0.932	0.944
	Support vector machine	1,530	0.927	0.909	0.925

TABLE 4 Performance of feasible classifiers on different feature list.

Feature list	Classification algorithm	Number of features	Weighted F1	MCC	ACC
Lasso feature list	Random forest	290	0.906	0.886	0.907
LightGBM feature list	Random forest	60	0.951	0.939	0.950
MCFS feature list	Random forest	70	0.932	0.916	0.932
mRMR feature list	Random forest	80	0.931	0.916	0.932
PFI feature list	Random forest	60	0.919	0.901	0.919

As for the rest three feature lists, the IFS curves are shown in Figures 4–6. The optimal DT/KNN/RF/SVM classifier can be set up on each feature list. The numbers of top features used in these classifiers are listed in Table 3, where the performance of these classifiers is also provided. Similar to the results on the Lasso and LightGBM feature lists, the optimal RF classifier was also better than other three optimal classifiers on each feature list.

To make full use of the utility of five algorithms, the best features should be extracted from each feature list, thereby obtaining the latent essential gene features. As mentioned above, the optimal RF classifier was best for each feature list. Thus, the features used in these classifiers can be picked up as important candidates. However, such feature numbers (9,150 for Lasso feature list, 2,730 for LightGBM feature list, 17,750 for MCFS feature list, 5,150 for mRMR feature list, 3,630 for PFI feature list) were too large to make detailed analyses. In view of this, we tried to find out another RF classifier, which adopted much less features and provided a little lower performance than the optimal RF classifier, on each feature list. By carefully checking the IFS results on RF on each feature list, such RF classifiers adopted the top 290 features in the Lasso feature list, top 60 features in the LightGBM feature list, top 70 features in the MCFS feature list, top 80 features in the mRMR feature list, and top 60 features in the PFI feature list. The corresponding points have been marked on the IFS curves of RF, as illustrated in Figures 2–6. The detailed performance of these RF classifiers is listed in Table 4. It can be observed that their performance was still quite high, the weighted F1 values were all higher than 0.900. Compared with the weighted F1 yielded by the optimal RF classifier on the same feature list, this RF classifier provided a little lower weighted F1. However, their efficiencies were sharply improved because much less features were involved. This indicated the extreme importance of features used in these RF classifiers. For easy descriptions, these RF classifiers were called feasible RF classifiers. Furthermore, the performance of the feasible RF classifier on one feature list was generally better than the optimal DT/KNN/SVM classifier on the same feature list, further confirming the importance of features in the feasible RF classifiers. To clear show the relationship between the feature sets used in five feasible RF classifiers, a Venn diagram was plotted, as shown in Figure 7. The detailed results of the intersection are shown in Supplementary Table S3.

10.3389/fmicb.2023.1138674

Some gene features occurred in multiple subsets, meaning that they were deemed to be important by multiple feature ranking algorithms. They may have strong associations with antiviral immunity. Some of them would be discussed in detail in the subsequent sections.

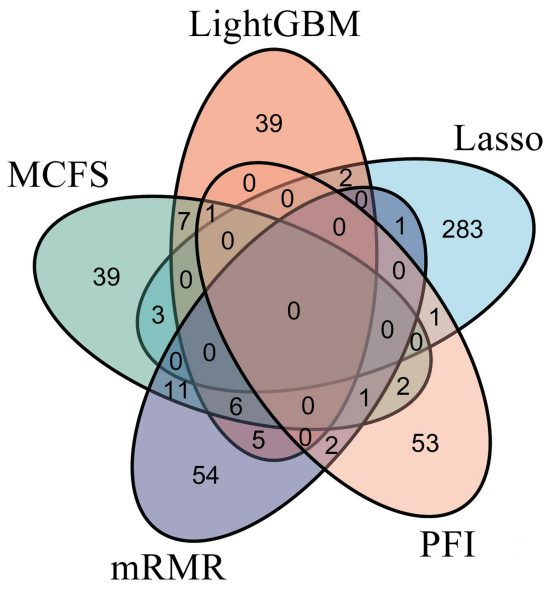


FIGURE 7
Venn diagram of the feature sets used to construct feasible random forest classifiers on five feature lists that were obtained by Lasso, LightGBM, MCFS, mRMR, and PFI, respectively. The overlapping circles indicated genes that occurred in multiple sets. These genes were deemed to be important by multiple feature ranking algorithms.

3.3. Classification rules

Although the performance of DT was much lower than RF and SVM according to the IFS results on five feature lists, DT has an exclusive merit as it is a white-box algorithm. It can provide quantitative rules that can be interpreted to aid in the analysis. On the Lasso, LightGBM, MCFS, mRMR, and RF feature lists, the optimal DT classifier adopted the first 11,950, 15,650, 20,120, 15,930, and 5,440 gene features. Based on the samples represented by these features, five trees were obtained, from which five groups of classification rules can be extracted. Supplementary Table S4 shows these classification rule groups. Some conditions in major rules would be discussed in detail later.

3.4. Enrichment analysis

Five feature sets used to construct five feasible RF classifiers were combined into one set. To uncover the underlying biological meanings behind gene features in such set, the enrichment analysis was conducted on these genes. Figure 8 visualizes top five GO terms in three GO clusters and top five pathways. The GO terms, such as thioester and fatty acid metabolic processes, were enriched, along with peroxisomes and some terms related to metabolism and transport. KEGG enriched pathways included fatty acid biosynthesis, catabolism, and metabolism. Thioesters can be directly involved in the immune response as carriers of antigen presentation and thioesterified fatty acids or other lipid products can be involved in the regulation of immune cells as signaling molecules. Their metabolism is inseparable from the peroxisome.

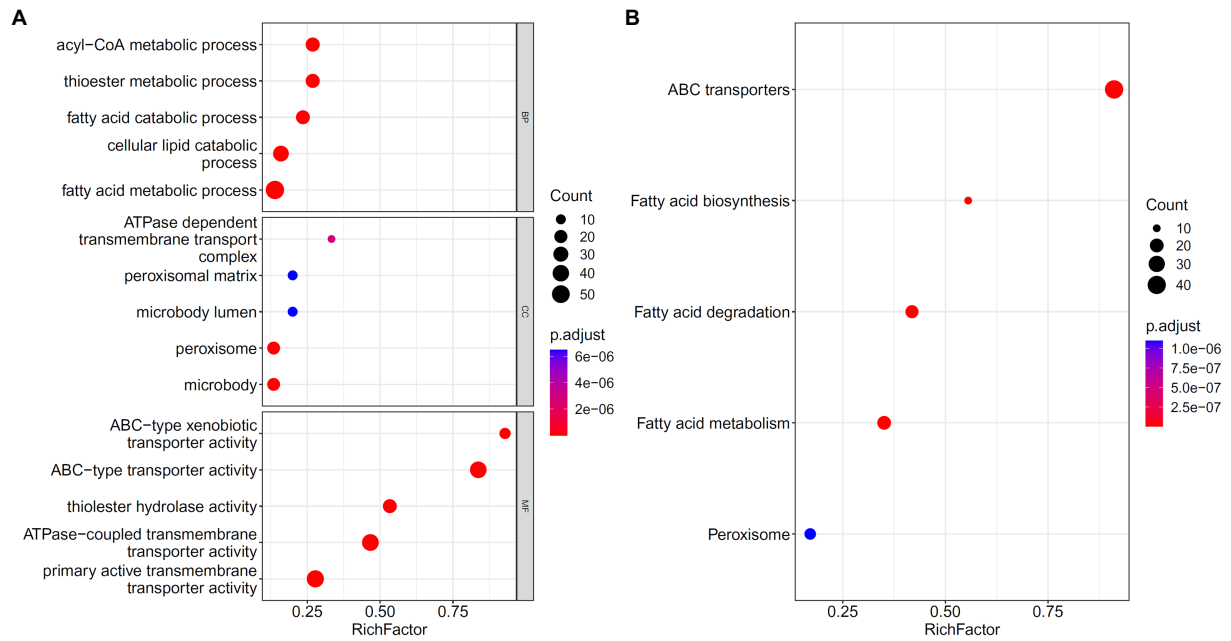


FIGURE 8
Gene ontology (GO) and KEGG pathway enrichment analysis on the union of five feature sets used to construct feasible random forest classifiers. The FDR<0.05 criterion was used to filter GO terms and KEGG pathways. The top five significant GO terms in three GO clusters **(A)** and top five KEGG pathways **(B)** were shown.

4. Discussion

As listed in "Results", some essential genes and classification rules were discovered. As they can be strongly related to the response to vaccination in antitumor viral immunity, they were discussed in this section. We collected the scientific findings of other researchers and initially summarized the experimental evidence of the aforementioned genes and rules, proving the accuracy of the findings.

4.1. Analysis of essential conditions in rules

Five rule groups were discovered as listed in Supplementary Table S4. As each rule contained multiple gene features and thresholds on expression levels, it was not easy to confirm the special pattern expressed by each rule through existing publications. Thus, we divided each rule into multiple conditions and analyzed the reasonability of some essential conditions. If the conditions used the same gene and same expression trend, they were deemed to be identical. The occurrence number of each condition in five rule groups was counted, which represented how many feature ranking methods identified the condition to be important. Some representative conditions with such numbers larger than two were discussed.

4.1.1. Analysis of conditions identified *via* four methods

IFI27 occurred in four rule groups, including rule groups on Lasso, LightGBM, mRMR, and MCFS feature lists. The study found that the expression levels of antiviral-related genes such as *IFI27* decreased during the vaccinations. This result is consistent with the dynamically enhanced inflammatory response in vaccinated individuals. *IFI27* is considered a biomarker with high sensitivity and specificity (AUC > 0.85) (Wang et al., 2022). Vaccination can improve the body's ability to fight viruses. Our analysis results show that the expression level of *IFI27* gradually increased within 2–4 days of the first injection and decreased 7 days after vaccination. However, after the second injection, the expression level of *IFI27* gradually increased within 1–4 days after the injection. Compared with the first injection, some patients had the fastest response times earlier than the first injection. The expression level of *IFI27* decreased 7–10 days after vaccination. The peak duration of the second injection is speculated to be longer than that of the first injection. The antiviral immune-related molecular mechanism of IFI27 has been reported. As a common interferon (IFN)-stimulated gene, *IFI27* encodes a mitochondrial protein that is normally induced by IFN to express and function in most responding cells. It may regulate apoptosis through the stability of mitochondrial membrane, thereby affecting immune response (Cheriyath et al., 2011). In addition, IFI27 can inhibit viral DNA replication and gene expression (Ullah et al., 2021). *In vitro* studies have shown that IFI27 is up-regulated in plasmacytoid dendritic cells, which are antigen-presenting cells sensitive to viral infection (Tang et al., 2017). Transcriptome results showed that vaccinated patients had significantly attenuated IFN responses compared to unvaccinated Omicron and Alpha-infected patients, represented by IFI27, which controls antiviral responses (Lee et al., 2022b). The results of RNA sequencing data analysis showed that macrophages in the blood of SARS-CoV-2-infected patients released a large number of IFNs, activated mitochondrial *IFI27* expression, and disrupted energy metabolism in immune cells, ultimately aggravating viral immune evasion and replication (Duan et al., 2022). Based on existing research reports and our analysis, we speculate that after vaccination, the release of IFN increases, which promotes an increase in mitochondrial protein IFI27, inhibits SARS-CoV-2 replication and gene expression, and enhances antiviral immunity. In addition, after two vaccine doses, some people's antiviral immunity takes effect earlier than after the first dose, and vaccine efficacy lasts longer. Therefore, *IFI27* may be used as a biomarker for antiviral immunity of vaccines.

4.1.2. Analysis of conditions identified *via* three methods

Syndecan-1 (*SDC1*) and *small nuclear ribonucleoprotein polypeptide G* (*SNRPG*) were found in rule groups on LightGBM, mRMR, and MCFS feature lists. *SDC1* encodes a transmembrane (type I) heparan sulfate proteoglycan protein that belongs to the syndecan proteoglycan family. As a component of glycocalyx (GAC), SDC1 plays an important role in cell proliferation, cell migration, and other processes through extracellular matrix protein receptors (Reszegi et al., 2022). SDC1 was found to be elevated in COVID-19 patients (Goonewardena et al., 2021). SDC1 may contribute to early risk stratification of staged diseases such as COVID-19 and provide a pathobiological reference (Goonewardena et al., 2021). Studies have confirmed that patients infected with COVID-19 can produce inflammation-induced degradation of the GAC layer of endothelial cells, and SDC1 can be used as an important parameter to assess GAC damage (Vollenberg et al., 2021). High levels of SDC1 may cause more severe endothelial damage and inflammation (Zhang et al., 2021). Molecular experiments demonstrate that *SDC1* acts as a target gene of miR-10a-5p during porcine hemagglutinating encephalomyelitis virus (PHEV) infection and is involved in host defense mechanisms. Decreased expression levels of *SDC1* lead to reduced viral replication, and downstream inhibition of *SDC1* exerts an antiviral effect in PHEV-induced disease (Hu et al., 2020). Transcriptome analysis showed that the expression level of *SDC1* increased only 7 days after the first dose of vaccination. After the second dose, the expression level remained low. On the one hand, this low level may help prevent endothelial damage and severe inflammatory response. On the other hand, it may inhibit viral replication and facilitate a more efficient antibody production.

SNRPG is a protein-coding gene involved in the formation of the U1, U2, U4, and U5 small nuclear ribonucleoprotein complexes. Related pathways include SARS-CoV-2 infection and gene expression.[5] Studies have shown that SNRPG-related risk models are associated with infiltration of immune cells such as T cells and M2 macrophages (Liu et al., 2022). The specific mechanism between *SNRPG* and SARS-CoV-2 infection is limited. Transcriptome analysis showed that the *SNRPG* expression level was high on the day of the first vaccine injection, whereas the expression level was lower on the day of the second vaccine injection. The low *SNRPG* level continued until day 10 after vaccination. The obvious differences in *SNRPG* levels after different injections suggest that the gene can be regarded as an indicator of the effectiveness of vaccination. However, the molecular mechanism needs to be further explored.

5 https://pathcards.genecards.org/Card/sars-cov-2_infection?queryString=SNRPG

4.1.3. Analysis of conditions identified *via* two methods

Rules found in two methods included *TPX2, CCDC28A, FAM227B, PKN2-AS1, NEK2, USP46, C22orf15, SLC20A1, TMSB15A, C2,* and *ZFC3H1*. Some of these genes are associated with antiviral immunity. For example, *TPX2* (microtubule nucleation factor) is a gene whose encoded product is involved in the activation of protein kinase activity, DNA damage, gene transcription, and other physiological processes. PPI network analysis from STRING revealed that as a hub gene, *TPX2* may be a novel COVID-19 intervention target and biomarker (Hasan et al., 2022). As one of the antigen components of a multivalent recombinant fusion protein prophylactic vaccine (rBmHAXT), *TPX2* can promote the production of high titers of antigen-specific antibodies and their isotypes. Animals vaccinated with the *TPX2* antigen secreted higher levels of blood IFN-γ and showed better immune protection compared with unvaccinated animals (Khatri et al., 2018). Studies have shown that *TPX2* can activate Aurora A kinase (AURKA), which is involved in cell cycle regulation. *TPX2* overexpression enhanced cell proliferation and migration (Zou et al., 2018). The *TPX2* gene may be a potential target for diagnosis and prognosis in patients already infected with hepatitis B virus (HVB) (Ji et al., 2020). Transcriptome data analysis showed that *TPX2* expression levels increased within 7–10 days after the patients received the second vaccine dose. This is consistent with activation of IFN-induced responses, increased transcripts of specific IGHV clones, and a trend toward memory B cell enrichment (Lee et al., 2022a). *TPX2* may be related to antiviral immunity caused by different doses. However, the correlation and mechanism of action need to be further verified.

4.2. Top features identified *via* multiple methods

On the basis of the features identified by the five feature ranking algorithms (Figure 7), an intersection of results obtained by multiple methods (≥3) was selected as important candidates. We summarized the evidence for some vital gene features, listed in Table 5, based on the broad studies shown below.

NFE2-like bZip transcription factor 2 (*NRF2*), also called *NFE2L2*, encodes a cap'n'collar (CNC) transcription factor and belongs to the small family of basic leucine zipper (bZIP) proteins (Khan et al., 2021). *NRF2* can bind to antioxidant response elements and participate in the transcription of downstream target genes. Thus, it plays an important role in physiological processes such as cellular redox, tissue damage, and metabolic homeostasis. The encoded protein of *NRF2* is involved in various injury and inflammatory responses involving class I MHC-mediated antigen presentation and KEAP1-NFE2L2 pathway, among others. *NRF2* contributes to GSH metabolism and stress response and is associated with the pro-inflammatory effects of SARS-CoV-2 in host cells (Galli et al., 2022). The protein synthesis of SARS-CoV-2 may increase Cys and activate endoplasmic reticulum stress of transcription factors, which ultimately promotes changes in cellular oxidation, cellular metabolism, and GSH transmembrane flux (Galli et al., 2022). Importantly, *NRF2* activation has been shown to benefit respiratory infections in various animal models (Muchtaridi et al., 2022). *NRF2* exerts anti-inflammatory effects by inhibiting pro-inflammatory genes such as *IL6* and *IL1B* (Huang et al., 2022). *NRF2* induces the expression of genes that promote specificity of macrophages such as the macrophage receptor, which is responsible for bacterial phagocytosis (Schaefer et al., 2022), and the cluster of differentiation gene 36 (CD36), which resists viral infection (Hillier et al., 2022). *NRF2* Activation is involved in inflammatory cascade (Jayakumar et al., 2022), regulation of innate immune responses, and antiviral cytosolic DNA sensing. *NRF2* inhibits pro-inflammatory signaling pathways such as TNF-α signaling and is involved in regulating the innate immune response during sepsis. *NRF2* increases susceptibility to DNA virus infection by inhibiting the expression of the adaptor protein STING1, thereby inhibiting antiviral cytosolic DNA sensing (Olagnier et al., 2018). After SARS-CoV-2 infection, *NRF2* is activated and restricts the release of pro-inflammatory cytokines by inhibiting IRF3 dimerization. In addition, *NRF2* inhibits the replication of SARS-CoV-2 and other viruses through a type I IFN-independent pathway (Olagnier et al., 2020).

Regulation of nuclear pre-mRNA domain containing 1B (*RPRD1B*), also named cell-cycle-related and expression-elevated protein in tumor (*CREPT*) or *C20ORF77*, is located on chromosome 20q11 and can bind to RNA polymerase on the cyclin D1 gene, resulting in the formation of a cyclin D1 ring structure, which can promote transcription (Lu et al., 2012; Wang et al., 2014). *RPRD1B* can also participate in the transcription of genes related to the Wnt/β-catenin signaling pathway (Wu et al., 2010). GO annotation results showed that *RPRD1B* can bind to the RNA polymerase II complex and play a role in pathways such as TCR signaling and T-cell activation. The mRNA and protein expression of *RPRD1*B in patients under 50 years old were significantly different from those in patients over 50 years of age. *RPRD1B* expression levels correlate with human papillomavirus infection and may be affected by age (Wen et al., 2021). The expression level of *RPRD1B* in peripheral blood T cells of psoriasis, lichen planus (LP), and atopic dermatitis (AD) was found higher than that of healthy subjects. *RPRD1B* is involved in the pathogenesis of inflammatory diseases by regulating the transcription of genes such as *IL-4*, *RGS16*, and *CD30* (Li et al., 2013). Our analysis showed that the *RPRD1B* expression level changed in patients who received different vaccinations. Combined with existing evidence, we speculate that *RPRD1B* uses T cells as a carrier to play a role in antiviral immunity.

Neuraminidase 3 (*NEU3*) is a protein-encoding gene whose product is located in the plasma membrane and belongs to the glycohydrolase family. Its activity is specific to gangliosides and may be involved in gangliosides in lipid bilayer adjustment. Pathways associated with *NEU3* include protein metabolism and glycosphingolipid metabolism. It can directly interact with signaling receptors such as EGFR to regulate transmembrane signaling (Wada et al., 2007; Mozzi et al., 2015). Sialidase activity in human polymorphonuclear leukocytes plays a key role in infection and inflammatory responses (Cross et al., 2003; Sakarya et al., 2004). Sialidase activity is determined by membrane-associated sialidase (*NEU3*), which promotes cell adhesion and cell proliferation.

TABLE 5 Essential genes identified by three feature ranking algorithms.

Index	Gene symbol	Description
1	RPRD1B	Regulation of nuclear pre-mRNA domain containing 1B
2	NFE2L2	NFE2-like bZip transcription factor 2
3	SMC5	Structural maintenance of chromosome 5
4	NEU3	Neuraminidase 3

Combined with existing evidence, our results indicate that after vaccination, the body produces antibodies against SARS-CoV-2 that regulate the host immune response by affecting the activity of *NEU3*.

The encoded product of structural maintenance of chromosome 5 (*SMC5*) has ATP-binding activity and is involved in physiological processes such as DNA recombination, cellular senescence, protein metabolism, and transport of mature mRNAs. In addition, SMC*5* can bind to SMC6, participate in the repair of DNA double-strand breaks through homologous recombination, and prevent the transcription of free DNA such as circular virus DNA genomes (Decorsière et al., 2016). Proteomic analysis revealed that Epstein–Barr virus infection disrupts the adhesion proteins SMC5/6, thereby affecting DNA damage repair. In the absence of the involucrin protein BNRF1, SMC5/6 interferes with the formation and encapsidation of viral replication compartments (RCs), ultimately affecting viral lytic replication. SMC5/6 may act as intrinsic immunosensors and restriction factors of human herpes virus RC in viral infectious diseases (Yiu et al., 2022). The SMC5/6 complex compresses viral chromatin to silence gene expression; thus, its depletion enhances viral expression. The SMC5/6 complex also functions in immunosurveillance of extrachromosomal DNA (Dupont et al., 2021). As an intrinsic antiviral restriction factor, Smc5/6, when localized to nuclear domain 10 (ND10) in primary human hepatocytes, inhibits HBV transcription without inducing an innate immune response (Niu et al., 2017). We screened SMC5 signatures in populations vaccinated with different doses. The results suggest that SMC5 may serve as an indicator of vaccine effectiveness.

5. Conclusion

The purpose of this study was to analyze the blood transcriptome in response to different numbers and timing of vaccinations through a variety of machine learning algorithms. It also aimed to identify antiviral immunity-related molecules in different vaccinated populations. The feature intersection of multiple analysis methods reflects the effects of different vaccinations on host gene expression. The analysis results showed that the key gene features were highly consistent with existing research conclusions, which helped us to further clarify the possible mechanisms of these genes. The important antiviral immune characteristics obtained in this study will help in understanding the differences in mechanisms of action of different vaccinations and provide a reference for targeted COVID-19 intervention and for optimization of vaccine strategies.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE201533.

Author contributions

TH and Y-DC designed the study. JL, WG, and KF performed the experiments. JR and HL analyzed the results. JL, JR, and HL wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

Funding

This research was supported by the National Key R&D Program of China [2022YFF1203202], Strategic Priority Research Program of Chinese Academy of Sciences [XDA26040304, XDB38050200], the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences [202002], and Shandong Provincial Natural Science Foundation [ZR2022MC072].

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1138674/full#supplementary-material

SUPPLEMENTARY TABLE S1
Feature lists obtained using Lasso, LightGBM, MCFS, mRMR and PFI.

SUPPLEMENTARY TABLE S2
Performance of IFS with different classification algorithms.

SUPPLEMENTARY TABLE S3
Intersection of feature sets used to construct feasible random forest classifiers on five feature lists.

SUPPLEMENTARY TABLE S4
Classification rules generated by the optimal DT classifier on different feature lists.

References

Amano, M., Otsu, S., Maeda, K., Uemura, Y., Shimizu, Y., Omata, K., et al. (2022). Neutralization activity of sera/IgG preparations from fully BNT162b2 vaccinated individuals against SARS-CoV-2 alpha, Beta, gamma, Delta, and kappa variants. *Sci. Rep.* 12:13524. doi: 10.1038/s41598-022-17071-9

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* 37, 373–384. doi: 10.1080/00401706.1995.10484371

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

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. doi: 10.1613/jair.953

Cheriyath, V., Leaman, D. W., and Borden, E. C. (2011). Emerging roles of FAM14 family members (G1P3/ISG 6-16 and ISG12/IFI27) in innate immunity and cancer. *J. Interf. Cytokine Res.* 31, 173–181. doi: 10.1089/jir.2010.0105

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

Cross, A. S., Sakarya, S., Rifat, S., Held, T. K., Drysdale, B. E., Grange, P. A., et al. (2003). Recruitment of murine neutrophils in vivo through endogenous sialidase activity. *J. Biol. Chem.* 278, 4112–4120. doi: 10.1074/jbc.M207591200

Dai, Q., Bao, C., Hai, Y., Ma, S., Zhou, T., Wang, C., et al. (2018). MTGIpick allows robust identification of genomic islands from a single genome. *Brief. Bioinform.* 19, 361–373. doi: 10.1093/bib/bbw118

Decorsière, A., Mueller, H., Van Breugel, P. C., Abdul, F., Gerossier, L., Beran, R. K., et al. (2016). Hepatitis B virus X protein identifies the Smc5/6 complex as a host restriction factor. *Nature* 531, 386–389. doi: 10.1038/nature17170

Dramiński, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2007). Monte Carlo feature selection for supervised classification. *Bioinformatics* 24, 110–117. doi: 10.1093/bioinformatics/btm486

Duan, C., Ma, R., Zeng, X., Chen, B., Hou, D., Liu, R., et al. (2022). SARS-CoV-2 achieves immune escape by destroying mitochondrial quality: comprehensive analysis of the cellular landscapes of lung and blood specimens from patients with COVID-19. *Front. Immunol.* 13:946731. doi: 10.3389/fimmu.2022.946731

Dupont, L., Bloor, S., Williamson, J. C., Cuesta, S. M., Shah, R., Teixeira-Silva, A., et al. (2021). The SMC5/6 complex compacts and silences unintegrated HIV-1 DNA and is antagonized by Vpr. *Cell Host Microbe* 29, 792–805.e6. doi: 10.1016/j.chom.2021.03.001

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: learning a Variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* 20, 1–81.

Folegatti, P. M., Ewer, K. J., Aley, P. K., Angus, B., Becker, S., Belij-Rammerstorfer, S., et al. (2020). Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. *Lancet* 396, 467–478. doi: 10.1016/S0140-6736(20)31604-4

Galli, F. A.-O. X., Marcantonini, G., Giustarini, D. A.-O., Albertini, M. A.-O., Migni, A., Zatini, L., et al. (2022). How aging and oxidative stress influence the cytopathic and inflammatory effects of SARS-CoV-2 infection: the role of cellular glutathione and cysteine metabolism. *Antioxidants (Basel)* 11:1366. doi: 10.3390/antiox11071366

Goonewardena, S. N., Grushko, O. G., Wells, J., Herty, L., Rosenson, R. S., Haus, J. M., et al. (2021). Immune-mediated Glycocalyx remodeling in hospitalized COVID-19 patients. *Cardiovasc. Drugs Ther.*, 1–7. doi: 10.1007/s10557-021-07288-7

Hasan, M. I., Rahman, M. H., Islam, M. B., Islam, M. Z., Hossain, M. A., and Moni, M. A. (2022). Systems biology and bioinformatics approach to identify blood based signatures molecules and drug targets of patient with COVID-19. *Inform. Med. Unlocked* 28:100840. doi: 10.1016/j.imu.2021.100840

Hillier, J. A.-O., Allcott, G. J., Guest, L. A., Heaselgrave, W., Tonks, A. A.-O., Conway, M. A.-O., et al. (2022). The BCAT1 CXXC motif provides protection against ROS in acute myeloid Leukaemia cells. *Antioxidants (Basel)* 11:683. doi: 10.3390/antiox11040683

Hu, S., Li, Z., Lan, Y., Guan, J., Zhao, K., Chu, D., et al. (2020). MiR-10a-5p-mediated Syndecan 1 suppression restricts porcine Hemagglutinating encephalomyelitis virus replication. *Front. Microbiol.* 11:105. doi: 10.3389/fmicb.2020.00105

Huang, F., Fu, M., Li, J., Chen, L., Feng, K., Huang, T., et al. (2023a). Analysis and prediction of protein stability based on interaction network, gene ontology, and KEGG pathway enrichment scores. *BBA-Proteins and Proteomics* 1871:140889. doi: 10.1016/j.bbapap.2023.140889

Huang, F., Ma, Q., Ren, J., Li, J., Wang, F., Huang, T., et al. (2023b). Identification of smoking associated transcriptome aberration in blood with machine learning methods. *Biol. Med. Res. Int.* 2023:5333361. doi: 10.1155/2023/5333361

Huang, J., Zhang, Z., Hao, C., Qiu, Y., Tan, R., Liu, J., et al. (2022). Identifying drug-induced liver injury associated with inflammation-drug and drug-drug interactions in pharmacologic treatments for COVID-19 by bioinformatics and system biology analyses: the role of Pregnane X receptor. *Front. Pharmacol.* 13:804189. doi: 10.3389/fphar.2022.804189

Jayakumar, T., Huang, C. J., Yen, T. A.-O., Hsia, C. A.-O., Sheu, J. A.-O., Bhavan, P. A.-O., et al. (2022). Activation of Nrf2 by Esculetin mitigates inflammatory responses through suppression of NF-κB signaling Cascade in RAW 264.7 cells. *Molecules* 27:5143. doi: 10.3390/molecules27165143

Ji, Y., Yin, Y., and Zhang, W. (2020). Integrated Bioinformatic analysis identifies networks and promising biomarkers for hepatitis B virus-related hepatocellular carcinoma. *Int. J. Genom.* 2020, 1–18. doi: 10.1155/2020/2061024

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree, in Proceedings of the 31st International Conference on Neural Information Processing Systems. (Long Beach, California, USA: Curran Associates Inc.).

Khan, H., Patel, S., and Majumdar, A. (2021). Role of NRF2 and Sirtuin activators in COVID-19. *Clin. Immunol.* 233:108879. doi: 10.1016/j.clim.2021.108879

Khatri, V., Chauhan, N., Vishnoi, K., Von Gegerfelt, A., Gittens, C., and Kalyanasundaram, R. (2018). Prospects of developing a prophylactic vaccine against human lymphatic filariasis - evaluation of protection in non-human primates. *Int. J. Parasitol.* 48, 773–783. doi: 10.1016/j.ijpara.2018.04.002

Kong, R., Xu, X., Liu, X., He, P., Zhang, M. Q., and Dai, Q. (2020). 2SigFinder: the combined use of small-scale and large-scale statistical testing for genomic island detection from a single genome. *BMC Bioinform.* 21:159. doi: 10.1186/s12859-020-3501-2

Lee, H. K., Go, J., Sung, H., Kim, S. W., Walter, M., Knabl, L., et al. (2022a). Heterologous ChAdOx1-BNT162b2 vaccination in Korean cohort induces robust immune and antibody responses that includes omicron. *iScience* 25:104473 doi: 10.1016/j.isci.2022.104473

Lee, H. K., Knabl, L., Walter, M., Knabl, L. Sr., Dai, Y., Fussl, M., et al. (2022b). Prior vaccination exceeds prior infection in eliciting innate and humoral immune responses in omicron infected outpatients. *Front. Immunol.* 13:916686. doi: 10.3389/fimmu.2022.916686

Li, Z., Guo, W., Ding, S., Chen, L., Feng, K., Huang, T., et al. (2022b). Identifying key MicroRNA signatures for neurodegenerative diseases with machine learning methods. *Front. Genet.* 13:880997. doi: 10.3389/fgene.2022.880997

Li, H., Huang, F., Liao, H., Li, Z., Feng, K., Huang, T., et al. (2022a). Identification of COVID-19-specific immune markers using a machine learning method. *Front. Mol. Biosci.* 9:952626. doi: 10.3389/fmolb.2022.952626

Li, X., Li, J., Yang, Y., Hou, R., Liu, R., Zhao, X., et al. (2013). Differential gene expression in peripheral blood T cells from patients with psoriasis, lichen planus, and atopic dermatitis. *J. Am. Acad. Dermatol.* 69, e235–e243. doi: 10.1016/j.jaad.2013.06.030

Li, Z., Mei, Z., Ding, S., Chen, L., Li, H., Feng, K., et al. (2022c). Identifying methylation signatures and rules for COVID-19 with machine learning methods. *Front. Mol. Biosci.* 9:908080. doi: 10.3389/fmolb.2022.908080

Liu, J., Gu, L., Zhang, D., and Li, W. (2022). Determining the prognostic value of spliceosome-related genes in hepatocellular carcinoma patients. *Front. Mol. Biosci.* 9:759792. doi: 10.3389/fmolb.2022.759792

Liu, H. A., and Setiono, R. (1998). Incremental feature selection. *Appl. Intell.* 9, 217–230. doi: 10.1023/A:1008363719778

Lu, D., Wu, Y., Wang, Y., Ren, F., Wang, D., Su, F., et al. (2012). CREPT accelerates tumorigenesis by regulating the transcription of cell-cycle-related genes. *Cancer Cell* 21, 92–104. doi: 10.1016/j.ccr.2011.12.016

Masic, I., Naser, N., and Zildzic, M. (2020). Public health aspects of COVID-19 infection with focus on cardiovascular diseases. *Mast. Sociomed.* 32, 71–76. doi: 10.5455/msm.2020.32.71-76

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-protein. *Structure* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

Mizrahi, B., Lotan, R., Kalkstein, N., Peretz, A., Perez, G., Ben-Tov, A., et al. (2021). Correlation of SARS-CoV-2-breakthrough infections to time-from-vaccine. *Nat. Commun.* 12:6379. doi: 10.1038/s41467-021-26672-3

Mozzi, A., Forcella, M., Riva, A., Difrancesco, C., Molinari, F., Martin, V., et al. (2015). NEU3 activity enhances EGFR activation without affecting EGFR expression and acts on its sialylation levels. *Glycobiology* 25, 855–868. doi: 10.1093/glycob/cwv026

Muchtaridi, M. A.-O., Amirah, S. R., Harmonis, J. A., and Ikram, E. A.-O. (2022). Role of nuclear factor erythroid 2 (Nrf2) in the recovery of Long COVID-19 using natural antioxidants: A systematic review. *Antioxidants (Basel)* 11:1551. doi: 10.3390/antiox11081551

Niu, C., Livingston, C. M., Li, L., Beran, R. K., Daffis, S., Ramakrishnan, D., et al. (2017). The Smc5/6 complex restricts HBV when localized to ND10 without inducing an innate immune response and is counteracted by the HBV X protein shortly after infection. *PLoS One* 12:e0169648. doi: 10.1371/journal.pone.0169648

Olagnier, D., Brandtoft, A. A.-O., Gunderstofte, C., Villadsen, N. L., Krapp, C., Thielke, A. L., et al. (2018). Nrf2 negatively regulates STING indicating a link between antiviral sensing and metabolic reprogramming. *Nat. Commun.* 9:3506. doi: 10.1038/s41467-018-05861-7

Olagnier, D. A.-O., Farahani, E., Thyrsted, J., Blay-Cadanet, J., Herengt, A., Idorn, M. A.-O., et al. (2020). SARS-CoV2-mediated suppression of NRF2-signaling reveals potent antiviral and anti-inflammatory activity of 4-octyl-itaconate and dimethyl fumarate. *Nat. Commun.* 11:4938. doi: 10.1038/s41467-020-18764-3

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.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi: 10.1109/TPAMI.2005.159

Pozzetto, B., Legros, V., Djebali, S., Barateau, V., Guibert, N., Villard, M., et al. (2021). Immunogenicity and efficacy of heterologous ChAdOx1-BNT162b2 vaccination. *Nature* 600, 701–706. doi: 10.1038/s41586-021-04120-y

Ran, B., Chen, L., Li, M., Han, Y., and Dai, Q. (2022). Drug-drug interactions prediction using fingerprint only. *Comput. Math. Methods Med.* 2022, 1–14. doi: 10.1155/2022/7818480

Ren, J., Zhou, X., Guo, W., Feng, K., Huang, T., and Vcai, Y.-D. (2022). Identification of methylation signatures and rules for sarcoma subtypes by machine learning methods. *Biomed. Res. Int.* 2022, 1–11. doi: 10.1155/2022/5297235

Reszegi, A., Tatrai, P., Regos, E., Kovalszky, I., and Baghy, K. (2022). Syndecan-1 in liver pathophysiology. *Am. J. Physiol. Cell Physiol.* 323, C289–C294. doi: 10.1152/ajpcell.00039.2022

Safavian, S. R., and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* 21, 660–674. doi: 10.1109/21.97458

Sakarya, S., Rifat, S., Zhou, J., Bannerman, D. D., Stamatos, N. M., Cross, A. S., et al. (2004). Mobilization of neutrophil sialidase activity desialylates the pulmonary vascular endothelial surface and increases resting neutrophil adhesion to and migration across the endothelium. *Glycobiology* 14, 481–494. doi: 10.1093/glycob/cwh065

Schaefer, R. E. M., Callahan, R. C., Atif, S. M., Orlicky, D. J., Cartwright, I. M., Fontenot, A. P., et al. (2022). Disruption of monocyte-macrophage differentiation and trafficking by a heme analog during active inflammation. *Mucosal Immunol.* 15, 244–256. doi: 10.1038/s41385-021-00474-8

Tang, S., and Chen, L. (2022). iATC-NFMLP: identifying classes of anatomical therapeutic chemicals based on drug networks, fingerprints and multilayer perceptron. *Curr. Bioinforma.* 17, 814–824. doi: 10.2174/1574893617666220318093000

Tang, B. M., Shojaei, M., Parnell, G. P., Huang, S., Nalos, M., et al. (2017). A novel immune biomarker IFI27 discriminates between influenza and bacteria in patients with suspected respiratory infection. *Eur. Respir. J.* 49:1602098. doi: 10.1183/13993003.02098-2016

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *J. R. Stat. Soc.* 73, 273–282. doi: 10.1111/j.1467-9868.2011.00771.x

Ullah, H., Sajid, M., Yan, K., Feng, J., He, M., Shereen, M. A., et al. (2021). Antiviral activity of interferon alpha-inducible protein 27 against hepatitis B virus gene expression and replication. *Front. Microbiol.* 12:656353. doi: 10.3389/fmicb.2021.656353

Vollenberg, R. A.-O., Tepasse, P. A.-O., Ochs, K. A.-O., Floer, M., Strauss, M., Rennebaum, F., et al. (2021). Indications of persistent Glycocalyx damage in convalescent COVID-19 patients: A prospective multicenter Study and hypothesis. *Viruses* 13:2324. doi: 10.3390/v13112324

Wada, T., Hata, K., Yamaguchi, K., Shiozaki, K., Koseki, K., Moriya, S., et al. (2007). A crucial role of plasma membrane-associated sialidase in the survival of human cancer cells. *Oncogene* 26, 2483–2490. doi: 10.1038/sj.onc.1210341

Wang, R., and Chen, L. (2022). Identification of human protein subcellular location with multiple networks. *Curr. Proteom.* 19, 344–356. doi: 10.2174/1570164619666220531113704

Wang, H., and Chen, L. (2023). PMPTCE-HNEA: Predicting metabolic pathway types of chemicals and enzymes with a heterogeneous network embedding algorithm. *Curr. Bioinform.* doi: 10.2174/1574893618666230224121633

Wang, Y., Li, J., Zhang, L., Sun, H. X., Zhang, Z., Xu, J., et al. (2022). Plasma cell-free RNA characteristics in COVID-19 patients. *Genome Res.* 32, 228–241. doi: 10.1101/gr.276175.121

Wang, Y., Qiu, H., Hu, W., Li, S., and Yu, J. (2014). RPRD1B promotes tumor growth by accelerating the cell cycle in endometrial cancer. *Oncol. Rep.* 31, 1389–1395. doi: 10.3892/or.2014.2990

Wang, Y., Xu, Y., Yang, Z., Liu, X., and Dai, Q. (2021). Using recursive feature selection with random Forest to improve protein structural class prediction for low-similarity sequences. *Comput. Math. Methods Med.* 2021, 1–9. doi: 10.1155/2021/5529389

Wen, N., Bian, L., Gon, J., and Meng, Y.-A. (2021). RPRD1B is a potentially molecular target for diagnosis and prevention of human papillomavirus E6/E7 infection-induced cervical cancer: A case-control study. *Asia Pac. J. Clin. Oncol.* 17, 230–237. doi: 10.1111/ajco.13439

Wu, C., and Chen, L. (2023). A model with deep analysis on a large drug network for drug classification. *Math. Biosci. Eng.* 20, 383–401. doi: 10.3934/mbe.2023018

Wu, Y., Fau-Yang, X., Fau-Wang, Y., Fau-Ren, F., Fau-Liu, H., Fau-Zhai, Y., et al. (2010). p15RS attenuates Wnt/β-catenin signaling by disrupting {beta}-catenin·TCF4 interaction. *J. Biol. Chem.* 285, 34621–34631. doi: 10.1074/jbc.M110.148791

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovations* 2:100141. doi: 10.1016/j.xinn.2021.100141

Yang, S., Wang, Y., Chen, Y., and Dai, Q. (2020). MASQC: next generation sequencing assists third generation sequencing for quality control in N6-Methyladenine DNA identification. *Front. Genet.* 11:269. doi: 10.3389/fgene.2020.00269

Yang, Z., Yi, W., Tao, J., Liu, X., Zhang, M.Q., Chen, G., et al. (2022). HPVMD-C: a disease-based mutation database of human papillomavirus in China. Database (Oxford) 2022, baac018. doi: 10.1093/database/baac018

Yiu, S. P. T., Guo, R., Zerbe, C., Weekes, M. P., and Gewurz, B. E. (2022). Epstein-Barr virus BNRF1 destabilizes SMC5/6 cohesin complexes to evade its restriction of replication compartments. *Cell Rep.* 38:110411. doi: 10.1016/j.celrep.2022.110411

Zhang, D., Li, L., Chen, Y., Ma, J., Yang, Y., Aodeng, S., et al. (2021). Syndecan-1, an indicator of endothelial glycocalyx degradation, predicts outcome of patients admitted to an ICU with COVID-19. *Mol. Med.* 27:151. doi: 10.1186/s10020-021-00412-1

Zhao, X., Chen, L., and Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. *Math. Biosci.* 306, 136–144. doi: 10.1016/j.mbs.2018.09.010

Zou, J., Huang, R. Y., Jiang, F. N., Chen, D. X., Wang, C., Han, Z. D., et al. (2018). Overexpression of TPX2 is associated with progression and prognosis of prostate cancer. *Oncol. Lett.* 16, 2823–2832. doi: 10.3892/ol.2018.9016

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 21 March 2023
DOI 10.3389/fmicb.2023.1100873

OPEN ACCESS

EDITED BY
Taoyang Wu,
University of East Anglia,
United Kingdom

REVIEWED BY
Wei-Hua Chen,
Huazhong University of Science and Technology,
China
Shi Huang,
The University of Hong Kong,
Hong Kong SAR,
China
Gongchao Jing,
Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences (CAS),
China

*CORRESPONDENCE
Hui Liu
✉ hliu@xzhmu.edu.cn
Dongshen Ma
✉ madongshen89@163.com

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

RECEIVED 17 November 2022
ACCEPTED 28 February 2023
PUBLISHED 21 March 2023

CITATION
Xu Y, Zhao J, Ma Y, Liu J, Cui Y, Yuan Y, Xiang C, Ma D and Liu H (2023) The microbiome types of colorectal tissue are potentially associated with the prognosis of patients with colorectal cancer. *Front. Microbiol.* 14:1100873. doi: 10.3389/fmicb.2023.1100873

The microbiome types of colorectal tissue are potentially associated with the prognosis of patients with colorectal cancer

Yixin Xu[1], Jing Zhao[2], Yu Ma[3], Jia Liu[2], Yingying Cui[2], Yuqing Yuan[2], Chenxi Xiang[2], Dongshen Ma[2]* and Hui Liu[2,3]*

[1]Department of General Surgery, The Affiliated Hospital of Xuzhou Medical University, Xuzhou, Jiangsu, China, [2]Department of Pathology, The Affiliated Hospital of Xuzhou Medical University, Xuzhou, Jiangsu, China, [3]Department of Pathology, Xuzhou Medical University, Xuzhou, Jiangsu, China

As the second leading cause of cancer worldwide, colorectal cancer (CRC) is associated with a poor prognosis. Although recent studies have explored prognostic markers in patients with CRC, whether tissue microbes carry prognostic information remains unknown. Here, by assessing the colorectal tissue microbes of 533 CRC patients, we found that Proteobacteria (43.5%), Firmicutes (25.3%), and Actinobacteria (23.0%) dominated the colorectal tissue microbiota, which was different from the gut microbiota. Moreover, two clear clusters were obtained by clustering based on the tissue microbes across all samples. By comparison, the relative abundances of Proteobacteria and Bacteroidetes in cluster 1 were significantly higher than those in cluster 2; while compared with cluster 1, Firmicutes and Actinobacteria were more abundant in cluster 2. In addition, the Firmicutes/Bacteroidetes ratios in cluster 1 were significantly lower than those in cluster 2. Further, compared with cluster 2, patients in cluster 1 had relatively poor survival (Log-rank test, p=0.0067). By correlating tissue microbes with patient survival, we found that the relative abundance of dominant phyla, including Proteobacteria, Firmicutes, and Bacteroidetes, was significantly associated with survival in CRC patients. Besides, the co-occurrence network of tissue microbes at the phylum level of cluster 2 was more complicated than that of cluster 1. Lastly, we detected some pathogenic bacteria enriched in cluster 1 that promote the development of CRC, thus leading to poor survival. In contrast, cluster 2 showed significant increases in the abundance of some probiotics and genera that resist cancer development. Altogether, this study provides the first evidence that the tissue microbiome of CRC patients carries prognostic information and can help design approaches for clinically evaluating the survival of CRC patients.

KEYWORDS

colorectal cancer, tissue microbe, prognostic biomarkers, survival, pathogenic bacteria

1. Introduction

The incidence and mortality of colorectal cancer (CRC) have increased significantly in recent years, ranking the 3rd and 5th among all malignant tumors, respectively (Siegel et al., 2020; Zhao et al., 2020; Lu et al., 2021). Most patients are in the middle and late stages when diagnosed, which seriously threatens the survival and quality of life of patients (Dekker et al., 2019; Cienfuegos-Jimenez et al., 2021; Peng et al., 2022). The 5-year relative survival ranges from

more than 90% in stage I patients to slightly more than 10% in stage IV patients (Brenner et al., 2014; Biswas et al., 2021). Due to the frequent recurrence and metastasis, the prognosis of CRC is yet to be improved, especially for those with unknown tissue origin (He et al., 2020a,b; Liu et al., 2021). Accurate prediction of the prognosis of CRC patients is of great significance for targeted treatment and avoidance of overtreatment. However, at present, most studies are focused on identifying biomarkers for early screening of CRC (Ahlquist et al., 2000; Tanaka et al., 2020; Wu et al., 2021), and the exploration of biomarkers for patient prognosis is still limited, except for a few initial tries (Yang et al., 2022; Yuan et al., 2022).

Microbial communities are thought to influence the initiation, progression, metastasis, and response to the treatment of a variety of cancers (Cullin et al., 2021; Qi et al., 2022; Wang et al., 2022). In addition to gut microbes, microbes in other niches may influence host physiology. Many members of the microbial community can induce cell proliferation by activating certain signaling pathways. Microbial communities can act as a source of activating signals for aberrant epithelial cell proliferation, initiating cancer (Fulbright et al., 2017). This includes microbes on the outer surface and mucosal sites, as well as tissue-resident microbes (Heymann et al., 2021). Castellarin et al. (2012) found that *Fusobacterium nucleatum* transcripts were 400 times more abundant in CRC tumor tissues than in normal tissues. In addition, *F. nucleatum* has been associated with liver metastases (Bullman et al., 2017), amplifying its potential impact on cancer. *Bacteroides fragilis* is a commensal bacteria active in the whole colon, among which enterotoxigenic *B. fragilis* (ETBF) is believed to be associated with the induction of colitis and colon tumorigenesis due to its enrichment in stool and mucosal samples of cancer patients (Boleij et al., 2015; Haghi et al., 2019). Besides, healthy gut microbes are typically made up of dominant populations of *Lactobacilli*, *Bacteroides*, and *Bifidobacterium* (Nakatsu et al., 2015). In CRC, *Fusobacterium*, *Porphyromonas*, *Parvimonas*, *Peptostreptococcus*, and *Gemella* showed excessive dominance, indicating the occurrence of bacterial flora imbalance (Nakatsu et al., 2015; Wirbel et al., 2019; Cheng et al., 2022). However, there is no consensus that one or more microbes can be associated with the prognosis of CRC patients, whether it is intestinal flora or intratumoral microbes of tumor tissue. Consequently, there is an urgent need to study the association between microbial communities and the prognosis of patients with malignant tumors.

Enterotype is a new concept proposed by Arumugam et al. (2011) in the study of intestinal microbiota in 2011. Arumugam et al. (2011) found that the gut microbiota can be divided into three groups according to the dominant genera, with *Bacteroides*, *Prevotella*, and *Bifidobacteria* as the dominant types. Different enterotypes have different microbiota structures and functional genes, and people with different enterotypes have different ways of energy metabolism and storage. In recent years, more and more studies have shown that a large number of microbes are enriched in tumor tissues (Hu et al., 2017; Nejman et al., 2020; Wong-Rolle et al., 2021). Therefore, we wonder whether the colorectal tumor tissue microbiota of CRC patients can be classified similarly to the gut microbiota and whether this classification carries prognostic information of CRC patients, such as the propensity for recurrence and metastasis as well as survival time.

To this end, we collected colorectal microbiological samples from 533 CRC patients at The Cancer Genome Atlas (TCGA). By characterizing the microbial diversity of all samples, we found that the Shannon index of 533 samples showed bimodal distribution. Therefore, based on the clustering of tissue microbiota from all CRC patients, we obtained colorectal tissue microbiota typing. Further, we correlated tissue microbiota typing with prognosis in CRC patients and found that increased relative abundance of certain microbes was significantly associated with worse or better prognosis. This study provides new insights into inferences about the prognosis of CRC patients based on the composition of the dominant bacteria in the tissue microbiota.

2. Materials and methods

2.1. Data collection and preparation

A total of 533 tissue microbiome samples of CRC patients and the corresponding metadata were obtained in this study. Cancer microbiome data and the clinical metadata data used in this study were available at ftp://ftp.microbio.me/pub/cancer_microbiome_analysis/ (Poore et al., 2020). The microbial abundance matrix in the data set was annotated by two methods, Kraken and Shotgun. Given Kraken's high usage rate in metagenomic analysis, only the microbial abundance obtained from Kraken's annotation was used in this study. Microbiome data included six levels of microbial count including kingdom, phylum, class, order, family, and genus. We calculated the relative abundance of microbes at each level for subsequent analysis.

2.2. Clustering analysis

Based on the tissue microbiome abundance matrix, all samples were clustered using the "partitioning around medoids" (PAM) clustering method. Clustering was conducted with package "cluster" in R. Different from K-means clustering based on means, PAM is based on more robust partitioning around central points. In this study, we obtained five groups based on the microbial community at the phylum level by PAM clustering. To reduce the complexity and improve the rationality of the analysis, we further combined these five groups into two groups with significant differences in tissue microbes.

2.3. Survival analysis and dimension reduction

The overall survival between different groups was compared by Kaplan–Meier (KM) analysis, and the p value was generated with the log-rank test. In this study, we divided all samples equally into two groups (High vs. Low) based on the relative abundances of Proteobacteria, Actinobacteria, Firmicutes, and Bacteroidetes, respectively. Then, survival analysis was conducted on these two groups. Principal component analysis (PCA) was performed with packages "FactoMineR" and "factoextra" in R. The R^2 and p value were calculated by an ANOSIM test. Univariate cox regression was performed by the R package "survminer."

2.4. Linear discriminant analysis effect size analysis

Linear discriminant analysis effect size (LEfSe) (Segata et al., 2011), an analytical tool for discovering and interpreting high-dimensional data biometrics (genes, pathways, taxons, etc.) was used to determine the significantly different genera in relative abundance between the two clusters. LEfSe used linear discriminant analysis (LDA) to estimate the magnitude of the effect of the abundance of each component (species) on the differential effect. In this study, we identified 11 potential biomarkers at the genus level with an LDA score > 4 and $p < 0.05$.

2.5. Network analysis

We mapped the co-occurrence network of tissue microbiota in two groups of colorectal cancer patients. Correlation coefficients and p value between the microbes at the phylum level were generated by the R function "rcorr" in the "Hmisc" package. Further, the network was visualized by Gephi (Bastian et al., 2009), a software tool for building and visualizing bibliometric networks. Only the correlation p-values less than 0.01 were shown in the network. The network graph showed only edges with correlation coefficients greater than 0.2 and less than −0.2. Nodes in the network diagram represent microbes, and edges represent correlations between microbes. Node size indicates the relative abundance of microbes. The microbes whose names are shown in the network diagram are the important ones in the network, namely the nodes with a high degree.

3. Results and discussion

3.1. Tissue microbe profiles of colorectal cancer patients

Colorectal tissue has a different microbiota profile than the gut. Proteobacteria was the phylum with the highest relative abundance in CRC patient tissues with an average relative abundance of 43.5%, followed by Firmicutes, Actinobacteria, and Bacteroidetes, with the relative abundance of 25.3, 23.0, and 5.1%, respectively (Figure 1A). Similarly, the dominant flora in the gut is mainly composed of Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria, accounting for more than 97% of the intestinal flora (Eckburg et al., 2005). However, different from tissue microorganisms, the dominant phyla of gut microbiota are Firmicutes and Bacteroidetes, with only a small proportion of other phyla (Stopinska et al., 2021). An increase in Proteobacteria in the gut is considered a microbial marker of dysregulation of the gut microbiota and a potential diagnostic feature of disease risk (Shin et al., 2015). We detected a high abundance of Proteobacteria in the tissues of CRC patients, which also represents the deterioration of colorectal tumors in patients.

Next, to explore the microbial diversity of CRC patients' tissues, we calculated the Shannon index of all samples. Notably, the distribution of microbial diversity was bimodal (Figure 1B), with a smaller peak at 2.3 and a larger peak at 3.7. Further, we created a clustering heatmap based on the abundance matrix of phylum-level microbes for all samples (Figure 1C). Similarly, all samples could be clustered into two main groups based on phylum-level microbes across the samples. Preliminarily, we found that the abundances of the four dominant phyla (Proteobacteria, Firmicutes, Actinobacteria, and Bacteroidetes) in the tissues showed differences between the two groups. The large differences in the abundance of dominant phyla led us to wonder whether the tissue microbiota of CRC patients is classified as similar to the enterotype of gut microbiota.

3.2. CRC patients can be divided into two clusters based on tissue microbiome, and the prognosis of the two clusters is significantly different

We next investigated whether the tissue microbiome abundance reflected the same bimodal distribution as observed for the Shannon index. For this, we used a clustering method called "partitioning around medoids" (PAM) for the abundance of the four dominant phyla with the highest relative abundance. The clustering results showed that all samples were divided into five groups with silhouette widths of 0.58, 0.46, 0.50, 0.53, and 0.44, respectively (Figures 2A,B). We further verified the clustering quality with silhouette width, and the result showed that the silhouette width was the highest (0.53) with $k = 5$, suggesting that was the optimal number of clusters (Figure 2C). These two components explain 79.26% of the point variability. Besides, considering the bimodal distribution presented by the Shannon index of all samples (Figure 1B) and the clear two groups presented by clustering heatmap (Figure 1C), we further combined these five groups into two clusters according to the patient survival. Finally, we obtained two clusters of the five groups, with significant differences ($p = 0.0067$) in survival between the two clusters (Supplementary Table S1). PCA showed that the relative abundance of the four dominant phyla of the two clusters was significantly different (Figure 2D; ANOSIM, $p = 0.001$, $R^2 = 0.63$). Besides, consistent with the bimodal distribution (Figure 1B), the Shannon index of cluster 1 was significantly higher than that of cluster 2 (Supplementary Figure S1). In-depth, we compared the differences of single species between the two clusters separately. Results showed that the relative abundance of Proteobacteria, Actinobacteria, Firmicutes, and Bacteroidetes were significantly differences between cluster 1 and cluster 2 (Figures 2E–H, Wilcoxon test, $p < 4.2e\text{-}12$). Specifically, the relative abundance of Proteobacteria and Bacteroidetes in cluster 1 was significantly higher than that in cluster 2, while Actinobacteria and firmicutes were significantly enriched in cluster 2 compared with cluster 1. Besides, the Firmicutes/Bacteroidetes (F/B) ratios of cluster 2 were significantly higher than that of cluster 1 (Figure 2I). The low F/B ratio in the gut is usually considered a biomarker of obesity in humans and animals (Magne et al., 2020). Studies have found reduced F/B ratios in the gut in patients with a variety of diseases, including Alzheimer's disease, cholelithiasis, and rheumatoid arthritis (Grigor'eva, 2020; Artacho et al., 2021; Sheng et al., 2021). Consequently, we hypothesized that the reduced F/B ratio in colorectal tissues of CRC patients in cluster 1 may affect the tumorigenesis process and thus change the prognosis.

Next, we investigated whether there were differences in prognosis, such as survival, among CRC patients in the two

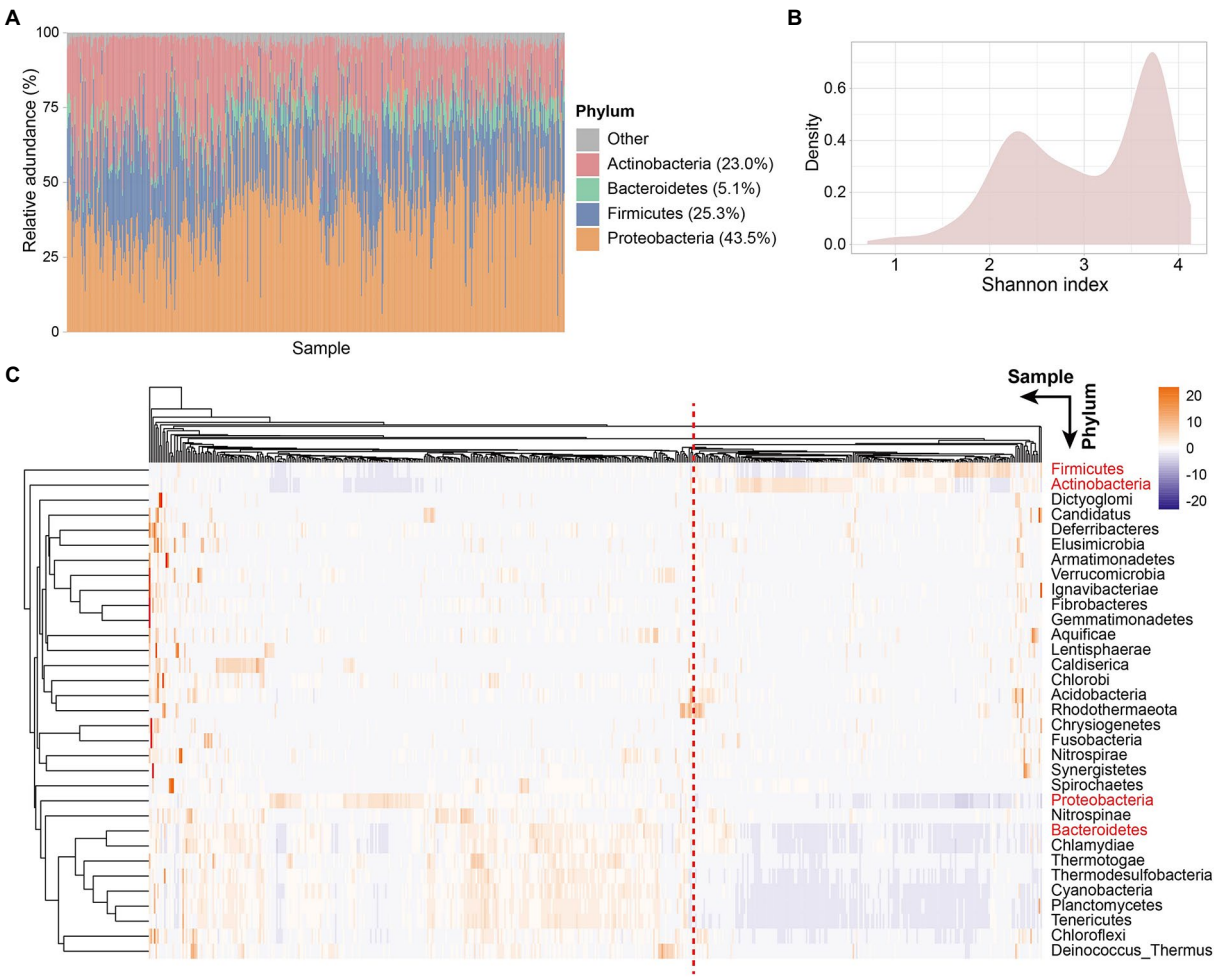


FIGURE 1
Tissue microbe profiles of CRC patients. **(A)** Tissue microbial community composition at phylum level across all samples. Different color represents different phyla. The four phyla with the highest relative abundance are shown in the figure. Each column represents a sample. **(B)** Density plot of Shannon index of all samples. **(C)** Clustering heatmap based on the relative abundance of 33 species at the phylum level in all samples. Rows represent species and columns represent samples. The names of the four phyla with the highest relative abundance are shown in red. Samples separated by red dashed lines differed in relative abundance at the phylum level species.

clusters. Since these two clusters were obtained based on the four dominant phyla with the highest relative abundance, we further compared the overall microbial communities of these two clusters. The results showed that the overall tissue microbial communities of cluster 1 and cluster 2 were also significantly different (Figure 2J; ANOSIM, $R^2 = 0.65$, $p = 0.001$). Then, the survival analysis of patients in these two clusters showed that compared with cluster 2, patients in cluster 1 had significantly worse survival (Figure 2K, $p = 0.0067$). Besides, to verify the computational stability of our results, we randomly selected 50% of the samples and repeated PAM clustering and survival analysis (Supplementary Figure S2). Repeated analysis based on a 50% sample size confirmed the consistency of the results. A significant difference in survival between the two groups could still be found even when the sample size was reduced.

Our results demonstrate that tissue microbiota in CRC patients potentially influences tumor development and that tissue microbiota characteristics carry patient prognostic information.

3.3. Microbes with significantly different abundance are responsible for the differentiation of prognosis between the two groups

Previously, we found that colorectal microbiota may affect the prognosis of CRC patients. Next, we focused on which microbiota plays a role in tumor progression. For this, all samples were equally divided into two groups (High and low) according to the relative abundance of the four dominant phyla (Proteobacteria, Actinobacteria, Firmicutes, and Bacteroidetes). Then, we performed survival curves for the two groups, respectively, and compared them (Figures 3A–D). Survival analysis showed that patients with a high abundance of Proteobacteria in colorectal tissue had significantly worse survival (Figure 3A, $p = 0.0025$). In contrast, patients with a high abundance of Firmicutes had significantly improved survival compared with patients with fewer Firmicutes in colorectal tissue (Figure 3C, $p = 0.035$). Similar to Proteobacteria, patients with more abundant

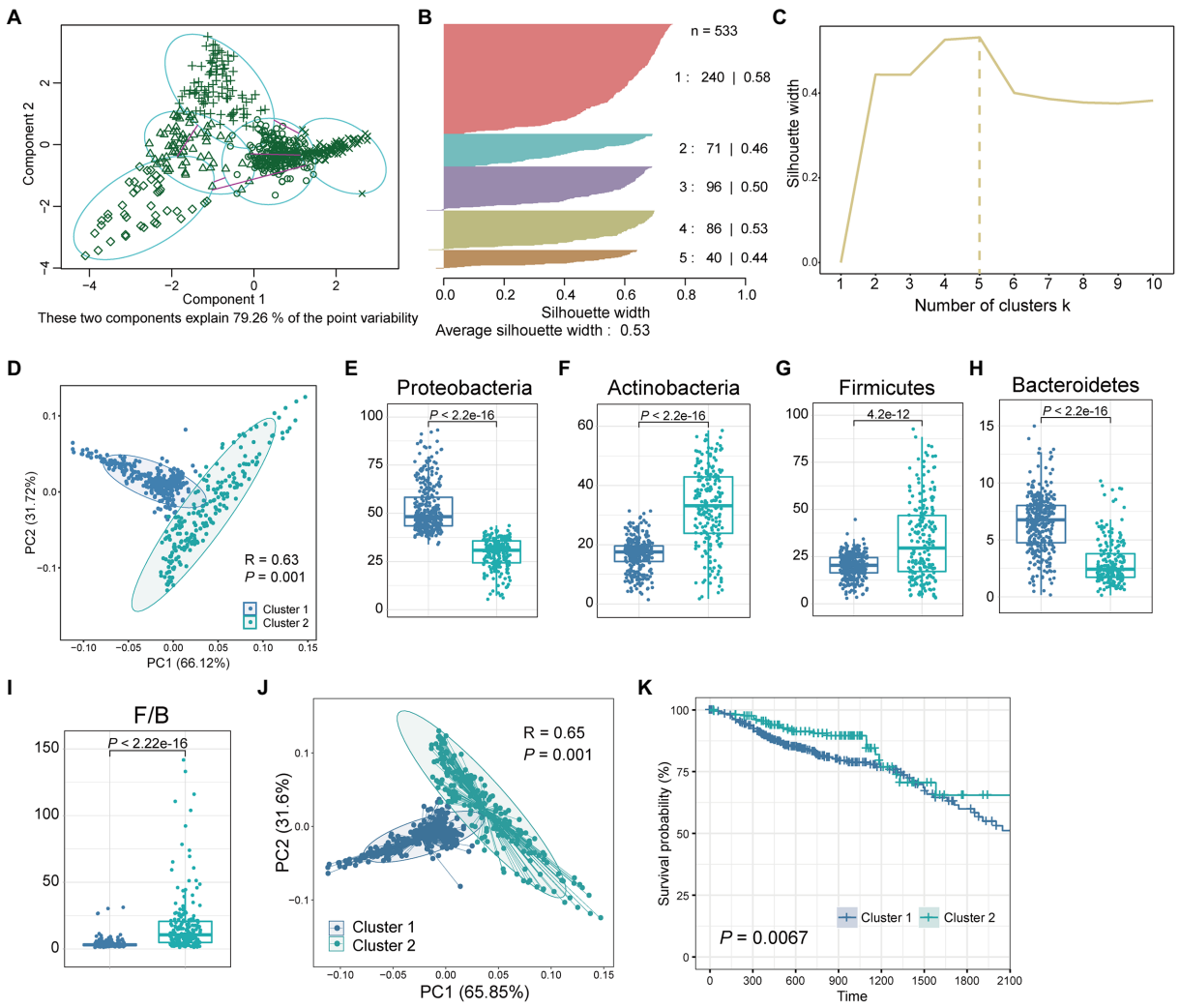


FIGURE 2
Two clusters obtained by clustering the tissue microbiome abundance. **(A)** All samples were clustered into five groups by PAM clustering method. These two components explain 79.26% of the point variability. **(B)** The silhouette width and sample size of the five groups. **(C)** The corresponding silhouette width when the clustering number is 1–10. **(D)** PCA plot of relative abundance of four dominant phyla for CRC samples reveals considerable variation between cluster 1 and cluster 2. The R^2 and p value was calculated by an ANOSIM test. Boxplot of differences in **(E)** Proteobacteria, **(F)** Actinobacteria, **(G)** Firmicutes, **(H)** Bacteroidetes, and **(I)** Firmicutes/Bacteroidetes (F/B) between cluster 1 and cluster 2. The p value was calculated by a Wilcoxon rank-sum test. **(J)** PCA plot of relative abundance of tissue microbe data at the phylum level for CRC samples reveals considerable variation between cluster 1 and cluster 2. **(K)** Kaplan–Meier survival curve for overall survival of cluster 1 and cluster 2. The p value was calculated by log-rank test.

Bacteroidetes had significantly better survival (Figure 3D, $p = 0.048$). Among the four dominant phyla, only Actinobacteria do not affect the survival of CRC patients through their actions (Figure 3B, $p = 0.83$). In conclusion, the significantly lower survival rate of patients in cluster 1 compared with patients in cluster 2 is most likely due to the high abundance of Proteobacteria and Bacteroidetes, as well as the low abundance of Firmicutes in tissue microbes of patients in cluster 1.

A study showed that with the development of health-polyp-adenomas-CRC, the relative abundance of Proteobacteria increased gradually, while the relative abundance of Firmicutes decreased gradually (Liu et al., 2020). A comparative analysis of bacterial phyla levels between groups in 40 samples showed a significant increase in Proteobacteria abundance and a significant decrease in Firmicutes in colorectal cancer tissue compared with normal intestinal mucosa (Yang et al., 2019). Liu et al. (2022) found that Proteobacteria had a positive promoting effect on the risk of colorectal cancer and other diseases. Besides, one study confirmed that compared with healthy individuals, inflammatory bowel disease (IBD) and CRC patients had reduced bacterial diversity and abundance, and significantly enriched Bacteroidetes (Quaglio et al., 2022). While our results are consistent with previous studies, more depth, our data suggest that increased Proteobacteria and Bacteroidetes, and decreased Firmicutes in colorectal tissue may be accompanied by poorer patient survival.

3.4. Genera belonging to these dominant phyla showed significant differences in abundance between the two clusters of patients

Having found significant differences in tissue microbial composition between the two clusters at the phylum level, we next

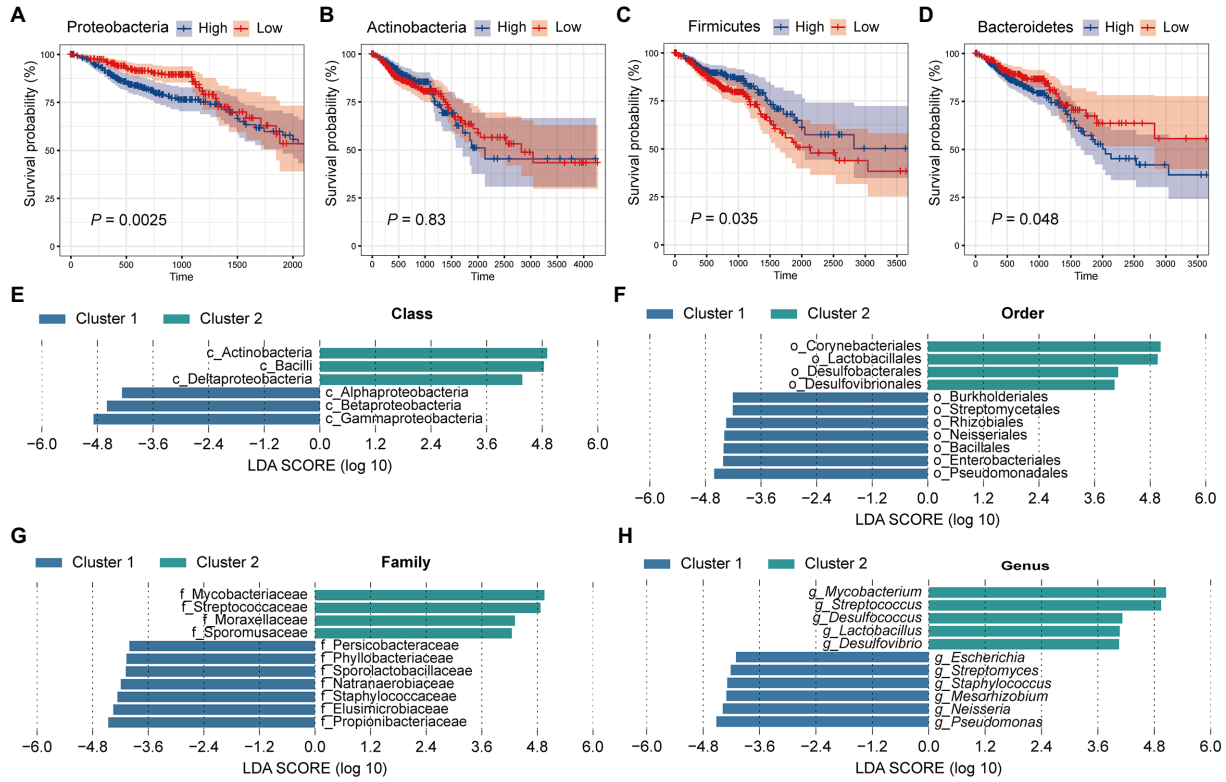


FIGURE 3
Tissue microbes are responsible for the significant difference in survival between cluster 1 and cluster 2. All samples were divided into two groups based on the relative abundance of **(A)** Proteobacteria, **(B)** Actinobacteria, **(C)** Firmicutes, and **(D)** Bacteroidetes, respectively, and survival curves were performed based on these two groups. The *p* value was calculated by log-rank test. LEfSe identified the significantly different species in relative abundance between the two clusters at the **(E)** class, **(F)** order, **(G)** family, and **(H)** genus level, respectively. The LDA threshold is set to 4.

aimed to explore the similarities and differences between the two groups at other levels. For this, LEfSe analysis with a linear discriminant analysis (LDA) threshold of 4 was used to identify significantly different species in the two clusters (Figures 3E–H). First, at the class level, 6 significantly different species were identified. Specifically, Alphaproteobacteria, Betaproteobacteria, and Gammaproteobacteria were significantly enriched in cluster 2, while Deltaproteobacteria and Bacilli were more abundant in cluster 1 (Figure 3E). At the order level, we identified 7 species enriched in cluster 1 and 4 species enriched in cluster 2. Lactobacillales was more abundant in cluster 2 and Burkholderiales and Enterobacteriales were enriched in cluster 1 (Figure 3F). At the family level, Staphylococcaceae, Propionibacteriaceae, and Elusimicrobiaceae were enriched in cluster 1, while Mycobacteriaceae and Streptococcaceae were enriched in cluster 2 (Figure 3G). At the genus level, a total of 11 significantly different genera were identified, of which 5 were significantly enriched in cluster 2 and 6 were significantly more abundant in cluster 1 (Figure 3H). It has been proposed that *Mycobacteria* as non-specific immune enhancers may have the potential to be effective agents for the prevention or treatment of gastrointestinal diseases, including CRC (Kim et al., 2022). The researchers indicated that heat-killed *Mycobacteria tuberculosis* had a protective effect in a model of inflammation-associated CRC. Meanwhile, we found that *Mycobacteria* were significantly enriched in the tissues of cluster 2 patients, and the better survival of cluster 2 patients confirmed this conclusion. Li et al. (2021) demonstrated that co-culture with *Streptococcus thermophilus* or its conditioned medium reduced the proliferation of CRC cells in culture, and oral gavage of *S. thermophilus* significantly reduced tumorigenesis. *Streptococcus*, a genus belonging to the phylum Firmicutes, similarly showed a significant increase in abundance in cluster 2 patients compared with cluster 1. *Lactobacillus*, a genus belonging to Firmicutes, was found to be significantly more abundant in cluster 2 patient tissues. *Lactobacillus* has long been considered an important probiotic for gut health. Studies have suggested that *Lactobacillus gallinarum* prevented intestinal tumors by producing protective metabolites that promoted CRC cell apoptosis (Sugimura et al., 2021). Besides, the F/B ratio in obese mice was reduced by the treatment of *Lactobacillus sakei* NR28 and *Lactobacillus rhamnosus* GG (Stojanov et al., 2020). In a human clinical trial, the beneficial influence of *Lactobacillus salivarius* was demonstrated (Larsen et al., 2013). Besides, univariate cox regression analysis was performed for the genera with the top 30 relative abundance (Supplementary Figure S3). Among them, four genera (*Escherichia*, *Streptococcus*, *Pseudomonas*, and *Bacteroides*) were significantly correlated with patient survival, which was consistent with KM survival analysis (Figure 3H). What's more, the four genera belong to Proteobacteria, Firmicutes, and Bacteroidetes, which was also consistent with our PAM clustering.

Most of the genera significantly enriched in the tissues of cluster 1 patients were pathogenic bacteria of CRC or harmful to intestinal health. For instance, recent studies have identified *Escherichia coli*, a species belonging to *Escherichia*, as one of the candidate pathogens for

CRC (Cheng et al., 2020). A metabolomic and 16S microbiome analysis of 224 stool samples showed a significant increase in *Staphylococcus* in CRC patients (Clos-Garcia et al., 2020). Besides, the relative abundance of *Enterococcus* and *Neisseria* was significantly higher in the fecal microbiota of patients with invasive cancer compared with early cancer. The genus *Pseudomonas* contains a series of pathogens, among which *Pseudomonas aeruginosa* is a common opportunistic pathogen, which is a common nosocomial infection pathogen in patients with immune deficiency (Mielko et al., 2019). The abnormal proportion of *Pseudomonas nucleomonas* produced a proinflammatory microenvironment, promoted the proliferation of CRC cells, and promotes the chemotherapy resistance of CRC (Chen et al., 2022). *Neisseria meningitides*, an aerobic gram-negative diplococcus, contribute to high morbidity in young adults through an epidemic or sporadic meningitis (Rouphael and Stephens, 2012). Taken together, our data demonstrate that the tissue microbes of CRC patients in cluster 1 tend to enrich some pathogenic bacteria that promote the development of CRC, thus leading to poor survival, while patients in cluster 2 have significantly more bacteria that resist the development of tumors.

3.5. The two clusters of patients had different tissue microbiome co-occurrence network properties

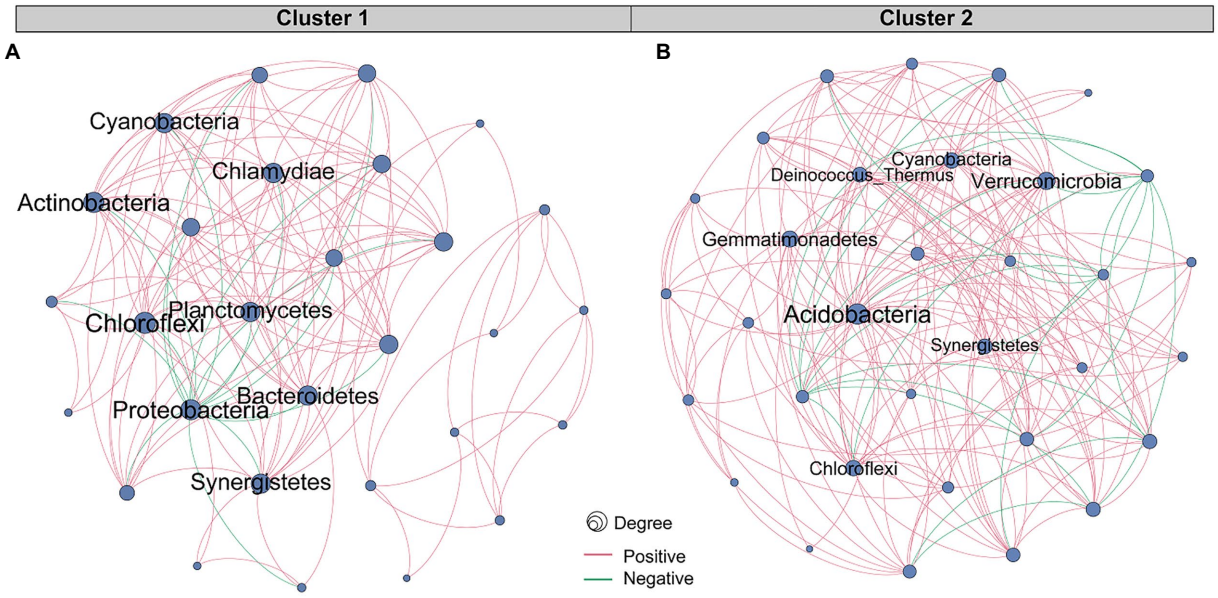
The role of a single or single class of microbes in affecting the occurrence and development of tumors is limited, and the synergistic or antagonistic effects of sufficient species in the microbial community cannot be ignored. Therefore, we constructed co-occurrence networks for the two clusters based on the correlation between species at the phylum level (Figures 4A,B). Network analysis revealed that the nodes and edges of cluster 1 were 29 and 147, respectively, while for cluster 2, they were 33 and 192. For cluster 1, the positive and negative correlations between phylum species were 11.6 and 88.4%, respectively, while for cluster 2, they were 13.0 and 87.0%, respectively. The proportion of positive and negative correlations between tissue microbes in the two clusters was similar. Further, we compared other important network properties between the two clusters, including average degree, diameter, and clustering coefficient (Figure 4C). The results showed that the diameter and clustering coefficient of cluster 1 (6 and 0.755, respectively) were higher than those of cluster 2 (3 and 0.687, respectively), while the average degree of cluster 2 (11.636) was higher than that of cluster 1 (10.138). The important species in the two networks, namely keystone, were significantly different (Table 1). In cluster 1, Chloroflexi, Proteobacteria, and Actinobacteria occupied an important position in the network. However, the keystone species in the network were Acidobacteria, Verrucomicrobia, and Gemmatimonadetes. Besides, compared with cluster 1, the keystone in cluster 2 had a higher degree and weight.

Our study demonstrated that compared with cluster 1, the network of cluster 2 was more complicated. Microbial communities in tumor tissues are not merely collections of independent individuals, but interconnected complexes that communicate, recombine, and coevolve with each other (Layeghifard et al., 2017). Yuan et al. (2022) compared the tissue microbiological co-occurrence networks in 134 lung cancer patients without recurrence or metastasis (non-RM) and 174 patients with recurrence or metastasis (RM) and found that the co-occurrence network of non-RM was more complicated than RM. Recurrence and metastasis as well as survival in our study are both important prognostic indicators of cancer patients (Usuda et al., 2014). Our study shows that the poorer survival of CRC patients is accompanied by a microbiome co-occurrence network of reduced complexity in tissues.

There are several limitations in this study. First, this cohort of 533 CRC patients included confounding factors such as race, country, sex, and age. A recent study looked at the intratumoral microbiota of different cancer types to better understand the influence of age, sex, body mass index (BMI), and ethnicity on the composition of the intratumoral microbiota (Luo et al., 2022). The authors found that race was strongly associated with microbiota abundance, while age, sex, and BMI had little to do with it. Consequently, further analyses should be conducted to distinguish patients of different races and to more accurately identify biologically meaningful microbial markers. A study divided patients with CRC into proximal and distal (Jin et al., 2021), which are not considered the same disease. They found differences in the association of microbes with these two subtypes in CRC patients. For instance, in patients with proximal colon cancers, a high abundance of Fusobacteria was associated with poor prognosis, but not in patients with distal CRC. However, in our study, we did not detect a significant association between Fusobacteria and patient survival. The possible reason is that there are many subtypes of colorectal cancer, and different subtypes may have different associations with tissue microbes. Second, though we show that clusters based on tissue microbiome are associated with survival, we did not provide any prediction model using related microbes. In the future, it will be interesting to develop microbe-based prognosis models. Third, recent studies suggested that tissue histopathological image is correlated with the prognosis of cancers (Liu et al., 2022; Yang et al., 2022; Yao et al., 2022). It would be interesting to study the relationship between tissue microbes and histopathology. Finally, the lack of a healthy control cohort in this study adds a barrier to further understanding changes in tissue microbiota abundance between CRC patients and the normal population. However, tissue from perfectly healthy populations is extremely difficult to obtain, so for colorectal cancer, future studies could consider a control cohort of patients with other intestinal diseases that do not significantly alter the microbial composition of colorectal tissue.

4. Conclusion

The present study advances the understanding of the colorectal microbiota in CRC patients, providing evidence for the critical role of tissue microbes influencing the prognosis of patients *via* the variation of the proportion of probiotics, pathogens, or bacteria that can alter the progression of CRC. Moreover, it provides one possible explanation for the heterogeneity of postoperative survival in CRC patients, such that differences in microbial community composition in colorectal tumor tissues of different patients. Thus, we recommend that before the treatment of CRC patients, it is considered to obtain the microbial content of the tumor tissue of the patients to determine the survival time and other prognosis index of the patients, and



C

Network parameters of cluster 1 and cluster 2

Network	Nodes/edges	Avg. Degree	Diameter	Clust. Coeff.
Cluster 1	29/147	10.138	6	0.755
Cluster 2	33/192	11.636	3	0.687

FIGURE 4
Comparison of co-occurrence network structure and properties between cluster 1 and cluster 2. Co-occurrence networks based on correlation between species at the phylum level of **(A)** cluster 1 and **(B)** cluster 2. The red and green edges represent positive and negative correlations, respectively. Node size is proportional to the relative abundance of species. **(C)** Comparison of co-occurrence network properties between cluster 1 and cluster 2. Avg. Degree, Average degree. Clust. Coeff., Clustering coefficient.

TABLE 1 Comparison of co-occurrence network properties between cluster 1 and cluster 2.

	Phylum	Degree	Eccentricity	Closeness centrality	Betweenness centrality	Clustering
Cluster 1	Chloroflexi	18	5	0.518519	31.274242	0.69281
	Proteobacteria	17	5	0.509091	15.065909	0.772059
	Actinobacteria	17	5	0.509091	17.482576	0.772059
	Bacteroidetes	16	5	0.5	10.774242	0.833333
	Chlamydiae	16	5	0.5	6.274242	0.858333
	Cyanobacteria	16	5	0.5	6.274242	0.858333
	Planctomycetes	16	5	0.5	6.274242	0.858333
	Synergistetes	16	5	0.5	19.848485	0.758333
Cluster 2	Acidobacteria	25	2	0.820513	121.520854	0.326667
	Verrucomicrobia	20	2	0.727273	36.331713	0.542105
	Gemmatimonadetes	18	2	0.695652	45.134963	0.424837
	Chloroflexi	17	2	0.680851	16.273766	0.661765
	Cyanobacteria	17	2	0.680851	20.251597	0.654412
	Deinococcus_Thermus	16	2	0.666667	9.267849	0.741667
	Synergistetes	16	2	0.666667	12.196489	0.691667

The degree represents the number of all edges connected by each node. Closeness centrality represents the sum of the number of nodes that a node can reach divided by the shortest path that can reach the node. Betweenness centrality indicates the ratio between the number of betweenness paths passed by a node by other nodes and the total number of shortest paths in the figure. The eccentricity represents the largest shortest path that a node can reach.

to assist clinicians in making accurate decisions to avoid overtreatment. Extrapolating from this concept, we suggest that for CRC therapy to be beneficial it needs to be coupled to the tissue microbiome profile of patients.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: ftp://ftp.microbio.me/pub/cancer_microbiome_analysis/.

Author contributions

DM and HL contributed to conception and design of the study. YY organized the data. JZ performed the statistical analysis. YX wrote the first draft of the manuscript. YM revised the manuscript. JL, YC, and CX wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version. All authors contributed to the article and approved the submitted version.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1100873/full#supplementary-material

References

Ahlquist, D. A., Skoletsky, J. E., Boynton, K. A., Harrington, J. J., Mahoney, D. W., Pierceall, W. E., et al. (2000). Colorectal cancer screening by detection of altered human DNA in stool: feasibility of a multitarget assay panel. *Gastroenterology* 119, 1219–1227. doi: 10.1053/gast.2000.19580

Artacho, A., Isaac, S., Nayak, R., Flor-Duro, A., Alexander, M., Koo, I., et al. (2021). The pretreatment gut microbiome is associated with lack of response to methotrexate in new-onset rheumatoid arthritis. *Arthritis Rheumatol.* 73, 931–942. doi: 10.1002/art.41622

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

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proc. Int. AAAI Conf. Web Soc. Media* 3, 361–362. doi: 10.1609/icwsm.v3i1.13937

Biswas, R., Ghosh, D., Dutta, B., Halder, U., Goswami, P., and Bandopadhyay, R. (2021). Potential non-coding RNAs from microorganisms and their therapeutic use in the treatment of different human cancers. *Curr. Gene Ther.* 21, 207–215. doi: 10.2174/1566523220999201230204814

Boleij, A., Hechenbleikner, E. M., Goodwin, A. C., Badani, R., Stein, E. M., Lazarev, M. G., et al. (2015). The Bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clin. Infect. Dis.* 60, 208–215. doi: 10.1093/cid/ciu787

Brenner, H., Kloor, M., and Pox, C. P. (2014). Colorectal cancer. *Lancet*. 383, 1490–1502. doi: 10.1016/S0140-6736(13)61649-9

Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., et al. (2017). Analysis of fusobacterium persistence and antibiotic response in colorectal cancer. *Science* 358, 1443–1448. doi: 10.1126/science.aal5240

Castellarin, M., Warren, R. L., Freeman, J. D., Dreolini, L., Krzywinski, M., Strauss, J., et al. (2012). *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome Res.* 22, 299–306. doi: 10.1101/gr.126516.111

Chen, S., Zhang, L., Li, M., Zhang, Y., Sun, M., Wang, L., et al. (2022). *Fusobacterium nucleatum* reduces METTL3-mediated m(6)a modification and contributes to colorectal cancer metastasis. *Nat. Commun.* 13:1248. doi: 10.1038/s41467-022-28913-5

Cheng, Y., Ling, Z., and Li, L. (2020). The intestinal microbiota and colorectal cancer. *Front. Immunol.* 11:615056. doi: 10.3389/fimmu.2020.615056

Cheng, L., Qi, C., Yang, H., Lu, M., Cai, Y., Fu, T., et al. (2022). gutMGene: a comprehensive database for target genes of gut microbes and microbial metabolites. *Nucleic Acids Res.* 50, D795–D800. doi: 10.1093/nar/gkab786

Cienfuegos-Jimenez, O., Vazquez-Garza, E., and Rojas-Martinez, A. (2021). CAR-NK cells for cancer therapy: molecular redesign of the innate antineoplastic response. *Curr. Gene Ther.* 22:1724. doi: 10.2174/1566523222666211217091724

Clos-Garcia, M., Garcia, K., Alonso, C., Iruarrizaga-Lejarreta, M., D'Amato, M., Crespo, A., et al. (2020). Integrative analysis of fecal metagenomics and metabolomics in colorectal cancer. *Cancers (Basel)* 12:1142. doi: 10.3390/cancers12051142

Cullin, N., Azevedo Antunes, C., Straussman, R., Stein-Thoeringer, C. K., and Elinav, E. (2021). Microbiome and cancer. *Cancer Cell* 39, 1317–1341. doi: 10.1016/j.ccell.2021.08.006

Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M., and Wallace, M. B. (2019). Colorectal cancer. *Lancet* 19, 1467–1480. doi: 10.1016/S0140-6736(19)32319-0

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., et al. (2005). Diversity of the human intestinal microbial flora. *Science* 308, 1635–1638. doi: 10.1126/science.1110591

Fulbright, L. E., Ellermann, M., and Arthur, J. C. (2017). The microbiome and the hallmarks of cancer. *PLoS Pathog.* 13:e1006480. doi: 10.1371/journal.ppat.1006480

Grigor'eva, I. N. (2020). Gallstone disease, obesity and the firmicutes/bacteroidetes ratio as a possible biomarker of gut dysbiosis. *J. Pers. Med.* 11:10013. doi: 10.3390/jpm11010013

Haghi, F., Goli, E., Mirzaei, B., and Zeighami, H. (2019). The association between fecal enterotoxigenic *B. fragilis* with colorectal cancer. *BMC Cancer* 19:6115. doi: 10.1186/s12885-019-6115-1

He, B., Dai, C., Lang, J., Bing, P., Tian, G., Wang, B., et al. (2020a). A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. *Biochim. Biophys. Acta Mol. basis Dis.* 1866:165916. doi: 10.1016/j.bbadis.2020.165916

He, B., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., et al. (2020b). TOOme: a novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression. *Front. Bioeng. Biotechnol.* 8:394. doi: 10.3389/fbioe.2020.00394

Heymann, C. J. F., Bard, J. M., Heymann, M. F., Heymann, D., and Bobin-Dubigeon, C. (2021). The intratumoral microbiome: characterization methods and functional impact. *Cancer Lett.* 522, 63–79. doi: 10.1016/j.canlet.2021.09.009

Hu, Y., Zhou, M., Shi, H., Ju, H., Jiang, Q., and Cheng, L. (2017). Measuring disease similarity and predicting disease-related ncRNAs by a novel method. *BMC Med. Genet.* 17:71. doi: 10.1186/s12920-017-0315-9

Jin, M., Shang, F., Wu, J., Fan, Q., Chen, C., Fan, J., et al. (2021). Tumor-associated microbiota in proximal and distal colorectal cancer and their relationships with clinical outcomes. *Front. Microbiol.* 12:727937. doi: 10.3389/fmicb.2021.727937

Kim, Y. M., Choi, J. O., Cho, Y. J., Hong, B. K., Shon, H. J., Kim, B. J., et al. (2022). Mycobacterium potentiates protection from colorectal cancer by gut microbial alterations. *Immunology*. doi: 10.1111/imm.13586

Larsen, N., Vogensen, F. K., Gobel, R. J., Michaelsen, K. F., Forssten, S. D., Lahtinen, S. J., et al. (2013). Effect of *Lactobacillus salivarius* Ls-33 on fecal microbiota in obese adolescents. *Clin. Nutr.* 32, 935–940. doi: 10.1016/j.clnu.2013.02.007

Layeghifard, M., Hwang, D. M., and Guttman, D. S. (2017). Disentangling interactions in the microbiome: a network perspective. *Trends Microbiol.* 25, 217–228. doi: 10.1016/j.tim.2016.11.008

Li, Z., Lu, G., Li, Z., Wu, B., Luo, E., Qiu, X., et al. (2021). Altered actinobacteria and firmicutes phylum associated epitopes in patients with Parkinson's disease. *Front. Immunol.* 12:632482. doi: 10.3389/fimmu.2021.632482

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. *Front. Cell Dev. Biol.* 9:619330. doi: 10.3389/fcell.2021.619330

Liu, X., Tong, X., Zou, Y., Lin, X., Zhao, H., Tian, L., et al. (2022). Mendelian randomization analyses support causal relationships between blood metabolites and the gut microbiome. *Nat. Genet.* 54, 52–61. doi: 10.1038/s41588-021-00968-y

Liu, X., Yuan, P., Li, R., Zhang, D., An, J., Ju, J., et al. (2022). Predicting breast cancer recurrence and metastasis risk by integrating color and texture features of histopathological images and machine learning technologies. *Comput. Biol. Med.* 146:105569. doi: 10.1016/j.compbiomed.2022.105569

Liu, W., Zhang, R., Shu, R., Yu, J., Li, H., Long, H., et al. (2020). Study of the relationship between microbiome and colorectal cancer susceptibility using 16SrRNA sequencing. *Biomed. Res. Int.* 2020:7828392. doi: 10.1155/2020/7828392

Lu, K., Wang, F., Ma, B., Cao, W., Guo, Q., Wang, H., et al. (2021). Teratogenic toxicity evaluation of bladder cancer-specific oncolytic adenovirus on mice. *Curr. Gene Ther.* 21, 160–166. doi: 10.2174/1566523220999201217161258

Luo, M., Liu, Y., Hermida, L. C., Gertz, E. M., Zhang, Z., Li, Q., et al. (2022). Race is a key determinant of the human intratumor microbiome. *Cancer Cell* 40, 901–902. doi: 10.1016/j.ccell.2022.08.007

Magne, F., Gotteland, M., Gauthier, L., Zazueta, A., Pesoa, S., Navarrete, P., et al. (2020). The Firmicutes/Bacteroidetes ratio: a relevant marker of gut Dysbiosis in obese patients? *Nutrients* 12:474. doi: 10.3390/nu12051474

Mielko, K. A., Jablonski, S. J., Milczewska, J., Sands, D., Lukaszewicz, M., and Mlynarz, P. (2019). Metabolomic studies of Pseudomonas aeruginosa. *World J. Microbiol. Biotechnol.* 35:178. doi: 10.1007/s11274-019-2739-1

Nakatsu, G., Li, X., Zhou, H., Sheng, J., Wong, S. H., Wu, W. K., et al. (2015). Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nat. Commun.* 6:8727. doi: 10.1038/ncomms9727

Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwang, Y., Geller, L. T., et al. (2020). The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* 368, 973–980. doi: 10.1126/science.aay9189

Peng, P., Luan, Y., Sun, P., Wang, L., Zeng, X., Wang, Y., et al. (2022). Prognostic factors in stage IV colorectal cancer patients with resection of liver and/or pulmonary metastases: a population-based cohort study. *Front. Oncol.* 12:850937. doi: 10.3389/fonc.2022.850937

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

Qi, C., Cai, Y., Qian, K., Li, X., Ren, J., Wang, P., et al. (2022). gutMDisorder v2.0: a comprehensive database for dysbiosis of gut microbiota in phenotypes and interventions. *Nucleic Acids Res.* 51, D717–D722. doi: 10.1093/nar/gkac871

Quaglio, A. E. V., Grillo, T. G., De Oliveira, E. C. S., Di Stasi, L. C., and Sassaki, L. Y. (2022). Gut microbiota, inflammatory bowel disease and colorectal cancer. *World J. Gastroenterol.* 28, 4053–4060. doi: 10.3748/wjg.v28.i30.4053

Rouphael, N. G., and Stephens, D. S. (2012). Neisseria meningitidis: biology, microbiology, and epidemiology. *Methods Mol. Biol.* 799, 1–20. doi: 10.1007/978-1-61779-346-2_1

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

Sheng, C., Lin, L., Lin, H., Wang, X., Han, Y., and Liu, S. L. (2021). Altered gut microbiota in adults with subjective cognitive decline: the SILCODE study. *J. Alzheimers Dis.* 82, 513–526. doi: 10.3233/JAD-210259

Shin, N. R., Whon, T. W., and Bae, J. W. (2015). Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends Biotechnol.* 33, 496–503. doi: 10.1016/j.tibtech.2015.06.011

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA Cancer J. Clin.* 70, 7–30. doi: 10.3322/caac.21590

Stojanov, S., Berlec, A., and Strukelj, B. (2020). The influence of probiotics on the Firmicutes/Bacteroidetes ratio in the treatment of obesity and inflammatory bowel disease. *Microorganisms* 8:715. doi: 10.3390/microorganisms8111715

Stopinska, K., Radziwon-Zaleska, M., and Domitrz, I. (2021). The microbiota-gut-brain Axis as a key to neuropsychiatric disorders: a mini review. *J. Clin. Med.* 10:4640. doi: 10.3390/jcm10204640

Sugimura, N., Li, Q., Chu, E. S. H., Lau, H. C. H., Fong, W., Liu, W., et al. (2021). *Lactobacillus gallinarum* modulates the gut microbiota and produces anti-cancer metabolites to protect against colorectal tumourigenesis. *Gut* 71, 2011–2021. doi: 10.1136/gutjnl-2020-323951

Tanaka, E., Uchida, D., Shiraha, H., Kato, H., Ohyama, A., Iwamuro, M., et al. (2020). Promising gene therapy using an adenovirus vector carrying REIC/Dkk-3 gene for the treatment of biliary cancer. *Curr. Gene Ther.* 20, 64–70. doi: 10.2174/1566523220666200309125709

Usuda, K., Sagawa, M., Motomo, N., Ueno, M., Tanaka, M., Machida, Y., et al. (2014). Recurrence and metastasis of lung cancer demonstrate decreased diffusion on diffusion-weighted magnetic resonance imaging. *Asian Pac. J. Cancer Prev.* 15, 6843–6848. doi: 10.7314/APJCP.2014.15.16.6843

Wang, P., Zhang, S., He, G., Du, M., Qi, C., Liu, R., et al. (2022). microbioTA: an atlas of the microbiome in multiple disease tissues of Homo sapiens and Mus musculus. *Nucleic Acids Res.* 51, D1345–D1352. doi: 10.1093/nar/gkac851

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* 25, 679–689. doi: 10.1038/s41591-019-0406-6

Wong-Rolle, A., Wei, H. K., Zhao, C., and Jin, C. (2021). Unexpected guests in the tumor microenvironment: microbiome in cancer. *Protein Cell* 12, 426–435. doi: 10.1007/s13238-020-00813-8

Wu, Y., Jiao, N., Zhu, R., Zhang, Y., Wu, D., Wang, A. J., et al. (2021). Identification of microbial markers across populations in early detection of colorectal cancer. *Nat. Commun.* 12:3063. doi: 10.1038/s41467-021-23265-y

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., et al. (2022). Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput. Struct. Biotechnol. J.* 20, 333–342. doi: 10.1016/j.csbj.2021.12.028

Yang, Y., Misra, B. B., Liang, L., Bi, D., Weng, W., Wu, W., et al. (2019). Integrated microbiome and metabolome analysis reveals a novel interplay between commensal bacteria and metabolites in colorectal cancer. *Theranostics* 9, 4101–4114. doi: 10.7150/thno.35186

Yang, M., Yang, H., Ji, L., Hu, X., Tian, G., Wang, B., et al. (2022). A multi-omics machine learning framework in predicting the survival of colorectal cancer patients. *Comput. Biol. Med.* 146:105516. doi: 10.1016/j.compbiomed.2022.105516

Yao, Y., Lv, Y., Tong, L., Liang, Y., Xi, S., Ji, B., et al. (2022). ICSDA: a multi-modal deep learning model to predict breast cancer recurrence and metastasis risk by integrating pathological, clinical and gene expression data. *Brief. Bioinform.* 23:bbac448c. doi: 10.1093/bib/bbac448

Yuan, X., Wang, Z., Li, C., Lv, K., Tian, G., Tang, M., et al. (2022). Bacterial biomarkers capable of identifying recurrence or metastasis carry disease severity information for lung cancer. *Front. Microbiol.* 13:1007831. doi: 10.3389/fmicb.2022.1007831

Zhao, T., Hu, Y., Zang, T., and Cheng, L. (2020). MRTFB regulates the expression of NOMO1 in colon. *Proc. Natl. Acad. Sci. U. S. A.* 117, 7568–7569. doi: 10.1073/pnas.2000499117

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 23 March 2023
DOI 10.3389/fmicb.2023.1159076

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology,
China

REVIEWED BY
Min Chen,
Hunan Institute of Technology,
China
Yuansheng Liu,
Hunan University,
China

*CORRESPONDENCE
Lei Wang
✉ wanglei@xtu.edu.cn
Xianyou Zhu
✉ zxy@hynu.edu.cn

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

RECEIVED 05 February 2023
ACCEPTED 02 March 2023
PUBLISHED 23 March 2023

CITATION
Hu W, Yang X, Wang L and Zhu X (2023) MADGAN:A microbe-disease association prediction model based on generative adversarial networks.
Front. Microbiol. 14:1159076.
doi: 10.3389/fmicb.2023.1159076

MADGAN:A microbe-disease association prediction model based on generative adversarial networks

Weixin Hu[1], Xiaoyu Yang[2], Lei Wang[2,3]* and Xianyou Zhu[1]*

[1]College of Computer Science and Technology, Hengyang Normal University, Hengyang, China, [2]Institute of Bioinformatics Complex Network Big Data, Changsha University, Changsha, China, [3]Big Data Innovation and Entrepreneurship Education Center of Hunan Province, Changsha University, Changsha, China

Researches have demonstrated that microorganisms are indispensable for the nutrition transportation, growth and development of human bodies, and disorder and imbalance of microbiota may lead to the occurrence of diseases. Therefore, it is crucial to study relationships between microbes and diseases. In this manuscript, we proposed a novel prediction model named MADGAN to infer potential microbe-disease associations by combining biological information of microbes and diseases with the generative adversarial networks. To our knowledge, it is the first attempt to use the generative adversarial network to complete this important task. In MADGAN, we firstly constructed different features for microbes and diseases based on multiple similarity metrics. And then, we further adopted graph convolution neural network (GCN) to derive different features for microbes and diseases automatically. Finally, we trained MADGAN to identify latent microbe-disease associations by games between the generation network and the decision network. Especially, in order to prevent over-smoothing during the model training process, we introduced the cross-level weight distribution structure to enhance the depth of the network based on the idea of residual network. Moreover, in order to validate the performance of MADGAN, we conducted comprehensive experiments and case studies based on databases of HMDAD and Disbiome respectively, and experimental results demonstrated that MADGAN not only achieved satisfactory prediction performances, but also outperformed existing state-of-the-art prediction models.

KEYWORDS

microbe-disease associations, graph convolution neural network, generative adversarial network, residual network, computational prediction model

1. Introduction

Microbes are far more numerous than human cells (Integrative HMP (iHMP) Research Network Consortium, 2014; Sender et al., 2016), and play an important role in human beings (Human Microbiome Project Consortium, 2012). The microorganisms parasitic on the human body constitute the human microbial community, and their composition varies from person to person (Human Microbiome Project Consortium, 2012). These microbial populations can not only protect the human body from foreign microorganisms and pathogens, but also participate in intestinal digestion and absorption, and promote metabolism (Guarner and Malagelada, 2003;

Kau et al., 2011). Therefore, to some extent, the human microbial population can even be regarded as human "forgotten organs"(Quigley, 2013), the imbalance of microorganisms will not only lead to the occurrence of nervous system diseases, but also affect the immune and metabolic functions of the human body (Cenit et al., 2017; Li et al., 2017). For example, changes in intestinal microbiota are highly correlated with the pathogenesis of various nervous system diseases, including depression, autism (Kim et al., 2018), asthma (Al-Moamary et al., 2021) and cancer (Schwabe and Jobin, 2013), etc. Of course, there is also evidence showing that microbial populations can help regulate disease as well (Cryan and Dinan, 2012). For instance, researches show that lactic acid bacteria and bifid bacteria play a positive role in regulating anxiety, cognition, pain and depression symptoms (Desbonnet et al., 2010). In addition, Huang pointed out that microorganisms can affect the hypersensitivity and asthma of susceptible people. Early intervention to promote the healthy composition of human microbiome may help prevent asthma (Huang, 2013). Hence, it is meaningful to infer potential relationships between microorganisms and diseases, which can not only help researchers understand the pathogenesis of diseases, but also help us to prevent, diagnose and treat diseases, thus promoting global human health. Utilizing biotechnology to identify microbe-disease associations is time-consuming, costly and blind, so it is meaningful to identify potential microbe-disease associations through computational methods. Up to now, representative calculative methods can be roughly divided into four categories, such as the network-based, binary local features-based, matrix factorization/completion-based and graph neural network-based methods. Among them, the network-based methods infer latent microbe-disease associations by mainly adopting the topology information of different networks. For example, Chen et al. (2017) proposed a KATZ-based model KATZHMDA to infer possible microbe-disease associations based on a newly constructed heterogeneous network, which scores potential disease related microbes by step size and path numbers. Zeng et al. (2022) introduced the knowledge graph into the field of drug discovery, integrated data information through a displayed structure, and strengthened the structured connection and semantic relationship between entities. However, the methods based on binary local features focus on taking microbes and diseases as local objects, and identify potential microbe-disease associations by combining the features between them. For instance, Huang et al. (2017) developed a combined recommendation algorithm based on neighborhood and graph by integrating two independent recommendation models to recommend disease related microbes. In addition, Matrix factorization/completion-based methods aim to decompose the known incidence matrix into two characteristic matrices, and approximate the incidence matrix with the product of the two matrices. For instance, Shen et al. (2017) proposed a matrix factorization-based model for microbe-disease association prediction, which integrated known microbe-disease associations and introduced a collaborative matrix factorization scheme to update the correlation matrix about microbes and diseases for inferring the most possible disease-related microbes. Finally, the graph neural network-based methods used to learn structural data by taking microbe and disease related data as the input of the neural networks, so as to extract and explore features and patterns in graph structural data. For example, Long et al. (2021) developed a graph attention network with inductive matrix completion to detect potential microbe-disease associations. Cheng et al. (2021) used the deep generative model as an entry point to discuss and study the *de novo* molecular design for drug discovery (*de novo* molecular design for drug discovery).

The emergence of generative adversarial networks is another milestone in the field of computer vision. It provides a new tool for solving various image prediction problems. For instance, in 2014, Lan et al. proposed a framework for estimating the generative adversarial network model through the confrontation process, and improved the ability of the model through the mutual game between generative adversarial networks (Goodfellow et al., 2020). However, the generative adversarial network still has problems such as unstable results and difficult training. Hence, Arjovsky et al. (2017) conducted a theoretical analysis of the generative adversarial network and provided an optimal solution. Later, new results appeared in the field of image processing, such as Style GAN (Karras et al., 2019), Cycle GAN (Zhu et al., 2017), SeCGAN (Wu et al., 2019), etc. In recent years, many researchers have begun to explore the application of generative adversarial networks in other fields. For example, Lei et al. (2019) applied it in the direction of dynamic information generation to build a nonlinear time link prediction model. Dai et al. (2021) introduced generative adversarial networks to natural language translation work. Zheng et al. (2022) utilized a generative adversarial network model to predict urban traffic flow.

In this paper, a generative adversarial network framework called MADGAN was designed for latent microbe-disease association prediction, in which, a GCN was adopted to obtain the microbe-disease association features first, and then, we would train the ability of MADGAN by games between the generation network and the decision network. And at the same time, inspired by the idea of residual network, we introduced the cross-level weight distribution structure to enhance the depth of the network to prevent over-smoothing during the model training process. Finally, intensive experiments based on the *k*-fold cross-validation framework were implemented to compare the prediction performance between MADGAN and state-of-the-art prediction models. And as a result, MADGAN was proved to be of satisfactory prediction ability and outperformed existing representative competing models.

2. Materials and methods

2.1. Construction of the microbe-disease association network

In this section, we would download known microbe-disease associations from two well-known public databases including HMDAD (Ma et al., 2017) and Disbiome (Janssens et al., 2018) respectively. Among them, HMDAD[1] is the first microbe-disease association database constructed by ma et al. in 2017, which contains 483 known microbe-disease associations. After removing duplicate data, we finally obtained 450 different known microbe-disease associations between 39 diseases and 292 microbes. Besides, Disbiome[2] is a public microbe-disease association database

1 http://www.cuilab.cn/hmdad

2 https://disbiome.ugent.be/home

constructed by Janssens et al., in which, there are 5,573 known associations between 240 diseases and 1,098 microbes collected from published academic papers. After removing duplicate data, we finally derived 4,351 known microbe-disease associations between 218 diseases and 1,052 microbes. For convenience, let n_d and n_m denote the numbers of newly-downloaded diseases and microbes respectively, then we can obtain a adjacency matrix $A \in \mathbb{R}^{n_d \times n_m}$ as follows: for any given disease d_i and a microbe m_j, if there is a known association between them, there is A_{ij} =1, otherwise, there is A_{ij} =0.

2.2. Multiple similarity calculation of disease

2.2.1. Gaussian interaction profile kernel similarity of disease

Based on the assumption that two similar diseases will show similar interaction and non-interaction relationship with the same microorganism (Chen et al., 2017), in this section, we will first calculate the Gaussian interaction profile kernel similarity between a pair of diseases d_i and d_j as follows:

$$GD(d_i,d_j) = \exp\left(-\lambda_d \| A(i,:) - A(j,:) \|^2\right) \quad (1)$$

Where $A(i,:)$ and $A(j,:)$ represent the i^{th} and j^{th} rows of the adjacency matrix A respectively, and λ_d denotes the normalized kernel bandwidths that can be calculated as follows:

$$\lambda_d = \frac{1}{\left(\frac{1}{n_d}\sum_{i=1}^{n_d} \| A(i,:) \|^2\right)} \quad (2)$$

2.2.2. Cosine similarity of disease

Based on the assumption that if two diseases are similar to each other, then their cosine curves will be more coincident, in this section, we will define the cosine similarity between a pair of diseases d_i and d_j as follows:

$$CD(d_i,d_j) = \left(A(i,:) \cdot A(j,:) \right) / \left(|A(i,:)| * |A(j,:)| \right) \quad (3)$$

The result of cosine similarity has good stability and certainty, the calculation speed is fast and the result is more intuitive. Suitable for large-scale information retrieval. Where $A(i,:) \cdot A(j,:)$ denotes multiplying the vectors of row i and row j, $|A(i,:)|$ represents the mode of $A(i,:)$, and $|A(j,:)|$ represents the mode of $A(j,:)$. $|A(i,:)| * |A(j,:)|$ represents the multiplication of two moduli, and then the value of the modulus is removed by the product of the vector, and finally the cosine value of the angle between the two diseases is obtained, that is, the cosine similarity. The calculation result of cosine similarity is between -1 and 1. When the similarity between two diseases is extremely high, the calculation result tends to be 1. When the similarity between two diseases is very low, the calculation result tends to -1.

2.2.3. Functional similarity of disease

Based on the assumption that similar diseases tend to interact with similar genes, in this section, we will calculate the disease functional similarity based on the functional associations between disease-related genes (Xu and Li, 2006; Wei and Liu, 2020) as follows: Firstly, we download the gene interactions from HumanNet database[3], in which, every interaction has an associated log-likelihood score (LLS). And then, for any given diseases d_i and d_j, let $G_i = \{g_{i_1}, g_{i_2}, \ldots, g_{i_m}\}$ and $G_j = \{g_{j_1}, g_{j_2}, \ldots, g_{j_n}\}$ denote the newly-obtained gene sets of d_i and d_j separately, we will define the functional similarity between d_i and d_j as follows:

$$DFS(d_i,d_j) = \frac{\sum_{g_k \in G_i} F_{G_j}(g_k) + \sum_{g_k \in G_j} F_{G_i}(g_k)}{m+n} \quad (4)$$

Where $F_{G_i}(g_p) = \max_{g_q \in G_i} \left(FSS(g_p, g_q) \right)$, and $FSS(g_p, g_q)$ is the functional similarity score between the genes g_p and g_q, which can be calculated as follows:

$$FSS(g_p,g_q) = \begin{cases} 1 & if\ p = q \\ \dfrac{LLS(g_p,g_q) - LLS_{\min}}{LLS_{\max} - LLS_{\min}} & if\ p \neq q \end{cases} \quad (5)$$

Where $LLS_{\max}$ and $LLS_{\min}$ represent the maximum value of *LLS* and the minimum value of *LLS* in HumanNet, respectively.

Thereafter, by combining above GIP kernel similarity, disease cosine similarity and functional similarity of disease, we can obtain an integrated similarity matrix of disease as follows:

$$DS = \frac{GD + CD + DFS}{3} \quad (6)$$

2.3. Multiple similarity calculation of microbe

2.3.1. Gaussian interaction profile kernel similarity of microbe

In the same way, we can calculate the gaussian interaction profile kernel similarity between any two microbes m_i and m_j as follows:

$$MD(m_i,m_j) = \exp\left(-\lambda_m \| A(:,i) - A(:,j) \|^2\right) \quad (7)$$

Where $A(:,i)$ and $A(:,j)$ represent the i^{th} and j^{th} columns of the adjacency matrix A respectively, and λ_m denotes the normalized kernel bandwidths that can be calculated as follows:

$$\lambda_m = \frac{1}{\left(\frac{1}{n_m}\sum_{i=1}^{n_m} \| A(:,i) \|^2\right)} \quad (8)$$

3 https://www.inetbio.org/humannet

2.3.2. Cosine similarity of microbe

Similarly, the cosine similarity between any two microbes m_i and m_j can be obtained as follows:

$$CM\left(m_i,m_j\right)=\left(A\left(:,i\right)\cdot A\left(:,j\right)\right)/\left(\left|A\left(:,i\right)\right|\times\left|A\left(:,j\right)\right|\right) \quad (9)$$

The calculation process of cosine similarity between two microorganisms is the same as that of disease cosine similarity. Similarly, when the similarity between two microorganisms is extremely high, the calculation result tends to be 1. When the similarity between two microorganisms is very low, the calculation result tends to −1.

2.3.3. Functional similarity of microbe

In this section, we will calculate the functional similarity of microbe by using the following method proposed in the reference (Zhang et al., 2018): for any given disease d_t, it is first represented by a Directed Acyclic Graph $DAG_{d_t}=\left(V_{d_t},E_{d_t}\right)$, where V_{d_t} includes the disease d_t and its ancestor diseases, E_{d_t} contains all the directed edges from parent nodes to children nodes (Wang et al., 2010), and then, the semantic contribution of the disease d_l in V_{d_t} to d_t is defined as:

$$SC_{d_t}\left(d_i\right)=\begin{cases}1 & if\ d_l=d_t\\ \max\left\{0.5\times SC_{d_t}\left(d_l'\right)\middle|d_l'\in children\ of\ d_l\right\} & otherwise\end{cases} \quad (10)$$

The semantic value of disease d_t is formulated by:

$$SV_{d_t}=\sum_{d_l\in V_{d_t}}SC_{d_t}\left(d_l\right) \quad (11)$$

Then, the semantic similarity between any two diseases d_i and d_j can be defined as follows:

$$DSS\left(d_i,d_j\right)=\frac{\sum_{d_l\in V_{d_i}\cap V_{d_j}}\left(SC_{d_i}\left(d_l\right)+SC_{d_j}\left(d_l\right)\right)}{SV_{d_i}+SV_{d_j}} \quad (12)$$

Besides, based on above formulae, we can further define the similarity between the disease d_i and a set of diseases D as follows:

$$DS\left(d_i,D\right)=\max_{d_j\in D}\left(DSS\left(d_i,d_j\right)\right) \quad (13)$$

Hence, for any two given microbes m_i and m_j, we can calculate the function similarity between them as follows:

$$MFS\left(m_i,m_j\right)=\frac{\sum_{d_j\in D_j}DS\left(d_j,D_i\right)+\sum_{d_j\in D_i}DS\left(d_j,D_j\right)}{\left|D_i\right|+\left|D_j\right|} \quad (14)$$

Where D_i denotes the set of diseases associated with the microbe m_i, and D_j represents the set of diseases associated with the microbe m_j.

Obviously, by combining above GIP kernel similarity, disease cosine similarity and functional similarity of microbe, we can obtain an integrated similarity matrix of microbe as follows:

$$MS=\frac{MD+CM+MFS}{3} \quad (15)$$

2.4. Construction of the heterogeneous network

Based on above descriptions, it is easy to see that we can construct a heterogeneous network Y through integrating the integrated similarity matrix DS of disease and the integrated similarity matrix MS of microbe with the adjacency matrix A as follows:

$$Y=\begin{bmatrix}DS & A\\ A^T & MS\end{bmatrix} \quad (16)$$

3. Methods

The main framework of this paper is generative adversarial networks. A generative adversarial network consists of a generative network and a decision network, and it works by enhancing the model's capabilities during the mutual gaming of the two networks. As shown in Figure 1, the information of known microbial-disease association data is extracted from the database, and after the calculation of similarity, it is input into the generative network. The core of the generative network consists of a GCN layer and an attention mechanism, which consists of a graph convolutional layer and a sparse graph convolutional layer. The data are passed through the generative network to generate prediction results, and the prediction results and the original sample data are input into the discriminator, which distinguishes the real results from the generated results and returns to update the model parameters of the generative network. This is a game process, in which the generative network needs to generate prediction results that are sufficient to confuse the judgment of the discriminator, while the discriminator needs to correctly distinguish the generated results from the true results. The ability of the generative network model is continuously improved during the game until the discriminator and the generative network reach an equilibrium, i.e., the probability of both the predicted and true outcomes is one half.

The generator network uses the information of the data set to output data samples, and the generator $G(\bullet)$ obtains a random sample z from the data samples, and z conforms to the $p(z)$ probability distribution. After the generator generates data, it will be sent to the discriminator $D(\bullet)$, and the discriminator will try to predict the authenticity of the data after receiving real data or generated data. At the same time, it also needs a sample x from the real data distribution $p_{data}(x)$, the discriminator uses the activation function to solve a binary classification task, and outputs a value of 0–1 to distinguish the real result from the predicted result.

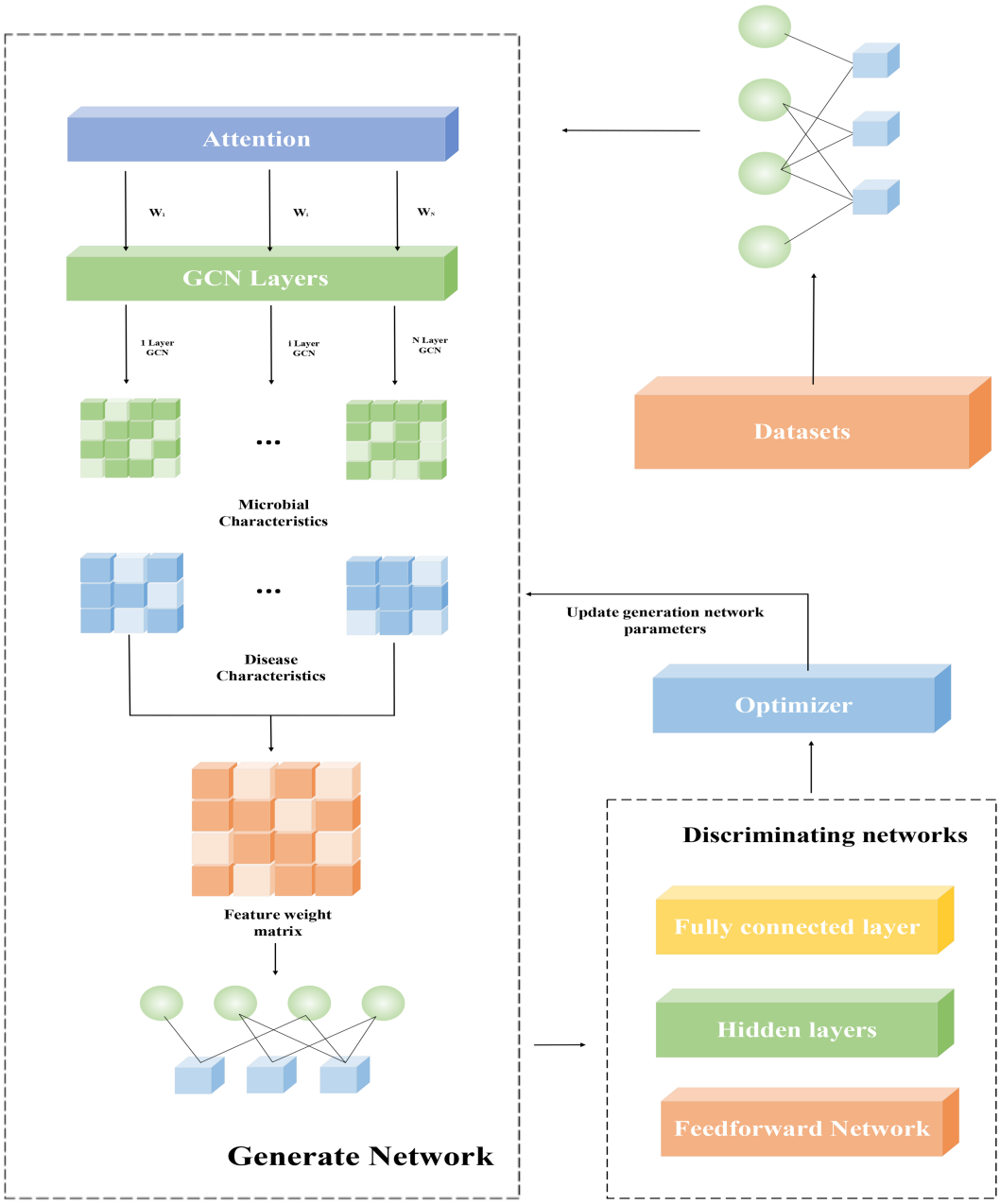


FIGURE 1
The general framework of the model.

The game process of generative adversarial networks can be expressed as follows:

$$\min \max V(D,G) = E_{x\sim p_{data}(x)}\left[logD(x)\right] + E_{z\sim p(z)}\left[1-logD(G(z))\right] \quad (17)$$

Among them, x is the real feature matrix, and $G(z)$ is the feature matrix generated by the generation network. $p_{data}(x)$ is the probability distribution of x, and $p(z)$ is the probability distribution of z. The optimization goal of training D to adjust its parameters is to maximize $D(x)$ and minimize $D(G(x))$, and the optimization goal of training G to adjust its parameters is to minimize $\max V(D,G)$. E stands for entropy, $x \sim p_{data}(x)$ stands for x is from $p_{data}(x)$ real data distribution. The meaning represented by $E_{x\sim p_{data}(x)}\left[logD(x)\right]$ is the entropy value from the real data distribution after passing the identifier. For data from the real data distribution, the ideal goal of the discriminator is to fully identify it, that is, predict the result as 1. Therefore, $E_{x\sim p_{data}(x)}\left[logD(x)\right]$ can also be regarded as the probability of the discriminator to distinguish real data, and the higher the probability, the better. The log function does not affect the relationship between variables, and its function is to amplify our loss to facilitate the calculation and optimization of the model. $E_{z\sim p(z)}\left[1-logD(G(z))\right]$ can be regarded as the entropy value after the input generated data passes through the discriminator, and also represents the probability

of the discriminator to distinguish the fake sample data. The smaller the probability, the better. $\min\max V(D,G)$ is expressed as a confrontation between the generator and the discriminator. The generator $G(\bullet)$ hopes that the discriminator cannot distinguish fake samples, so it hopes to minimize the result of $1-logD(G(z))$. The discriminator is the opposite, it hopes to better distinguish between true and false, that is, the result of maximizing $1-logD(G(z))$. This is also the origin of this formula. At the end of training, there will often be a balanced form.

The core of the principle of generative adversarial networks lies in the game between the generative network and the decision network. The core of the generative network is composed of GCN layers. In order to deepen the model depth of the generative network and thus generate more accurate prediction results, we use a residual network-like idea to optimize the model. We deepen the network while retaining the shallow features according to the weights, which makes the model less susceptible to phenomena such as oversmoothing and gradient explosion during the iterative process. As shown in Figure 2, the direct mapping is shown on the left, and the associated graph convolution operation and activation function are shown on the right.

The purpose of adding this structure is to increase the depth of the network. Under this premise, problems such as over-smoothing and gradient explosion are avoided. At the same time, combined with the attention mechanism, we have carried out weight ratios on both sides on the basis of similar residual ideas to achieve better results. Its formula derivation is as follows:

$$h_l = h_0 + \sum_{i,j=1}^{L} F(h_i, W_j) \tag{18}$$

Among them, h_L is the feature matrix output by each layer, and $l \in \{1,..,L\}$. W_j is the weight assigned to each layer, and $F(\bullet)$ is the graph convolution function.

And the relevant formula of $F(\bullet)$ is as follows:

$$F(z,W)_l = f(F(z)_{l-1}, Y) = \mu\left(D^{-\frac{1}{2}}YD^{-\frac{1}{2}}F(z)_{l-1}W_{l-1}\right) \tag{19}$$

Where $l \in \{1,..,L\}$, $F(z)_l$ is the feature matrix generated by the lth layer GCN network, $D = diag\left(\sum_{j=1}^{N_m+N_d} Y_{i,j}\right)$ is a diagonal matrix, and W_l is the weight matrix trained on the lth layer. And $\mu(\bullet)$ is an activation function. In this paper, the RELU function is used as the activation function. The formula is as follows:

$$RELU(x) = \begin{cases} x, x > 0 \\ 0, x \le 0 \end{cases} \tag{20}$$

The weight calculation formula of W_l is as follows:

$$W_l = \frac{1}{L} \tag{21}$$

Graph Convolution (GCN) is a convolutional model applied by CNN in the field of graph structure. Different from CNN to achieve feature extraction by processing pixels, graph convolution uses spectral graph theory to map the graph structure transformation to the frequency domain through Fourier transform for processing, and finally perform inverse transformation. Compared with CNN that handles neat pixels, GCN can more effectively extract the correlation features between two points. For data with associated structures, the ability to effectively extract spatial features brought by GCN can better help them complete their tasks. In our model, the reconstructed heterogeneous network feature matrix is input into the generative network and processed as the input of the GCN model. Formula (19) reflects the training process of the GCN model, and $\mathbf{z}$ is the input data. The function of $\mathrm{D}^{-\frac{1}{2}}\mathrm{YD}^{-\frac{1}{2}}$ is to dilute the importance of nodes with high degrees, and to balance the weight information of nodes with different degrees. Therefore, formula (19) can also be simplified as:

$$F(z,W)_l = \mu\left(\tilde{Y}F(z)_{l-1}W_{l-1}\right) \tag{22}$$

Among them, the role of $\tilde{Y}F(z)_{l-1}$ is to retain the information inherited by the upper layer nodes during the information transmission process, that is, to aggregate the information of the surrounding nodes to update the information of its own nodes.

The role of the discriminator is to distinguish between real and fake samples, and our discriminator consists of a fully connected feed-forward network, a hidden layer and an output layer. The discriminator alternately receives generated samples and real samples, and updates the parameters of the generated network through the discriminative results. Here we adopt the framework of WassersteinGAN to train the discriminator. The biggest difference between WGAN and traditional GAN is that the output layer is a linear layer and does not require a nonlinear activation function. Expressed in a formula it is:

$$D(z) = \mu(z'W_h + b_h)W_o + b_o \tag{23}$$

Among them, z is the input data, and z is the long vector after dimension reconstruction. $\mu(\bullet)$ is the activation function of the hidden layer, W_h and b_h are the hidden layer parameters, and W_o and b_o are the output layer parameters.

As shown in Algorithm 1, the input is a known microbial-disease association matrix A. The similarity matrix of microorganisms and diseases is computed to construct the heterogeneous network Y. The new feature matrix is fed into the generative network. After initializing the optimizer, the generated prediction results are output after N rounds of training. The generated prediction results and sample data are input into the discriminator, and the parameter information of the generative network is updated according to the output results of the discriminator, and the completed generative network model is saved after several rounds of training.

Algorithm 1: Algorithm of our proposed method

Inputs: Known associations matrix $A \in \mathbb{R}^{n_m \times n_d}$, microbe similarity matrix $K_S^m \in \mathbb{R}^{N_m \times N_m}$, disease similarity matrix $K_S^d \in \mathbb{R}^{n_d \times n_d}$;

Output: The completed training of the generative network model

Step 1: Constructing the heterogeneous network $Y \in \mathbb{R}^{(n_m+n_d)\times(n_d+n_m)}$ according to Formula (16);

Step 2: Input the feature matrix into the generative network, initializing Optimizer Parameter Information;

Step 3: for $i = 1 \rightarrow N$ do (N is the number of training rounds of the generative adversarial network)

for $l = 1 \rightarrow L$ do (L is the depth of the graph convolution model)

Compute the feature embedding of the L layer and output the generated prediction results

end for

Input the generated results and sample data into the decision network

Update optimizer parameter information

end for

Step 4: Save the model of the generative network

4. Experiments and results

4.1. Experimental setup

In this section, we adopted 5-fold cross validation(5cv) and 2-fold cross validation to assess the performance of our model. In the k-fold cross validation framework, all known microbe-disease associations in HMDAD and Disbiome were divided to k-subsets. In the process of model training, (k-1)-subsets are selected as the training set, and the remaining one as the test set. It is worth noting that there are no known negative samples, we regarded unknown associations as negative samples. After the training samples are input into MADGAN, all association pairs will get a predictive value. If the prediction score is higher than the given threshold, it will be considered as successful prediction. Obviously, different true positive rate and false positive rate can be obtained when setting different thresholds. The specific calculation formula is as follows:

$$\mathrm{TPR} = \frac{TP}{TP + FN}$$

$$\mathrm{FPR} = \frac{FP}{FP + TN} \tag{24}$$

Where TP and TN represent the numbers of positive samples correctly judged as positive samples and negative samples correctly judged as negative samples, respectively; FP and FN are the numbers of negative samples incorrectly judged as positive samples and positive samples incorrectly judged as negative samples. By setting different thresholds, we can get multiple groups of different TPRs and FPRs. Then, TPR and FPR under different thresholds are taken as the x-axis and y-axis respectively, the receiver operating characteristics (ROC) can be further plotted, and the area under the line is taken to evaluate the prediction performance of the model.

4.2. Parameter analysis

We performed multiple experimental and parametric analyses on the HMDAD database and the Disbiome database, respectively. As shown in Figure 3, we analyzed the experimental results generated by HMDAD in terms of the number of layers and embedding. We used a similar idea of residual network to deepen the number of layers of GCN to 4. After several rounds of training, the experimental results and loss values were maintained at a certain level, but we could see from the experimental results that after the number of layers was raised to 5, the experimental results could not be maintained at a certain level as in the previous layers, which we judged to be due to the limitation of the size of the dataset that made it impossible to deepen the network further. We judge that this is due to the limitation of the dataset size, which makes it impossible to deepen the network further, otherwise the phenomenon of oversmoothing will occur. We also compared different embedding values. Different embedding values take different time to train. When the embedding value is 128, the training time cost is greater than when the embedding value is 32. However, when the model depth is deepened to 5 layers, the embedding value of 128 cannot maintain good experimental results, and the embedding values of 32 and 64 are not affected much, but we think that further deepening the model depth and embedding values of 32 and 64 is also oversmoothing can occur, resulting in poor results.

For the Disbiome database, we also conducted multiple experiments, but the Disbiome database is much larger than the HMDAD database, and we were able to maintain the results at a certain level after deepening the GCN layers with our network up to 20 layers, without reaching the limit. We did not find the limit value due to the limitation of the experimental equipment, but we can understand that the experimental results did not deteriorate after deepening to more than 20 layers.

4.3. Comparison with state-of-the-art methods

In order to evaluate the performance of MADGAN, we compare our model with six state-of-the-art methods that includes network-based methods, binary local features-based methods, matrix factorization/completion-based methods and graph neural network-based methods. KATZHMDA and NTSHMDA are network-based methods, NGRHMDA and BiRWMP are binary local features-based methods, GRNMFHMDA is matrix factorization-based method, and GATMDA is graph neural network-based method. The comparison results of all these methods were shown in Tables 1, 2 respectively.

As shown in Tables 1, 2, we used 5 times of cross-validation and 2 times of cross-validation to conduct comparative experiments on the two databases. In experiments on the HMDAD database, our model performs better than other models. The 5-fold cross-validation method makes better use of the data set than the 2-fold cross-validation method, so it performs better. The data sample size of the Disbiome database is much larger than that of HMDAD, and its training time is also much longer than that of HMDAD. However, compared with HMDAD, the experimental results of all models have declined. We believe that part of the reason is that the depth of the model cannot support the training of a large number of sample data. Even if we use

10.3389/fmicb.2023.1159076

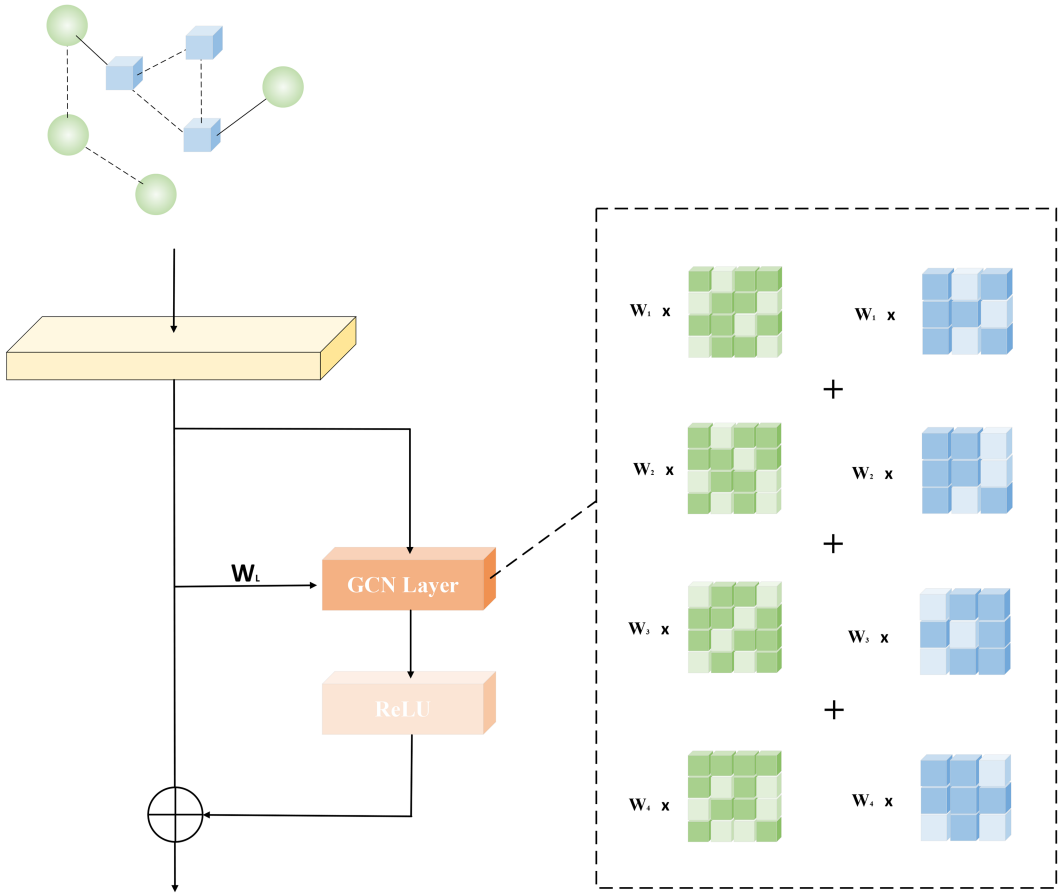


FIGURE 2
Generate network core model structure diagram.

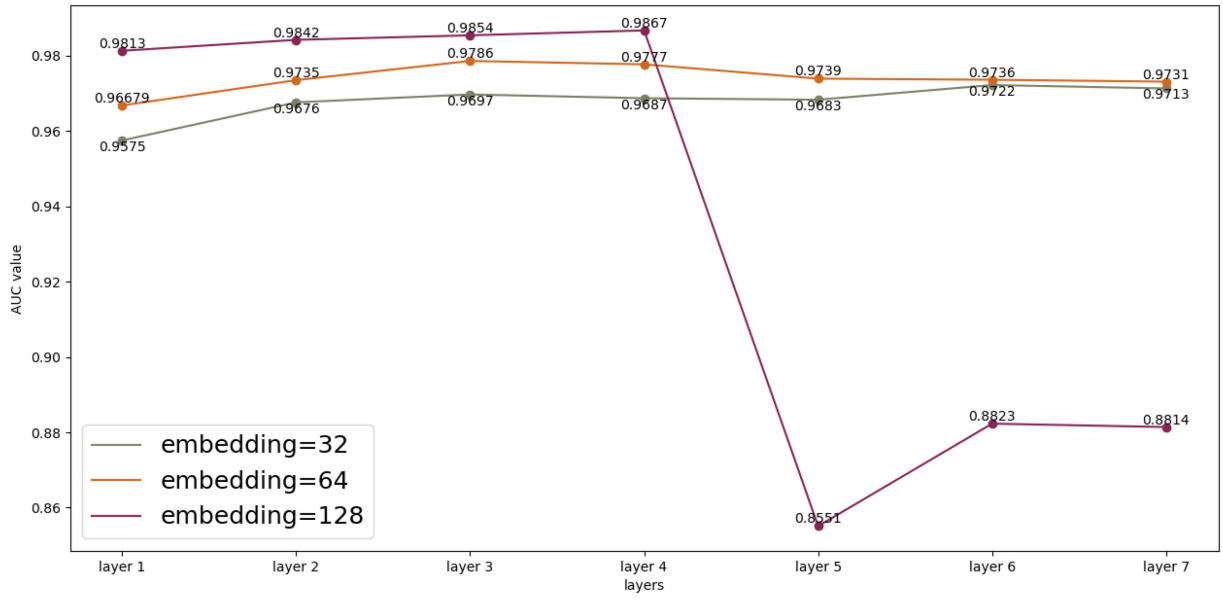


FIGURE 3
Model parameters analysis on the HMDAD dataset.

TABLE 1 Comparison performance between our model and state-of-the-art models based on HMDAD dataset.

Methods	AUC(5-fold cv)	AUC(2-fold cv)
KATZHMDA (Zhu et al., 2021) (network-based)	0.8703±0.0199	0.8755±0.0103
NTSHMDA (Luo and Long, 2018) (network-based)	0.8982±0.0312	0.8615±0.0151
NGRHMDA (Huang et al., 2017) (binary local features-based)	0.8921±0.0327	0.8929±0.0059
BiRWMP (Luo and Xiao, 2017) (binary local features-based)	0.8777±0.0089	0.8698±0.0079
GRNMFHMDA (He et al., 2018) (matrix factorization-based)	0.8806±0.0156	0.8756±0.0164
GATMDA (Long et al., 2021) (graph neural network-based)	0.9554±0.0184	0.9538±0.0049
Our model	0.9867±0.0078	0.9708±0.0117

TABLE 2 Comparison performance between our model and state-of-the-art models based on Disbiome dataset.

Methods	AUC(5-fold cv)	AUC(2-fold cv)
KATZHMDA (Zhu et al., 2021) (network-based)	0.6779±0.0141	0.6696±0.0058
NTSHMDA (Luo and Long, 2018) (network-based)	0.8294±0.0071	0.8086±0.0058
NGRHMDA (Huang et al., 2017) (binary local features-based)	0.8313±0.0052	0.8233±0.0046
BiRWMP (Luo and Xiao, 2017) (binary local features-based)	0.8344±0.0089	0.8139±0.0060
GRNMFHMDA (He et al., 2018) (matrix factorization-based)	0.8609±0.0047	0.8501±0.0017
GATMDA (Long et al., 2021) (graph neural network-based)	0.9307±0.0079	0.9296±0.0154
Our model	0.9428±0.0026	0.9290±0.0068

the method to deepen the depth of the model, it can only slightly improve the experimental effect. Another part of the reason may be because of the equipment environment.

5. Case study

In this section, we choose three diseases of asthma, Chronic Obstructive Pulmonary Disease (COPD) and Type 2 Diabetes (T2D) for case studies on the HMDAD to further verify the performance of our model. Specifically, we rank the above three related microorganisms in the predicted score results, and then select the top 20 microorganisms and evaluate the prediction performance of MADGAN through literature retrieval.

Asthma is a disease with heterogeneous process, accompanied by recurrent wheezing, chest tightness, dyspnea, indirect cough and other symptoms(Al-Moamary et al., 2021). It is reported that in 2010, about 8% of people were affected by asthma, especially in children, and the incidence rate is still rising(Guilbert et al., 2014). Asthma has been proved to be closely related to microorganisms(Çalışkan et al., 2013). For example, Haemophilia, Neisseria and Moraxella in the lungs of asthmatic patients have been proved to be closely related to the increased risk of neonatal oral and pharyngeal asthma, and Staphylococcus has been found in the respiratory tract of asthmatic children(Sullivan et al., 2016). These findings may provide a new method for the treatment of asthma. We choose the top 20 microorganisms related to asthma predicted by our model and then search the literature for further verification. The results are shown in the Table 3.

COPD is a lung disease that worsens over time, as long as the symptoms are shortness of breath and cough. By 2015, COPD patients accounted for about 2.4% of the global population (James et al., 2018). Due to the high smoking rate and aging population in developing countries, the death toll of COPD patients is rising rapidly. Although the treatment can delay the deterioration of COPD, there is no cure. Considering that there is a lot of evidence indicating the association between microbiome and COPD, for example, Galiana et al. (2014) found that the diversity of patients with high COPD was lower than that of patients with mild and moderate COPD. Therefore, we select the top 20 microorganisms related to COPD predicted by our model and then search the literature for further verification. The results are shown in the Table 4.

6. Conclusion

Deeply understanding the relationship between microorganisms and diseases can not only reveal the pathogenesis of more human diseases, but also provide new insights into disease prevention, diagnosis and treatment, thus promoting human health. Predicting the potential microbe-disease associations can help biologists to screen the most relevant microorganisms that cause diseases, thus reducing the time and cost of biological verification experiment (Zhou et al., 2017; Uchiyama et al., 2019). In this paper, we developed a deep learning model, named MADGAN, to predict potential microbe-disease associations. We adequately exploit multi-sources of abundant biological data to capture similarity features of microbes and diseases. This helps to predict new microbes (or new diseases) with few or no known association. In order to derive more informative representations, we propose graph convoluted neural network to learn representations for microbes and diseases. Meanwhile, the model is trained through the game between the generation network and the decision network. Finally, we utilized residual network and the cross-level weight distribution structure to enhance the depth of the network to prevent over-smoothing during model training. Comprehensive experiments demonstrated that MADGAN achieved satisfactory predictive performance.

However, although our model has good prediction performance, it still has some limitations and is expected to be further improved in the future. On the one hand, our model is a supervised learning framework, which means that our model cannot predict all new microorganisms

TABLE 3 The top 20 asthma-associated microbes predicted by MADGAN.

Rank	Microbe	Evidence
1	*Clostridium innocuum*	PMID:18672296
2	*Staphylococcus epidermidis*	PMID:6694502
3	*Streptobacillus*	PMID:6326694
4	*Burkholderiales bacterium* Smarlab 3,302,047	Unconfirmed
5	*Dorea*	PMID:30937143
6	*Stenotrophomonas maltophilia*	PMID:20537287
7	*Mannheimia*	PMID:10967288
8	Rikenellaceae	PMID:33204702
9	*Streptococcus parasanguinis*	PMID:17950502
10	*Yersinia*	PMID:10719781
11	*Alistipes*	PMID:33759390
12	*Corynebacterium*	PMID:22994424
13	*Erysipelotrichales*	PMID:22994424
14	*Mobiluncus*	Unconfirmed
15	*Cronobacter*	Unconfirmed
17	*Eubacteriaceae*	Unconfirmed
18	Unidentified bacterium ZF3	Unconfirmed
19	*Prevotellaceae*	PMID: 34422359
20	*Oxalobacteraceae*	PMID: 21194740

TABLE 4 The top 20 COPD-associated microbes predicted by MADGAN.

Rank	Microbe	Evidence
1	*Bacteroides*	PMID: 36498063
2	*Bacteroides* sp. CJ78	Unconfirmed
3	*Bacteroides vulgatus*	Unconfirmed
4	*Bacteroidetes*	PMID: 33063421
5	*Clostridiales bacterium* 80/3	Unconfirmed
6	*Clostridium cocleatum*	PMID:20857523
7	*Clostridium ramosum*	Unconfirmed
8	*Enterococcus*	PMID:24629344
9	*Erwinia*	Unconfirmed
10	*Escherichia*	PMID: 21605476
11	*Eubacteriaceae*	Unconfirmed
12	*Firmicutes*	PMID: 32353489
13	*Firmicutes bacterium* EG14	Unconfirmed
14	*Fusobacterium*	PMID: 35034433
15	*Verrucomicrobia*	PMID: 32295442
17	*Actinomyces*	PMID: 31174538
18	*Lachnospiraceae bacterium* A2	Unconfirmed
19	*Enterococcus faecalis*	PMID: 26623628
20	*Clostridia bacterium* TSW07CA7	Unconfirmeda

and diseases. In the future, we will consider integrating multiple prior biological information, such as microbe-drug disease association and drug-disease association, to develop an unsupervised learning framework. On the other hand, it is still a huge challenge for MADGAN to forecast on large-scale datasets. In the future, we will consider integrating the results of multiple datasets to build datasets, so as to improve the prediction performance of the model on large datasets.

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

WH and XY produced the main ideas, and did the modeling, computation and analysis and also wrote the manuscript. LW and XZ provided supervision and effective scientific advice and related ideas, research design guidance, and added value to the article through editing and contributing completions. All authors contributed to the article and approved the submitted version.

Funding

This work was partly sponsored by the Hunan Provincial Natural Science Foundation of China (No. 2022JJ50138), the National Natural Science Foundation of China (No. 62272064), the Key project of Changsha Science and technology Plan (No. KQ2203001), the Science and Technology Innovation Program of Hunan Province (No. 2016TP1020), and the Hunan Provincial Education Department Scientific Research Project (No.20B080).

Acknowledgments

The authors thank the referees for suggestions that helped improve the paper substantially.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1159076/full#supplementary-material

References

Al-Moamary, M. S., Alhaider, S. A., Alangari, A. A., Idrees, M. M., Zeitouni, M. O., Al Ghobain, M. O., et al. (2021). The Saudi initiative for asthma-2021 update: guidelines for the diagnosis and management of asthma in adults and children. *Ann. Thorac. Med.* 16, 4–56. doi: 10.4103/atm.ATM_697_20

Arjovsky, M., Chintala, S., and Bottou, L.. (2017). *Wasserstein Generative Adversarial Networks*. International Conference on Machine Learning, pp. 214–223.

Çalışkan, M., Bochkov, Y. A., Kreiner-Møller, E., Bønnelykke, K., Stein, M. M., Du, G., et al. (2013). Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. *N. Engl. J. Med.* 368, 1398–1407. doi: 10.1056/NEJMoa1211592

Cenit, M. C., Sanz, Y., and Codoñer-Franch, P. (2017). Influence of gut microbiota on neuropsychiatric disorders. *WJG* 23, 5486–5498. doi: 10.3748/wjg.v23.i30.5486

Chen, X., Huang, Y.-A., You, Z.-H., Yan, G. Y., and Wang, X. S. (2017). A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 33, 733–739. doi: 10.1093/bioinformatics/btw715

Cheng, Y., Gong, Y., Liu, Y., Song, B., and Zou, Q. (2021). Molecular design in drug discovery: a comprehensive review of deep generative models. *Brief. Bioinform.* 22:bbab344. doi: 10.1093/bib/bbab344

Cryan, J. F., and Dinan, T. G. (2012). Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nat. Rev. Neurosci.* 13, 701–712. doi: 10.1038/nrn3346

Dai, H., Chen, C., Li, Y., and Yuan, Y. (2021). GCNGAN: translating natural language to programming language based on GAN. *J. Phys.* 1873:012070. doi: 10.1088/1742-6596/1873/1/012070

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

Galiana, A., Aguirre, E., Rodriguez, J. C., Mira, A., Santibanez, M., Candela, I., et al. (2014). Sputum microbiota in moderate versus severe patients with COPD. *Eur. Respir. J.* 43, 1787–1790. doi: 10.1183/09031936.00191513

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Commun. ACM* 63, 139–144. doi: 10.1145/3422622

Guarner, F., and Malagelada, J.-R. (2003). Gut flora in health and disease. *Lancet* 361, 512–519. doi: 10.1016/S0140-6736(03)12489-0

Guilbert, T. W., Mauger, D. T., and Lemanske, R. F. (2014). Childhood asthma-predictive phenotype. The journal of allergy and clinical immunology. *In Pract.* 2, 664–670. doi: 10.1016/j.jaip.2014.09.010

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

Huang, Y. J. (2013). Asthma microbiome studies and the potential for new therapeutic strategies. *Curr Allergy Asthma Rep* 13, 453–461. doi: 10.1007/s11882-013-0355-y

Huang, Y.-A., You, Z.-H., Chen, X., Huang, Z. A., Zhang, S., and Yan, G. Y. (2017). Prediction of microbe–disease association from the integration of neighbor and graph with collaborative recommendation model. *J. Transl. Med.* 15, 1–11. doi: 10.1186/s12967-017-1304-7

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Integrative HMP (iHMP) Research Network Consortium (2014). The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* 16, 276–289. doi: 10.1016/j.chom.2014.08.014

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., et al. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *Lancet* 392, 1789–1858. doi: 10.1016/S0140-6736(18)32279-7

Janssens, Y., Nielandt, J., Bronselaer, A., Debunne, N., Verbeke, F., Wynendaele, E., et al. (2018). Disbiome database: linking the microbiome to disease. *BMC Microbiol.* 18:50. doi: 10.1186/s12866-018-1197-5

Karras, T., Laine, S., and Aila, T.. (2019). *A Style-Based Generator Architecture for Generative Adversarial Networks*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410.

Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., and Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature* 474, 327–336. doi: 10.1038/nature10213

Kim, N., Yun, M., Oh, Y. J., and Choi, H. J. (2018). Mind-altering with the gut: modulation of the gut-brain axis with probiotics. *J. Microbiol.* 56, 172–182. doi: 10.1007/s12275-018-8032-4

Lei, K., Qin, M., Bai, B., Zhang, G., and Yang, M.. (2019). *GCN-GAN: A Non-linear Temporal Link Prediction Model for Weighted Dynamic Networks*. IEEE INFOCOM 2019-IEEE Conference on Computer Communications. IEEE, pp. 388–396.

Li, X., Watanabe, K., and Kimura, I. (2017). Gut microbiota Dysbiosis drives and implies novel therapeutic strategies for diabetes mellitus and related metabolic diseases. *Front. Immunol.* 8:1882. doi: 10.3389/fimmu.2017.01882

Long, Y., Luo, J., Zhang, Y., and Xia, Y. (2021). Predicting human microbe–disease associations via graph attention networks with inductive matrix completion. *Brief. Bioinform.* 22:bbaa146. doi: 10.1093/bib/bbaa146

Luo, J., and Long, Y. (2018). NTSHMDA: prediction of human microbe-disease association based on random walk by integrating network topological similarity. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 1341–1351. doi: 10.1109/TCBB.2018.2883041

Luo, J., and Xiao, Q. (2017). A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on heterogeneous network. *J. Biomed. Inform.* 66, 194–203. doi: 10.1016/j.jbi.2017.01.008

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2017). An analysis of human microbe–disease associations. *Brief. Bioinform.* 18, 85–97. doi: 10.1093/bib/bbw005

Quigley, E. M. M. (2013). Gut bacteria in health and disease. *Gastroenterol. Hepatol.* 9, 560–569.

Schwabe, R. F., and Jobin, C. (2013). The microbiome and cancer. *Nat. Rev. Cancer* 13, 800–812. doi: 10.1038/nrc3610

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

Shen, Z., Jiang, Z., and Bao, W. (2017). CMFHMDA: collaborative matrix factorization for human microbe-disease association prediction. *Intell. Comput. Theor. Appl.*, 261–269. doi: 10.1007/978-3-319-63312-1_24

Sullivan, A., Hunt, E., MacSharry, J., and Murphy, D. M. (2016). The microbiome and the pathophysiology of asthma. *Respir. Res.* 17:163. doi: 10.1186/s12931-016-0479-4

Uchiyama, I., Mihara, M., Nishide, H., Chiba, H., and Kato, M. (2019). MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons. *Nucleic Acids Res.* 47, D382–D389. doi: 10.1093/nar/gky1054

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241

Wei, H., and Liu, B. (2020). iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. *Brief. Bioinform.* 21, 1356–1367. doi: 10.1093/bib/bbz057

Wu, H., Feng, J., Tian, X., Xu, F., Liu, Y., Wang, X. F., et al. (2019). *secGAN: A Cycle-Consistent GAN for Securely-recoverable Video Transformation*. Proceedings of the 2019 Workshop on Hot Topics in Video Analytics and Intelligent Edges, pp. 33–38.

Xu, J., and Li, Y. (2006). Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* 22, 2800–2805. doi: 10.1093/bioinformatics/btl467

Zeng, X., Tu, X., Liu, Y., Fu, X., and Su, Y. (2022). Toward better drug discovery with knowledge graph. *Curr. Opin. Struct. Biol.* 72, 114–126. doi: 10.1016/j.sbi.2021.09.003

Zhang, W., Yang, W., Lu, X., Huang, F., and Luo, F. (2018). The bi-direction similarity integration method for predicting microbe-disease associations. *IEEE Access* 6, 38052–38061. doi: 10.1109/ACCESS.2018.2851751

Zheng, H., Li, X., Li, Y., Yan, Z., and Li, T. (2022). GCN-GAN: integrating graph convolutional network and generative adversarial network for traffic flow prediction. *IEEE Access* 10, 94051–94062. doi: 10.1109/ACCESS.2022.3204036

Zhou, T., Tan, L., Cederquist, G. Y., Fan, Y., Hartley, B. J., Mukherjee, S., et al. (2017). High-content screening in hPSC-neural progenitors identifies drug candidates that inhibit Zika virus infection in fetal-like organoids and adult brain. *Cell Stem Cell* 21, 274–283.e5. doi: 10.1016/j.stem.2017.06.017

Zhu, L., Duan, G., Yan, C., and Wang, J. (2021). Prediction of microbe-drug associations based on chemical structures and the KATZ measure. *Curr. Bioinforma.* 16, 807–819. doi: 10.2174/1574893616666210204144721

Zhu, J. Y., Park, T., Isola, P., and Efros, A. A.. (2017). *Unpaired Image-to-image Translation Using Cycle-consistent Adversarial Networks*. Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232.

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 28 April 2023
DOI 10.3389/fmicb.2023.1170559

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Sha Tian,
Macau University of Science and Technology, China
Chenxing Xia,
Anhui University of Science and Technology, China

*CORRESPONDENCE
Hao Chen
✉ chenhao@hnu.edu.cn
Linlin Zhuo
✉ zhuoninnin@163.com

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

RECEIVED 21 February 2023
ACCEPTED 21 March 2023
PUBLISHED 28 April 2023

CITATION
Liao Q, Ye Y, Li Z, Chen H and Zhuo L (2023) Prediction of miRNA-disease associations in microbes based on graph convolutional networks and autoencoders. *Front. Microbiol.* 14:1170559. doi: 10.3389/fmicb.2023.1170559

Prediction of miRNA-disease associations in microbes based on graph convolutional networks and autoencoders

Qingquan Liao[1], Yuxiang Ye[2], Zihang Li[3], Hao Chen[1]* and Linlin Zhuo[2]*

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, [2]School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou, China, [3]School of Computing and Data Science, Xiamen University Malaysia, Sepang, Selangor, Malaysia

MicroRNAs (miRNAs) are short RNA molecular fragments that regulate gene expression by targeting and inhibiting the expression of specific RNAs. Due to the fact that microRNAs affect many diseases in microbial ecology, it is necessary to predict microRNAs' association with diseases at the microbial level. To this end, we propose a novel model, termed as GCNA-MDA, where dual-autoencoder and graph convolutional network (GCN) are integrated to predict miRNA-disease association. The proposed method leverages autoencoders to extract robust representations of miRNAs and diseases and meantime exploits GCN to capture the topological information of miRNA-disease networks. To alleviate the impact of insufficient information for the original data, the association similarity and feature similarity data are combined to calculate a more complete initial basic vector of nodes. The experimental results on the benchmark datasets demonstrate that compared with the existing representative methods, the proposed method has achieved the superior performance and its precision reaches up to 0.8982. These results demonstrate that the proposed method can serve as a tool for exploring miRNA-disease associations in microbial environments.

KEYWORDS

miRNA-disease association, microbial ecology, dual-autoencoder, graph convolutional network, insufficient information, topological information, robust representations

1. Introduction

MiRNAs are a class of endogenous short RNAs that have multiple important regulatory functions in the microbial environment. MiRNAs exert a significant influence in microbial ecology such as metabolism (Karp and Ambros, 2005), cell growth (Ambros, 2003), immune response (Jung et al., 2006), proliferation (Miska, 2005), cell cycle regulation (Liu et al., 2022a) and tumor invasion (Meng et al., 2007). Moreover, miRNAs completes the process of regulating gene expression by base-pairing with target RNA (Jopling et al., 2005; Vasudevan et al., 2007). As a result, miRNAs can effectively predict the occurrence of diseases in microbial ecology and contribute in prevention and diagnosis. HMDD and Human Cancer Differentially Expressed miRNA Database (dbDEMC) contains miRNA-disease related information (Li et al., 2014). However, the data available for research are relatively scarce, and the choice of wet assays to determine miRNA-disease associations is expensive. Thus, it is crucial to design an effective model to handle the experimental testing process (Chen et al., 2019a, 2021; Wang et al., 2019; Zhu et al., 2021).

In the field of biocomputing, correlation studies between various molecules have been conducted. For example, researchers predict the interaction between circRNA and disease (Wang et al., 2021), miRNA and lncRNA (Zhang et al., 2021), lncRNA and protein (Hu et al., 2018), etc. The aforementioned methods are necessary to predict miRNA-diseases, and most of them are based on complex networks. This line of research works builds one or multi networks on the original interaction datasets, and predicts disease-related miRNAs by integrating multi-level data. In general, these approaches can make reasonable predications about miRNA relatedness based on similar disease phenotypes and similar functions, and vice versa (You et al., 2017; Chen et al., 2018a,d, 2019b). For instance, Jiang et al. established a scoring mechanism for predicting disease-miRNA correlations based on miRNA-disease heterogeneous networks, and applied hypergeometric distribution to predict the strength of miRNA-disease associations (Jiang et al., 2010). Guided by global information of the data, Chen et al. proposed a strategy based on random walk to predict the association between diseases and miRNAs (Chen et al., 2012). Considering the fact that most of models cannot accurately predict miRNAs associated with isolated disease individuals, Zeng et al. added some perturbations to the network to train the predictor (Zeng et al., 2018). Recently, researchers have explored a wide range of miRNA functions, which increases the complexity of analyzing gene expression and regulatory networks in common diseases today (Vickers et al., 2014). Moreover, studies have shown that miRNAs participate in the regulation of many cardiovascular-related diseases. These studies demonstrate new aspects of miRNAs in the field of life sciences, and analyzing the regulation of these miRNAs on cardiovascular-related diseases is extremely valuable for proposing new diagnostic and preventive strategies.

Some studies based on statistical methods to predict miRNA-disease associations are attracting more and more attention from the researchers. For example, Li et al. constructed an SVM classifier based on miRNAs associated with specific tumor phenotypes (Li et al., 2012). This model is only for the prediction of diseases such as tumors and may not be suitable for other diseases. Considering the shortage of negative samples in supervised learning models, Yan et al. proposed a model that can reveal the interaction between diseases and miRNAs based on the principle of regularized least squares (Chen and Yan, 2014). This model can predict the associated miRNAs of emerging diseases, thanking to its semi-supervised learning strategy. Chen et al. demonstrated a computational model of matrix decomposition and heterogeneity network inference for predicting miRNA-disease associations (Chen et al., 2018c). In this model, similarities in disease signatures and disease-miRNA associations are integrated into a unified network. However, model parameters are relatively large, and how to reasonably set the parameters is a very challenging task. Xu et al. developed a novel model based on probabilistic matrix factorization (Xu et al., 2019). This model firstly integrates the similarity in the miRNA-disease network; And then performs a probability matrix factorization operation based on the interaction matrix and the similarity matrix.

However, the aforementioned models cannot still achieve promising performance in predicting miRNA-disease associations. Note that deep learning technology has recently been applied to the field of biological computing (Fu et al., 2020; Cai et al., 2021a,b; Liu et al., 2022c; Peng et al., 2022a,b,c; Tian et al., 2022; Xu et al., 2023; Zhang et al., 2023). For instance, Chen et al. constructed a restricted Boltzmann model that can predict associations in different domains (Chen et al., 2015). Because the variability among multiple types cannot be fully modeled, the prediction accuracy is not promising. Chen et al. pre-trained all miRNA-disease pairs on a restricted Boltzmann model and fine-tuned on DBN on the same proportion of positive and negative samples to obtain prediction scores (Chen, 2021). Peng et al. extract features based on a three-autoencoder and then apply a convolutional network to predict the final label (Peng et al., 2019).

Recently, graph neural networks have received much attention from the researchers. For instance, Chen et al. developed a method for miRNA disease association determination based on heterogeneous graphs (Vickers et al., 2014). Furthermore, Chen et al. proposed a network-integrated miRNA-disease-associated internal and external score prediction method (Chen and Zhang, 2014). Chen et al. proposed a predictive model integrating matrix deconstruction and heterogeneous graph aggregation (Chen et al., 2016). Chen et al. utilized matrix factorization to alleviate the influence of noise in adjacent matrices, and then perform node aggregation operations on heterogeneous networks. Mugunga proposed a predictive model based on path features and random walk to obtain correlation scores for miRNA-associated diseases, and potential miRNA-disease associations would be associated with high prediction scores (Mugunga et al., 2017). Guo et al. used a decision fusion strategy to prioritize the results of existing methods, and then verified the effectiveness of the decision fusion strategy (Guang, 2018). Zeng et al. constructed a heterogeneous network to predict potential associations between miRNAs and disease, while also accounting for dataset imbalance (Zeng, 2017). The model also uses a multi-layer perceptron-based approach to predict miRNA-disease pairs, integrating a variety of biological data resources.

Although the aforementioned methods are outstanding in predicting miRNA-disease associations, few studies consider the similarity and topological information comprehensively. Generally speaking, when the topological structure is very sparse, feature information becomes more important in association prediction; when feature information is incomplete, topological information can also play an auxiliary role. Inspired by this guidance, we propose a GCN and autoencoder-based approach that can comprehensively consider both feature and topological information in miRNA-disease networks. Our contributions can be summarized as follows:

1. We develop a GCNA-MDA model to predict miRNA-disease association based on GCN and autoencoders, which achieves the excellent performance. We employ dual-autoencoders to extract disease and miRNA features, which improves the robustness of node presentation. At the same time, we apply a 2-layer GCN to further aggregate disease and miRNA node features by fully considering the topological information.
2. We propose a robust strategy for constructing miRNA and disease basic feature matrix. Combining feature similarity and Gaussian similarity, a unified similarity matrix is constructed. Adding association information to the disease and

miRNA nodes respectively make the feature representation more abundant, thus alleviate the negative impact of insufficient data.

3. We conduct multiple comparison experiments on the HMDD dataset to verify that the GCNA-MDA model can accurately perform the prediction task. Moreover, we construct case studies to verify that the GCNA-MDA model can indeed be applied to examine the specific miRNA-disease associations.

2. Materials and methods

2.1. Dataset

The dataset used in the experiment could be downloaded from the HMDD v2.0 database (Li et al., 2014). The dataset includes 5430 validated associations generated by 495 miRNAs and 383 diseases. It can be abbreviated as adjacency matrix A, in which there are 495×383 miRNA disease associations. If disease d is associated with miRNA m, the association relationship is satisfied, that is, $A(m, d) = 1$, otherwise its value is 0.

2.2. Constructing miRNA and disease basic feature matrix

In this section, we describe in detail the process of constructing robust initial feature for miRNAs and diseases. These similarity matrices can be used as the input matrices for the autoencoder in the next stage. The main process will be introduced below.

2.2.1. Disease feature similarly

Based on the collected disease original feature information, its feature similarity network can be constructed (Schriml et al., 2012). Specifically, we apply the strategy of DAG to denote these diseases. For a disease node d, it is denoted by$DAG(d) = (d, v(d), e(d))$. $v(d)$ represents the set of nodes reached to d, and $e(d)$ represents all edges linked to d. In the DAG graph, the feature contribution weight W of the upper node x to d is calculated as follows:

$$W1_d(x) = \begin{cases} 1 & if \quad x = d \\ max\{\bigtriangledown * W1_d(x') | x' \in x_{children}\} & if \quad x \neq d, \end{cases} \tag{1}$$

where $\bigtriangledown$ represents the adjustment parameter of W, which is empirically set to 0.5 (Chen and Yan, 2013). Based on d and its upper nodes, the feature representation value of d can be calculated as follows:

$$Df1(d) = \sum_{x \in v(d)} W1_d(x). \tag{2}$$

We hypothesize that the greater the number of DAGs shared between two disease nodes, the smaller the difference between the two nodes may be. Thus, the feature similarity of two disease nodes A and B can be calculated as:

$$FS1(A, B) = \frac{\sum_{x \in v(A) \bigcap v(B)} W1_A(x) + W1_B(x)}{Df1(A) + Df1(B)} \tag{3}$$

For disease node d, if two nodes involve approximately the same $DAG(d)$ level, then two nodes should have different occurrence ratios and their contribution to the feature weight of disease d should be different. Thus, we propose the following equation to compute the influence of disease x on d:

$$W2_d(x) = -log \frac{|DAG(x)|}{|D|}, \tag{4}$$

where D denotes the disease set, and $|\cdot|$ denotes the operation of calculating the number of elements in the set. Similarly, the feature representation value of d and the feature similarity of two disease nodes A and B can be calculated as Equations 5 and 6, respectively:

$$Df2(d) = \sum_{x \in v(d)} W2_d(x), \tag{5}$$

$$FS2(A, B) = \frac{\sum_{x \in v(A) \bigcap v(B)} W2_A(x) + W2_B(x)}{Df2(A) + Df2(B)}. \tag{6}$$

Combining the two measure methods to obtain a more reasonable feature similarity, the calculation equation is as follows:

$$FS(A, B) = \frac{FS1(A, B) + FS2(A, B)}{2}. \tag{7}$$

2.2.2. Similarity based on Gaussian

We hypothesize that two miRNAs with small functional differences should be associated with diseases with similar properties (Van Laarhoven et al., 2011). Based on this assumption, we apply the Gaussian kernel distance calculation equation to calculate the similarity between disease nodes D_a and D_b:

$$GD(D_a, D_b) = exp(-\gamma_d \| Index(D_a) - Index(D_b) \|^2), \tag{8}$$

where

$$-\gamma_d = -\gamma'_d (\frac{1}{|D|} \sum_{i=1}^{|D|} \| Index(D_i) \|^2), \tag{9}$$

and γ_d represents the Gaussian kernel parameter, and represents the index function, which can index the row vector of the matrix. Similarly, the Gaussian kernel distance formula between miRNA nodes miR_a and miR_b is as follows:

$$GM(miR_a, miR_b) = exp(-\gamma_m \| Index(miR_a) - Index(miR_b) \|^2), \tag{10}$$

where

$$-\gamma_m = -\gamma'_m (\frac{1}{|M|} \sum_{i=1}^{|M|} \| Index(miR_i) \|^2), \tag{11}$$

and M represents the miRNA node set, and γ_d and γ_m are often set to 1 empirically (Chen and Yan, 2013).

2.2.3. Similarity integration

Due to missing data, some disease pairs may not exist in the feature similarity. For this case, using Gaussian kernel distance to measure the distance between diseases can robustly reflect the differences between diseases. Therefore, the calculation formula of the overall similarity between disease nodes A and B is formulated as

$$SD(A,B) = \begin{cases} \frac{GD(A,B) + FS(A,B)}{2} & if \quad x = d \\ GD(A,B) & if \quad x \neq d. \end{cases} \tag{12}$$

Similarly, the calculation equation of the overall similarity between miRNA nodes X and Y is representated as follows:

$$SM(X,Y) = \begin{cases} \frac{GM(X,Y) + FM(X,Y)}{2} & if \quad FM(X,Y) \quad exists \\ GM(X,Y) & otherwise, \end{cases} \tag{13}$$

where $FM(\cdot,\cdot)$ denotes the functional similarity score between two miRNA nodes.

2.3. Model design

In this section, we propose GCNA-MDA model for predicting miRNA-disease associations based on GCNs and dual-autoencoders. It mainly consists of three parts: firstly, a new similarity calculation strategy is used to obtain the initial basic feature matrix of miRNA (or disease); secondly, a dual-autoencoder is applied to extract the robust expression of miRNA and disease respectively; finally, a 2-layer GCN is applied to predict miRNA-disease associations. Next, the GCNA-MDA model architecture will be introduced in detail, and its overall framework is shown in Figure 1.

2.3.1. Node representation

In this subsection, a novel signature expression for miRNA (or disease) nodes is proposed. Considering that the direct interaction information between miRNA and disease is very important, we add disease-related information to the features of miRNA nodes. Similarly, we also add the corresponding miRNA information to the disease node. Specifically, according to formulas (13) and (12), we calculate the respective feature vectors based on miRNAs and diseases, respectively. Based on the above formula, the fusion with the miRNA-disease association matrix can be obtained:

$$F_d = (SD_1R_1, ..., SD_1R_{495}, ..., SD_{383}R_1, ..., SD_{383}R_{495})^T, \tag{14}$$

$$F_m = (SM_1C_1, ..., SM_1C_{495}, ..., SM_{383}C_1, ..., SM_{383}C_{495})^T, \tag{15}$$

where R_i and C_j represent the $i-th$ row and $j-th$ column vectors of the miRNA-disease association matrix, respectively. Subsequently, the matrices F_m and F_d of miRNAs and diseases were fed into a dual-autoencoder, respectively.

2.3.2. Feature extraction with dual-autoencoders

Based on the above presentation, the node expression of the miRNA (or disease) node fused with the correlation relationship can be obtained. Obviously, the number of nodes is small (383 and 495), but the vector length of each node is high (equal to twice the number of nodes of each type). In this case, the deep neural network may suffer from insufficient samples. Fortunately, autoencoders can play their unique role in this situation. With the strategy of unsupervised learning, the automatic encoding machine no longer needs a large number of samples for its training. This is convenient for us to extract more robust features for the next stage of association prediction tasks.

We extract features of miRNAs and disease nodes separately based on a symmetric dual-autoencoder. The process is mainly divided into two stages of encoding and decoding. During the encoding phase, the basis vectors of the nodes obtained in the previous section is fed into the encoder network. By setting a reasonable number of dimensions, low-rank feature vectors of miRNAs and diseases can be obtained. The calculation method in the encoder is:

$$Y = \sigma_e(W_eX + b_e), \tag{16}$$

where $\sigma_e()$ represents the sigmod activation function. W_e and b_e represent the weight and bias matrices in the encoder, respectively. Both matrices can be efficiently trained in the encoder. Thus, the low-rank vectors obtained from the encoding stage are fed into the decoder network. By setting a reasonable number of dimensions, robust feature vectors for miRNAs and diseases can be obtained. The calculation method in the decoder is:

$$F = \sigma_d(W_dX + b_d), \tag{17}$$

where $\sigma_d(\cdot)$ represents the sigmod activation function. W_d and b_d represent the weight and bias matrices in the decoder, respectively. F is stored as the final feature vector and is fed to the GCN in the next stage for association prediction tasks. To minimize the final feature distribution and the node's initial basic feature distribution, an optimization objective of the dual-autoencoder can be set as:

$$Loss = \sum_{x \in X} \|x - F_x\|^2. \tag{18}$$

In our research, we apply the common square loss function as the optimization objective. The X matrix covers all miRNA and disease nodes, and x is a row vector in the X matrix, which can be regarded as a certain node. In the last layer of the decoder, the node vector length is empirically set to 128.

2.3.3. Predict miRNA–Disease association by GCN

Through the aforementioned process, we can obtain robust features of miRNAs and disease nodes. It is well known that graph neural networks can well aggregate node features and fully consider the topological information of miRNA-disease networks. Therefore, this study uses GCN to predict whether there is an association between miRNA nodes and disease nodes. Since GCN is suitable for tasks on graphs with only one type of nodes and one type of links. Therefore, in order to obtain a unified node adjacency matrix, it is necessary to splice miRNA nodes and disease nodes. For adjacency matrix A, the first 495 indexes of its row (or column) represent miRNA, and the last 383 indexes

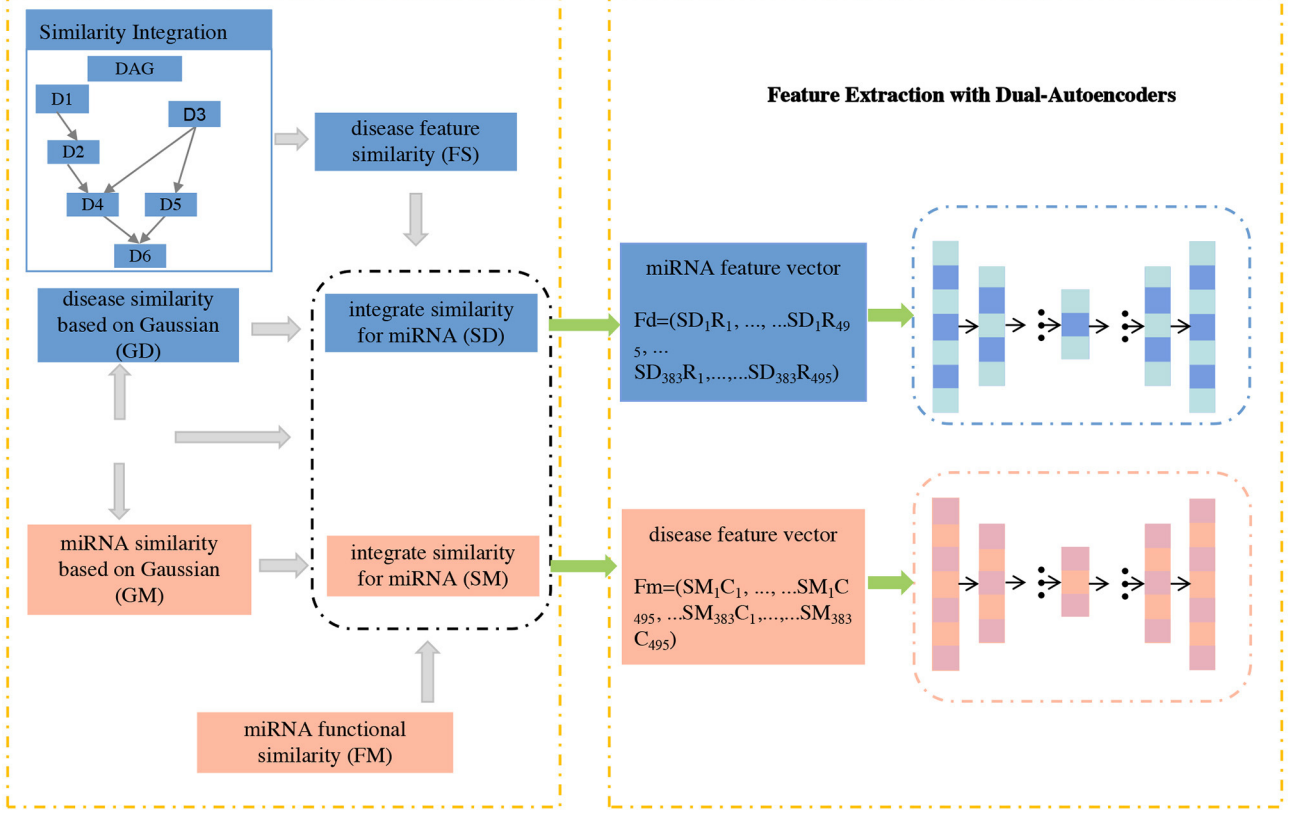


FIGURE 1
The concatenated feature representations are input to dual-autoencoders.

represent disease. For the elements in the matrix, the sub-matrix composed of elements from 1 to 495 rows and 496 to 878 columns represents miRNA-disease association. The specific calculation is as follows:

$$A = \begin{pmatrix} N_{MM} & N_{MD} \\ N_{DM} & N_{DD} \end{pmatrix}. \tag{19}$$

In the above equation, the size of the adjacency matrix A is 878 × 878. N_{MD} and N_{DM} represent miRNA-disease association, and N_{DD} and N_{MM} are set to 0. In GCN, the feature matrix F obtained in the previous section is fed into the GCN network as the initial node embedding matrix. Along with it, matrix A participates in GCN. GCN can aggregate nodes based on topology information to obtain more effective node embedding. The node embedding aggregation calculation is as follows:

$$H_{i+1} = \sigma(\hat{\Gamma}^{-\frac{1}{2}}\hat{A}\hat{\Gamma}^{-\frac{1}{2}}H_i W_i), \tag{20}$$

where H_i represents the node embedding of the i-th layer, H_0 comes from Fd or Fm. $\hat{A}$ represents the adjacency matrix with self-loops, and $\hat{\Gamma}$ represents the degree matrix of $\hat{A}$, W_i represents the trainable matrix. In this study, we design a 2-layer GCN to predict miRNA-disease associations as shown in Figure 2.

3. Results

In this section, our model compares the performance of several typical models on the HMDD dataset. In order to verify the reliability of the model, we also conducted 5-fold and 10-fold cross-validation experiments. At the same time, to demonstrate that the proposed model has certain practical significance, such as preliminary prevention and guidance for diseases, we also constructed corresponding case studies for certain diseases.

3.1. Evaluation strategy

We used common AUC and precision metrics to validate the performance of our model. Among them, AUC is a comprehensive indicator, which can reflect the comprehensive performance of the model. Since the sparse rate in the dataset is $((495X383) - 5430) \div (495X383) \approx 97.14\%$, in other words, the number of negative samples is far more than that of positive samples. However, from a practical point of view, we need to pay more attention to the performance of the model in the positive sample. Therefore, we use Precision to evaluate the performance of the model. Its calculation formula is as follows:

$$Precision = \frac{True\ Positive\ rate}{True\ Positive\ rate + False\ Negative\ rate}. \tag{21}$$

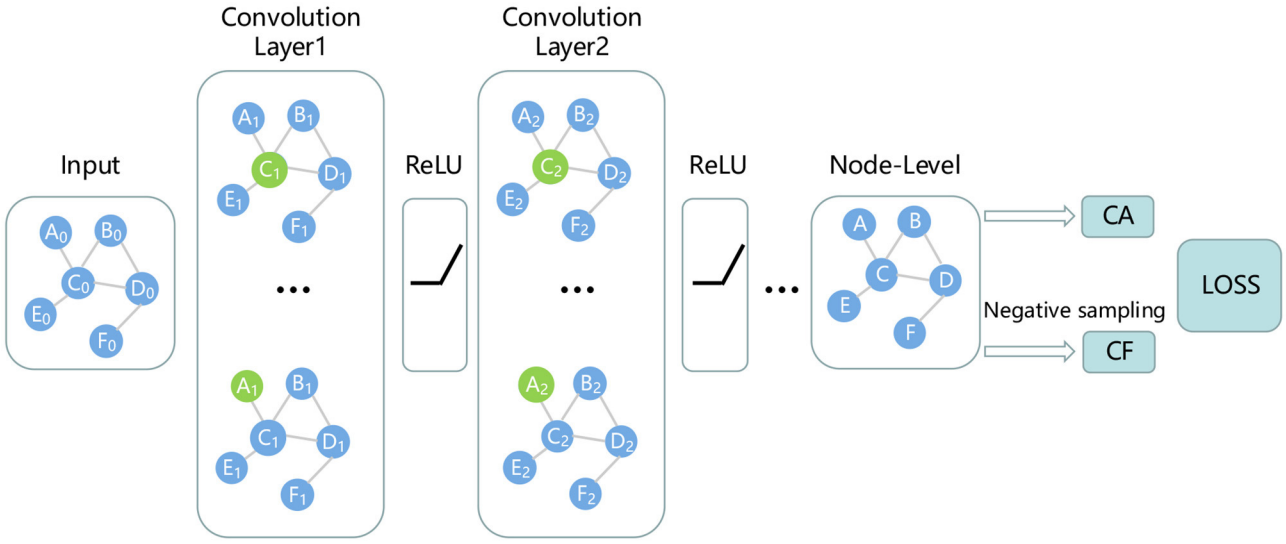


FIGURE 2
Prediction of miRNA-disease associations using GCN.

TABLE 1 Precision of six methods in miRNA-disease classification task.

Models	Precision (%)
RFMDA (Chen et al., 2018b)	62.53
LMTRDA (Wang et al., 2019)	80.13
ABMDA (Zhao et al., 2019)	81.52
GAEMDA (Li et al., 2021)	81.37
GBDT_LR (Zhou et al., 2020)	83.15
GCNA-MDA	87.80

Furthermore, in N-fold cross-validation experiments, we perform N-fold cross-validation by randomly splitting the sample into N equal parts. $N-1$ parts are used as the training set, and the rest are used as the test set. According to this strategy, N parts are used in turn as test sets, and the remaining parts are used as training sets to complete all cross-validation experiments. In the experiment, we consider the AUC metric to measure the performance of the model.

3.2. Comparative evaluation

We compare the GCNA-MDA model with GAEMDA (Li et al., 2021), GBDT_LR (Zhou et al., 2020), ABMDA (Zhao et al., 2019), LMTRDA (Wang et al., 2019), RFMDA (Chen et al., 2018b) models. The GAEMDA (Li et al., 2021) model fuses similarity information and topological neighborhood information in the miRNA-disease network, and integrates GCN and autoencoder for prediction tasks. GBDT_LR (Zhou et al., 2020), ABMDA (Zhao et al., 2019) and RFMDA (Chen et al., 2018b) use ensemble learning strategies to obtain high-quality features and then make corresponding predictions. Besides, GBDT_LR (Zhou et al., 2020), ABMDA (Zhao et al., 2019) used a new negative sample collection strategy to weaken the impact of negative sample coverage. LMTRDA (Wang et al., 2019) combined multi-way data for prediction tasks. Table 1 lists the results of performance comparison, indicating that the GCNA-MDA model obtains the highest Precision value of 89.82%. Our model fully incorporates multi-level information, while applying a dual-autoencoder to further refine the features. Meanwhile, we applies GCN to predict miRNA-disease associations, making the good use of topological information. Combining the above two reasons, our model has achieved the best accuracy results.

For the compared models, RFMDA (Chen et al., 2018b) achieves the worst performance. The main reason is attributed that although the model adopts the strategy of integrated learning, RFMDA (Chen et al., 2018b) does not consider the skew caused by excessive negative samples and it does not synthesize information from multiple sources. While the rest of the models employing multiple information significantly outperform the RFMDA (Chen et al., 2018b) model, which exhibits the importance of integrating multiple information. In addition, GBDT_LR (Zhou et al., 2020) combined with ABMDA (Zhao et al., 2019) applied the strategy of ensemble learning and weakening negative samples, resulting in a significant performance improvement.

3.3. Scalability evaluation

To measure the scalability of the GCNA-MDA model, we perform 5- and 10-fold cross-validation on the HMDD dataset. The results of 5-fold cross-validation are shown in Figure 3. The GCNA-MDA model achieved AUC values of 0.867, 0.878, 0.875, 0.878, and 0.867 in five experiments. The average of 5 AUCs is 0.8730, and the standard deviation is 0.00526. This shows that our model has good scalability and its performance is not easily affected by

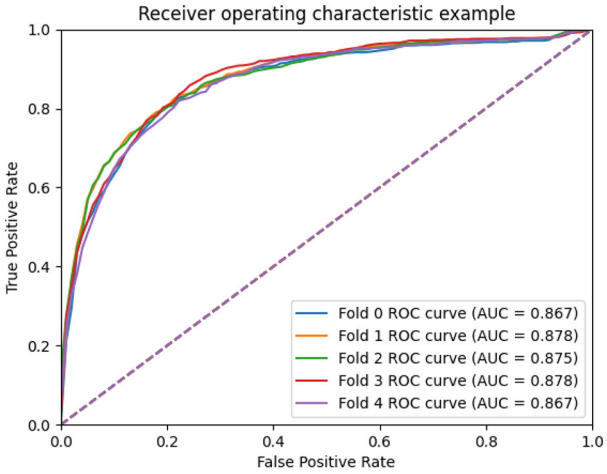


FIGURE 3
AUC performance of GCNA-MDA model on 5-fold cross-validation.

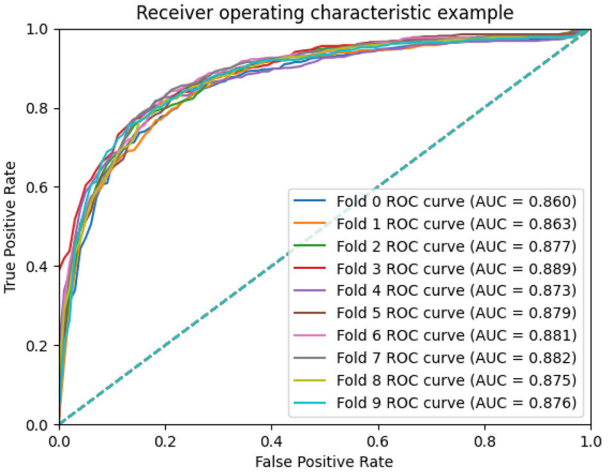


FIGURE 4
AUC performance of GCNA-MDA model on 10-fold cross-validation.

random factors. In order to further eliminate the interference of other factors, our GCNA-MDA model was subjected to a 10-fold cross-validation experiment on the HMDD dataset. Figure 4 shows the AUC performance of 10-fold cross-validation. The GCNA-MDA model achieved AUC values of 0.860, 0.863, 0.877, 0.889, 0.873, 0.879, 0.881, 0.882, 0.875, and 0.876 in 10 experiments. It can be calculated that the average value of the AUC indicator is 0.8755, and the standard deviation is 0.00561. We can find that there is only a difference of 0.0003 between the means of the two groups of experiments, and a difference of 0.00338 between the standard deviations of the two groups. Such variance is perfectly acceptable because random sampling is not controllable. It shows that the performance of the GCNA-MDA model is very stable, and it also shows that its accuracy will not be affected by random sampling. In addition, this may also be due to the local sampling strategy adopted in our research, so that the distribution and ratio of positive and negative samples tend to be similar at the same time.

3.4. Evaluation of different forecasting methods

Table 2 compares the performance of two autoencoder-based methods. The DFELMDA model (Liu et al., 2022b) employs autoencoders for feature extraction and random forests for miRNA-disease association prediction. While it performs well on the AUC indicator, its performance on other indicators is unsatisfactory, possibly due to overfitting caused by random forests. Moreover, the extreme imbalance of positive and negative samples further contributes to the low indicators. In contrast, the GCNA-MDA model performs consistently across all indicators, likely because it utilizes GCN in the prediction module, which effectively incorporates topological information. Additionally, we address the issue of imbalanced samples by maintaining a 1:1 ratio of positive and negative samples.

TABLE 2 Performance comparison of two models using autoencoders (%).

Models	AUC	AUPR	MCC	F1-score	Precision
GCNA-MDA	86.66	86.80	55.90	73.97	85.78
	87.80	88.42	58.33	73.61	90.24
	87.54	88.60	58.77	74.57	89.33
	87.75	87.99	57.19	75.49	85.19
	86.73	87.23	53.51	69.86	88.43
Average	87.30	87.81	56.74	73.50	87.80
DFELMDA (Liu et al., 2022b)	95.56	58.49	13.17	14.23	20.57

3.5. Case analysis

In order to verify the validity of our model, we conduct case analysis of 10 related diseases on the miRNA numbered hsa-mir-29a. In a more detailed operation, we selected the best model parameters in a 5-fold cross-validation experiment, and then selected these diseases in Table 3 as an external test set to predict the association with hsa-mir-29a. We picked 7 positive samples associated with hsa-mir-29a and 3 negative samples not associated with hsa-mir-29a. Table 3 presents the results of the case analysis. By comparing the results in the original database, the GCNA-MDA model correctly predicted all associations in the case analysis. This shows that the GCNA-MDA model does have certain reliability and can be further used as a reference for disease prediction.

We also performed a case analysis of the model on the disease side. For instance, we analyzed miRNAs potentially associated with Renal Cell-related cancer. Table 4 presents the analysis results, indicating that the GCNA-MDA model accurately identifies miRNAs associated with the disease by comparing databases. Thus, our model is effective for case studies involving both miRNAs and diseases.

TABLE 3 A case study of the association of miRNA named hsa-mir-29a with various diseases.

Diseases	Predicted	Diseases	Predicted
Carcinoma, hepatocellular	Verified	Heart failure	Verified
Liver neoplasms	verified	Cerebral infarction	Unverified
Influenza, human	verified	Colonic neoplasms	Verified
Scleroderma, localized	Verified	Gerstmann-Straussler-Scheinker disease	Verified
Skin neoplasms	Unverified	Carcinoma, Small cell	Unverified

TABLE 4 A case study of the association of disease named Carcinoma, Renal Cell with various miRNAs.

miRNAs	Predicted	miRNAs	Predicted
hsa-mir-132	Verified	hsa-mir-1303	Verified
hsa-mir-378b	Verified	hsa-mir-378e	Verified
hsa-mir-141	Verified	hsa-mir-218	Verified
hsa-mir-19b	Verified	hsa-mir-196b	Unverified
hsa-mir-498	Unverified	hsa-mir-3196	Verified

4. Conclusion

In this paper, a GCNA-MDA model that accurately predicts miRNA-disease associations is proposed based on dual autoencoders and GCN. We proposed a novel feature integration strategy based on the combination of multi-way data such as association similarity and feature similarity. This allows for a more complete initial representation of the node. Furthermore, we further perform feature extraction on these initial node representations with higher dimensions based on the dual-autoencoder. The self-supervised learning strategy alleviates the problem of insufficient positively correlated data, resulting in a more robust initial node embedding matrix. Finally, based on GCN, we perform corresponding aggregation operations on all miRNAs and disease nodes, and perform association prediction tasks. We constructed comparative experiments and scalability experiments to verify the effectiveness and scalability of our model. The case analysis of hsa-mir-29a shows that the GCNA-MDA model has certain practical significance.

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. Our data and code are available at https://github.com/Lqingquan/GCNA-MDA.

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

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

References

Ambros, V. (2003). Microrna pathways in flies and worms: growth, death, fat, stress, and timing. *Cell* 113, 673–676. doi: 10.1016/S0092-8674(03)00428-8

Cai, L., Ren, X., Fu, X., Peng, L., Gao, M., and Zeng, X. (2021a). ienhancer-xg: interpretable sequence-based enhancers and their strength predictor. *Bioinformatics* 37, 1060–1067. doi: 10.1093/bioinformatics/btaa914

Cai, L., Wang, L., Fu, X., Xia, C., Zeng, X., and Zou, Q. (2021b). Itp-pred: an interpretable method for predicting, therapeutic peptides with fused features low-dimension representation. *Briefings Bioinformat.* 22, bbaa367. doi: 10.1093/bib/bbaa367

Chen, X. (2021). Deep-belief network for predicting potential mirna-disease associations. *Briefing Bioinformat.* 22, bbaa186. doi: 10.1093/bib/bbaa186

Chen, X., Clarence Yan, C., Zhang, X., Li, Z., Deng, L., Zhang, Y., et al. (2015). Rbmmmda: predicting multiple types of disease-microrna associations. *Sci. Rep.* 5, 13877. doi: 10.1038/srep13877

Chen, X., Huang, L., Xie, D., and Zhao, Q. (2018a). Egbmmda: extreme gradient boosting machine for mirna-disease association prediction. *Cell Death Dis.* 9, 3. doi: 10.1038/s41419-017-0003-x

Chen, X., Liu, M.-X., and Yan, G.-Y. (2012). Rwrmda: predicting novel human microrna–disease associations. *Mol. Biosyst.* 8, 2792–2798. doi: 10.1039/c2mb25180a

Chen, X., Sun, L.-G., and Zhao, Y. (2021). Ncmcmda: mirna–disease association prediction through neighborhood constraint matrix completion. *Briefings Bioinformat.* 22, 485–496. doi: 10.1093/bib/bbz159

Chen, X., Wang, C.-C., Yin, J., and You, Z.-H. (2018b). Novel human mirna-disease association inference based on random forest. *Mol. Ther. Nucleic Acids* 13:568–579. doi: 10.1016/j.omtn.2018.10.005

Chen, X., Xie, D., Zhao, Q., and You, Z.-H. (2019a). Micrornas and complex diseases: from experimental results to computational models. *Briefings Bioinformat.* 20, 515–539. doi: 10.1093/bib/bbx130

Chen, X., Yan, C. C., Zhang, X., You, Z.-H., Deng, L., Liu, Y., et al. (2016). Wbsmda: within and between score for mirna-disease association prediction. *Sci. Rep.* 6, 1–9. doi: 10.1038/srep21106

Chen, X., and Yan, G.-Y. (2013). Novel human lncrna–disease association inference based on lncrna expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

Chen, X., and Yan, G.-Y. (2014). Semi-supervised learning for potential human microrna-disease associations inference. *Sci. Rep.* 4, 5501. doi: 10.1038/srep05501

Chen, X., Yin, J., Qu, J., and Huang, L. (2018c). Mdhgi: matrix decomposition and heterogeneous graph inference for mirna-disease association

prediction. *PLoS Comput. Biol.* 14, e1006418. doi: 10.1371/journal.pcbi.1006418

Chen, X., Zhu, C.-C., and Yin, J. (2018d). Predicting mirna-diseaseassociation based on inductive matrix completion. *Bioinformatics* 34, 4256–4265. doi: 10.1093/bioinformatics/bty503

Chen, X., Zhu, C.-C., and Yin, J. (2019b). Ensemble of decision tree reveals potential mirna-disease associations. *PLoS Comput. Biol.* 15, e1007209. doi: 10.1371/journal.pcbi.1007209

Chen, X. Y. C., and Zhang, X. (2014). Hgimda: heterogeneous graphinference for mirna-disease association prediction. *Oncotarget* 7(10):65257–65269. doi: 10.18632/oncotarget.11251

Fu, X., Cai, L., Zeng, X., and Zou, Q. (2020). Stackcppred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. *Bioinformatics* 36, 3028–3034. doi: 10.1093/bioinformatics/btaa131

Guang, H. (2018). Predicting microrna-disease associations using label propagation based on linear neighborhood similarity. *J. Biomed. Informat.* 82, 169–177. doi: 10.1016/j.jbi.2018.05.005

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). Hlpi-ensemble: prediction of human lncrna-protein interactions based on ensemble strategy. *RNA Biol.* 15, 797–806. doi: 10.1080/15476286.2018.1457935

Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., et al. (2010). Prioritization of disease microrna through a human phenome-microrna network. *BMC Syst. Biol.* 4, 1–9. doi: 10.1186/1752-0509-4-S1-S2

Jopling, C. L., Yi, M., Lancaster, A. M., Lemon, S. M., and Sarnow, P. (2005). Modulation of hepatitis c virus rna abundance by a liver-specific microrna. *Science* 309, 1577–1581. doi: 10.1126/science.1113329

Jung, Baltimore David, T. K. D., P, B. M., and Kuang, C. (2006). *NF-KappaB-Dependent Induction of microRNA miR-146, an Inhibitor Targeted to Signaling Proteins of Innate Immune Responses* (Thesis).

Karp, X., and Ambros, V. (2005). Encountering microrna in cell fate signaling. *Science* 310, 1288–1289. doi: 10.1126/science.1121566

Li, X., Xu, J., and Li, Y. (2012). Prioritizing candidate disease mirnas by topological features in the mirna-target dysregulated network. *Syst. Biol. Cancer Res. Drug Discov.* 2012, 289–306. doi: 10.1007/978-94-007-4819-4_12

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2014). Hmdd v2. 0: a database for experimentally supported human microrna and disease associations. *Nucleic Acids Res.* 42, D1070–D1074. doi: 10.1093/nar/gkt1023

Li, Z., Li, J., Nie, R., You, Z.-H., and Bao, W. (2021). A graph auto-encoder model for mirna-disease associations prediction. *Briefings Bioinformat.* 22, bbaa240. doi: 10.1093/bib/bbaa240

Liu, W., Jiang, Y., Peng, L., Sun, X., Gan, W., Zhao, Q., et al. (2022a). Inferring gene regulatory networks using the improved markov blanket discovery algorithm. *Interdiscipl. Sci. Computat. Life Sci.* 2022, 1–14. doi: 10.1007/s12539-021-00478-9

Liu, W., Lin, H., Huang, L., Peng, L., Tang, T., Zhao, Q., et al. (2022b). Identification of mirna–disease associations via deep forest ensemble learning based on autoencoder. *Briefings Bioinformat.* 23, bbac104. doi: 10.1093/bib/bbac104

Liu, W., Sun, X., Yang, L., Li, K., Yang, Y., and Fu, X. (2022c). Nscgrn: a network structure control method for gene regulatory network inference. *Briefings Bioinformat.* 23,bbac156. doi: 10.1093/bib/bbac156

Meng, F., Henson, R., Wehbe–Janek, H., Ghoshal, K., Jacob, S. T., and Patel, T. (2007). Microrna-21 regulates expression of the pten tumor suppressor gene in human hepatocellular cancer. *Gastroenterology* 133, 647–658. doi: 10.1053/j.gastro.2007.05.022

Miska, E. A. (2005). How micrornas control cell division, differentiation and death. *Curr. Opin. Genet. Dev.* 15, 563–568. doi: 10.1016/j.gde.2005.08.005

Mugunga, I., Ju, Y., Liu, X., and Huang, X. (2017). Computational prediction of human disease-related micrornas by path-based random walk. *Oncotarget* 8, 58526. doi: 10.18632/oncotarget.17226

Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., et al. (2019). A learning-based framework for mirna-disease association identification using neural networks. *Bioinformatics* 35, 4364–4371. doi: 10.1093/bioinformatics/btz254

Peng, L., Wang, C., Tian, G., Liu, G., Li, G., Lu, Y., et al. (2022a). Analysis of ct scan images for covid-19 pneumonia based on a deep ensemble framework with densenet, swin transformer, and regnet. *Front. Microbiol.* 13, 995323. doi: 10.3389/fmicb.2022.995323

Peng, L., Wang, F., Wang, Z., Tan, J., Huang, L., Tian, X., et al. (2022b). Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies. *Briefings Bioinformat.* 23, bbac234. doi: 10.1093/bib/bbac234

Peng, L., Yang, C., Huang, L., Chen, X., Fu, X., and Liu, W. (2022c). Rnmflp: predicting circrna–disease associations based on robust nonnegative matrix factorization and label propagation. *Briefings Bioinformat.* 23, bbac155. doi: 10.1093/bib/bbac155

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., et al. (2012). Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res.* 40, D940–D946. doi: 10.1093/nar/gkr972

Tian, G., Wang, Z., Wang, C., Chen, J., Liu, G., Xu, H., et al. (2022). A deep ensemble learning-based automated detection of covid-19 using lung ct images and vision transformer and convnext. *Front. Microbiol.* 13, 1024104. doi: 10.3389/fmicb.2022.1024104

Van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

Vasudevan, S., Tong, Y., and Steitz, J. A. (2007). Switching from repression to activation: micrornas can up-regulate translation. *Science* 318, 1931–1934. doi: 10.1126/science.1149460

Vickers, K. C., Rye, K.-A., and Tabet, F. (2014). Micrornas in the onset and development of cardiovascular disease. *Clin. Sci.* 126, 183–194. doi: 10.1042/CS20130203

Wang, C.-C., Han, C.-D., Zhao, Q., and Chen, X. (2021). Circular rnas and complex diseases: from experimental results to computational models. *Briefings Bioinformat.* 22, bbab286. doi: 10.1093/bib/bbab286

Wang, L., You, Z.-H., Chen, X., Li, Y.-M., Dong, Y.-N., Li, L.-P., et al. (2019). Lmtrda: Using logistic model tree to predict mirna-disease associations by fusing multi-source information of sequences and similarities. *PLoS Comput. Biol.* 15, e1006865. doi: 10.1371/journal.pcbi.1006865

Xu, J., Cai, L., Liao, B., Zhu, W., Wang, P., Meng, Y., et al. (2019). Identifying potential mirnas–disease associations with probability matrix factorization. *Front. Genet.* 10, 1234. doi: 10.3389/fgene.2019.01234

Xu, J., Xu, J., Meng, Y., Lu, C., Cai, L., Zeng, X., et al. (2023). Graph embedding and gaussian mixture variational autoencoder network for end-to-end analysis of single-cell rna sequencing data. *Cell Rep. Methods* 2023, 100382. doi: 10.1016/j.crmeth.2022.100382

You, Z.-H., Huang, Z.-A., Zhu, Z., Yan, G.-Y., Li, Z.-W., Wen, Z., et al. (2017). Pbmda: A novel and effective path-based computational model for mirna-disease association prediction. *PLoS Computat. Biol.* 13, e1005455. doi: 10.1371/journal.pcbi.1005455

Zeng, X. (2017). Inferring microrna-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE-ACM Transact. Comput. Biol. Bioinformat.* 14, 905–915. doi: 10.1109/TCBB.2016.2550432

Zeng, X., Liu, L., Lü, L., and Zou, Q. (2018). Prediction of potential disease-associated micrornas using structural perturbation method. *Bioinformatics* 34, 2425–2432. doi: 10.1101/223693

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021). Using network distance analysis to predict lncrna–mirna interactions. *Interdiscipl. Sci. Comput. Life Sci.* 13, 535–545. doi: 10.1007/s12539-021-00458-z

Zhang, Z., Xu, J., Wu, Y., Liu, N., Wang, Y., and Liang, Y. (2023). Capsnet-lda: predicting lncrna-disease associations using attention mechanism and capsule network based on multi-view data. *Briefings Bioinformat.* 24, bbac531. doi: 10.1093/bib/bbac531

Zhao, Y., Chen, X., and Yin, J. (2019). Adaptive boosting-based computational model for predicting potential mirna-disease associations. *Bioinformatics* 35, 4730–4738. doi: 10.1093/bioinformatics/btz297

Zhou, S., Wang, S., Wu, Q., Azim, R., and Li, W. (2020). Predicting potential mirna-disease associations by combining gradient boosting decision tree with logistic regression. *Comput. Biol. Chem.* 85, 107200. doi: 10.1016/j.compbiolchem.2020.107200

Zhu, C.-C., Wang, C.-C., Zhao, Y., Zuo, M., and Chen, X. (2021). Identification of mirna–disease associations via multiple information integration with bayesian ranking. *Briefings Bioinformat.* 22, bbab302. doi: 10.1093/bib/bbab302

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 22 May 2023
DOI 10.3389/fmicb.2023.1172340

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Zhenting Xiang,
University of Pennsylvania, United States
Peter Allan Jorth,
Cedars-Sinai Medical Center, United States

*CORRESPONDENCE
Tülay Yucel-Lindberg
✉ tulay.lindberg@ki.se

†These authors have contributed equally to this work

RECEIVED 23 February 2023
ACCEPTED 02 May 2023
PUBLISHED 22 May 2023

CITATION
Narayanan A, Söder B, Meurman J, Lundmark A, Hu YOO, Neogi U and Yucel-Lindberg T (2023) Composition of subgingival microbiota associated with periodontitis and diagnosis of malignancy—a cross-sectional study. *Front. Microbiol.* 14:1172340. doi: 10.3389/fmicb.2023.1172340

Composition of subgingival microbiota associated with periodontitis and diagnosis of malignancy—a cross-sectional study

Aswathy Narayanan[1,2†], Birgitta Söder[3†], Jukka Meurman[4], Anna Lundmark[5], Yue O. O. Hu[6,7], Ujjwal Neogi[8] and Tülay Yucel-Lindberg[5*]

[1]Division of Clinical Microbiology, Department of Laboratory Medicine, ANA Futura, Karolinska Institutet, Stockholm, Sweden, [2]Division of Infectious Diseases, Department of Medicine Huddinge, Karolinska Institutet, Stockholm, Sweden, [3]Division of Periodontology, Department of Dental Medicine, Karolinska Institutet, Huddinge, Sweden, [4]Department of Oral and Maxillofacial Diseases, University of Helsinki and Helsinki University Hospital, Helsinki, Finland, [5]Division of Pediatric Dentistry, Department of Dental Medicine, Karolinska Institutet, Huddinge, Sweden, [6]Department of Microbiology, Tumor and Cell Biology, Centre for Translational Microbiome Research, Karolinska Institutet, Stockholm, Sweden, [7]School of Environmental Science and Engineering, Hubei Polytechnic University, Huangshi, China, [8]The Systems Virology Lab, Division of Clinical Microbiology, Department of Laboratory Medicine, ANA Futura, Karolinska Institutet, Stockholm, Sweden

Periodontitis is one of the world's most prevalent infectious conditions, affecting between 25 and 40% of the adult population. It is a consequence of the complex interactions between periodontal pathogens and their products, which trigger the host inflammatory response, chronic inflammation, and tissue destruction. Chronic systemic low-grade inflammation is involved in numerous diseases, and it is also known that long-lasting inflammation and chronic infections predispose one to cancer. Here, we characterized and compared the subgingival microbiota associated with periodontitis and diagnosis of malignancy in a longitudinal 10-year follow-up study. The study was conducted on 50 patients with periodontitis and 40 periodontally healthy individuals. The recorded clinical oral health parameters were periodontal attachment loss (AL), bleeding on probing (BOP), gingival index (GI), probing depth (PD), and plaque index (PI). Subgingival plaque was collected from each participant, from which DNA was extracted, and 16S rRNA gene amplicon sequencing performed. Cancer diagnoses data were collected between the years 2008–2018 from the Swedish Cancer Registry. The participants were categorized based on having cancer at the time of sample collection (CSC), having developed cancer later (DCL), and controls without any cancer. The most abundant phyla across all 90 samples were *Actinobacteria*, *Proteobacteria*, *Firmicutes*, *Bacteroidetes*, and *Fusobacteria*. At the genus level, *Treponema*, *Fretibacterium*, and *Prevotella* were significantly more abundant in samples of periodontitis patients compared to non-periodontitis individuals. With regard to samples of cancer patients, *Corynebacterium* and *Streptococcus* were more abundant in the CSC group; *Prevotella* were more abundant in the DCL group; and *Rothia*, *Neisseria*, and *Capnocytophaga* were more abundant in the control group. In the CSC group, we also found that the presence of periodontal inflammation, in terms of BOP, GI, and PLI, significantly correlated with species belonging to the genera *Prevotella*, *Treponema*, and *Mycoplasma*. Our results revealed that several subgingival genera were differentially enriched among the

studied groups. These findings underscore the need for further research to fully understand the role that oral pathogens may play in the development of cancer.

KEYWORDS

periodontitis, supragingival plaque, cancer, malignancy, 16S rRNA gene sequencing, oral microbiota

Introduction

The oral cavity harbors thousands of different microbial species that can be found on soft tissue and teeth forming biofilms, or communities of microorganisms attached to a surface (Keijser et al., 2008). The most prevalent oral biofilm, dental plaque, exists on tooth surfaces in the form of complex multispecies communities. As the biofilm matures and develops, there is also a gradual shift from Gram-positive aerobic bacteria towards Gram-negative and anaerobic species, affecting the gingival environment with respect to pH and oxygen levels, which promotes species favored by this milieu (Asikainen and Chen, 1999; O'Toole et al., 2000). In addition, the inflammatory response from the host can enrich the environment with inflammatory mediators that enhance the growth of certain "inflammophilic" bacteria, which feed off inflammatory products (Hajishengallis, 2014). Such inflammation is generally resolved in normal healing processes, whereas insufficient resolution results in neutrophil-mediated chronic inflammation and destruction of tissue and bone structures (Serhan, 2014). Chronic inflammation involves several diseases, such as rheumatoid arthritis, periodontitis, type 2 diabetes mellitus, and cardiovascular disease. It is also known that long-lasting inflammation, secondary to chronic infections or infectious agents, predisposes one to cancer development (Coussens and Werb, 2002; de Martel et al., 2012; Garrett, 2015).

Periodontal disease (periodontitis) is a major cause of tooth loss in adults and one of the world's most prevalent chronic infectious inflammatory diseases, affecting up to 25–40% of the adult population. The most severe form of the disease affects 5–15% of the global population (Page and Eke, 2007; Dye, 2012; Eke et al., 2015). Periodontitis is characterized by the destruction of tooth-supporting tissue and bone, which may ultimately result in tooth loss. The disease results from the complex interactions between periodontal microorganisms and their products, triggering the host inflammatory response. The process is initiated when a biofilm forms near the gingiva and releases various substances, such as lipopolysaccharides, peptidoglycans, and toxins, which elicit a host response (Page and Kornman, 1997; Pollanen et al., 2012; Yucel-Lindberg and Bage, 2013). The "red complex" bacteria comprising *Porphyromonas gingivalis (P. gingivalis)*, *Treponema denticola*, and *Tannerella forsythia* has long been associated with the disease, but this view has changed with the emergence of new technologies towards a model where periodontitis is associated with a shift in the whole microbial composition rather than focusing on individual microbial species. As a consequence of bacterial challenge, the host immune response initiates the activation and stimulation of pro-inflammatory cytokines, chemokines, prostaglandins, toll-like receptors, and proteolytic enzymes, collectively contributing to the pathogenesis of periodontitis (Bascones et al., 2005). The expression and/or production of these factors have been demonstrated using gingival tissue biopsies, gingival fluid, and saliva, as well as different types of oral cells (Båge et al., 2011; Davanian et al., 2012; Cavalla et al., 2015). The ongoing "battle" of inflammation is not only measurable locally in the oral samples but also systemically, as increased levels of inflammatory mediators have been demonstrated in the blood of patients with oral diseases, particularly in those with periodontitis (Van Dyke, 2009; Hajishengallis and Chavakis, 2021).

Chronic inflammatory conditions associated with infections may lead to environments that promote genomic lesions and the initiation of tumors. Previous studies have reported an association between periodontitis and an increased risk of total cancer (Romandini et al., 2021; Kim et al., 2022). One meta-regression analysis based on seven case–control studies showed a statistically higher risk of oral cancer with increasing number of missing teeth, with the latter considered a proxy for chronic dental/oral infections (Virtanen et al., 2014). A systematic review and meta-analysis performed recently demonstrated that periodontal disease significantly increases the risk of colorectal cancer by 44% (Li et al., 2021). Additionally, it has been reported that oral squamous cell carcinoma, representing 95% of oral malignancies, is associated with alterations in the oral microbiome. Several studies have linked oral microbiota and periodontal pathogens to head and neck cancer, pancreatic cancers, and colorectal cancer (Ahn et al., 2012; Michaud, 2013; Flemer et al., 2018; Irfan et al., 2020). For example, both *in vivo* and *in vitro* studies have suggested that the key periodontal pathogen *P. gingivalis* contributes to oral carcinogenesis (Groeger et al., 2011; Gallimidi et al., 2015). In contrast, it was reported (Soder et al., 2021) that *P. gingivalis* and *Prevotella intermedia* were more prevalent among subjects without malignancy, whereas the periodontal bacteria *Aggregatibacter actinomycetemcomitans* was strongly associated with malignancy. According to different meta-analyses and reviews (Michaud et al., 2017; Nwizu et al., 2017; Li et al., 2022), the existing data provide support for an association between periodontal disease and risk of different types of cancer including head and neck, lung, colorectal, and pancreatic cancers, although additional research efforts are necessary to further identify the role of oral infections in malignancy. In this study, we aimed to characterize and compare subgingival microbiota associated with periodontitis and the diagnosis of cancer, data extracted from national register of malignancies, in a longitudinal 10-year follow-up study, using 16S rRNA gene sequence analysis.

Materials and methods

Sample collection, DNA extraction and sequencing

A total of 99 individuals divided into two groups, a periodontitis group ($n = 55$) and a non-periodontitis control group ($n = 44$), were

included in the present study. The participants of this study were derived from our Swedish cohort study, which was described in detail previously (Soder et al., 2007). In total, 1,676 participants (838 women and 838 men) were randomly selected from a database registry of all citizens of Stockholm County who were born on the 20th of the month (between years 1945–1954) and underwent an initial oral clinical examination (1985). In 2009, the participants were clinically reexamined for the prevalence of periodontal disease, from which 99 age- and gender-matched subjects with and without periodontitis were enrolled in the study. The included oral clinical parameters were gingival index (GI), pocket depth (PD), bleeding on probing (BOP), clinical attachment loss (CAL), and plaque index (PLI). For each tooth, BOP and CAL were assessed from six different surfaces using a periodontal probe (HU-FRIEDY Perio Probe). The criteria used for the classification of periodontitis were at least one site with PD ≥5 mm, CAL ≥5 mm, and BOP as described previously (Soder et al., 2007; Yakob et al., 2012). Subgingival plaque samples were carefully collected from four sites on each participant, from the second premolar in each quadrant, and stored at −80°C until microbiome analysis.

The diagnoses of malignancy were obtained from the Swedish Cancer Registry included in the registers of the National Board of Health and Welfare, Sweden. For the present study, the 10-year cumulative cancer diagnoses were collected between the years 2009–2018. The cancer cases that had been diagnosed were: orodigestive cancer, breast cancer, prostate cancer, gynaecological cancers, haematological malignancies, head and neck cancers and liver cancer. The study was approved by the Ethics Committee of the Karolinska University Hospital at Huddinge (Dnr 2007/1669-31; 2012/590-32; 2017/2204-32), and all participants gave their informed consent to be included in the study.

DNA extraction, 16S rRNA gene amplification, and sequencing

DNA was extracted from the 99 subgingival plaque samples (pooled together from all sites) using the QIAamp DNA Mini Kit (Qiagen, Valencia, CA, United States) and eluted into 50 µL H_2O. The V3–V4 regions of the bacterial 16S rRNA gene were amplified with 1.0 µM 341′F primer (CCTAHGGGRBGCAGCAG), 1.0 µM 805R primer (GACTACHVGGGTATCTAATCC) (Herlemann et al., 2011), KAPA HotStart ReadyMix (Biosystems, Wilmington, MA, United States), 0.5 ng/µL bovine serum albumin (New England Biolabs, Ipswich, MA, United States), and 2.0 ng of DNA. PCR was performed at 98°C for 2 min followed by 26 cycles of 98°C for 20 s, 54°C for 20 s, and 72°C for 15 s, and a final elongation step of 72°C for 2 min. The samples were purified with polyethylene glycol 6000 (Merck Millipore, Darmstadt, Germany) and carboxylic acid beads (Dynabeads® MyOne™, Thermo Fisher Scientific, Waltham, MA, United States) using the procedure described by Lundin et al. (2010). Thereafter, 12 µL of the amplified and purified product was used for indexing (0.4 µM forward and 0.4 µM reverse indexing primer and KAPA HotStart ReadyMix). The conditions for PCR cycling were 98°C for 2 min followed by 10 cycles of 98°C for 20 s, 62°C for 30 s, and 72°C for 30 s, and a final step of 72°C for 2 min. After amplification, the samples were quantified using a Qubit® 2.0 Fluorometer (Invitrogen, Carlsbad, CA, United States), diluted to 2.0 ng/µL, and pooled before purification by the same procedure as described above. An Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, United States) and a Qubit® 2.0 Fluorometer (Invitrogen, Carlsbad, CA, United States) were used for checking the amplicon fragment sizes and quantification. Equimolar amounts of the indexed samples were mixed and sequenced with Illumina MiSeq (Illumina Inc., San Diego, CA, United States) at the National Genomics Infrastructure/Science for Life Laboratory Stockholm.

After sequencing, nine samples with sequencing reads of less than 10,000 were excluded from further downstream analysis, resulting in a final dataset of 90 samples comprising 50 samples with periodontitis and 40 samples without periodontal disease (non-periodontitis). After the exclusion of low-depth libraries, the median depth of sequencing was 195,300 reads per sample [interquartile range (IQR): 146,700–218,600 reads].

Bioinformatics analysis

The raw paired-end sequences obtained from the Illumina sequencing were first checked for base call quality. The base quality checking was performed using the FastQC tool (Andrews, 2015). The Phred score (Q20) was used as a base quality score threshold for the analysis. Adapters were trimmed using TrimGalore (v0.6.4)[1], and primer sequences were removed with the help of the cutPrimers tool (Kechin et al., 2017). A rarefaction curve was generated to ensure sufficient sequencing depth in order to proceed with further downstream analysis (Supplementary Figure S1). The curves were generated by using R package phyloseq to plot the sequencing depth of the samples vs. the diversity indices, which showed that all the samples had sufficient sequencing depth to capture most of the microbial community, as the curves stabilized after 10,000× coverage.

Amplicon sequence variants estimation, taxonomic classification, and statistical analysis

The pre-processed paired-end sequences were used for further downstream analysis using various bioinformatic tools. First, the pre-processed reads were analyzed using Quantitative Insights into Microbial Ecology version 2 (QIIME2). Amplicon sequence variants (ASVs) generated using QIIME2 were used for functional interpretation of the microbiota (Bolyen et al., 2019).

To visualize the abundance of taxonomy, sample-wise stacked bar plots were constructed at phylum, family, and genus levels using the *ggplot2* (3.2.1) R package. The results were further analyzed with the *phyloseq* (1.28.0) R package to study the alpha and beta diversity of the samples. Alpha diversity was calculated using the *estimate_richness* R function and visualized using the *ggplot2* R package. Beta diversity was estimated using the *ordinate* R function and visualized using the *plot_ordination* R function. The clustering of the samples was presented with a *non-metric multidimensional scaling* (NMDS) plot based on Bray–Curtis distance. Rarefaction analyses were conducted

1 https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

using the *rarefy* (*Vegan* v2.6-2) R function. Permutational multivariate analysis of variance (PERMANOVA) was performed to test for significant differences between the two groups at the genus taxonomic level using the *vegan* (version 2.4-3) R package. The analysis compared the groups and provided the top organisms that were responsible for their differentiation (Anderson, 2017). Statistical tests were performed between NMDS1 and NMDS2 to obtain the significant coordinate, and Welch's *t*-test (two) was used to calculate the *p*-values. Correlation between microbial taxa and periodontal clinical parameters was assessed by the Spearman rank correlation coefficient (significance level $p < 0.05$) using the *psych* v2.2.3 R package. The graphical representation of the results was done using GraphPad Prism v8.4.2, where red indicates a positive correlation and blue indicates a negative correlation.

Results

Subgingival plaque microbiota composition and its relationship with periodontitis

The study cohort comprised 90 participants separated into two groups: the periodontitis group ($n = 50$) and the non-periodontitis group ($n = 40$). The mean age was 58.4 ± 2.7 years for the periodontitis group and 59.8 ± 2.9 years for the non-periodontitis group. We also categorized the patients based on their longitudinal follow-up of 10 years as having cancer ($n = 35$) at the time of sample collection (CSC, $n = 13$), those who developed cancer later (DCL, $n = 22$), and controls who did not have any cancer at the time of sampling but also did not develop cancer during the follow-up period ($n = 55$).

Figure 1A shows the relative abundance distribution of all 90 included plaque samples at the genus level. The most prominent genera (phylum in brackets) across all samples were *Rothia*, *Corynebacterium*, *Actinomyces* (*Actinobacteria*), Neisseria (*Proteobacteria*), *Streptococcus* (*Firmicutes*), *Capnocytophaga*, *Prevotella* (*Bacteroidetes*), and *Leptotrichia* (*Fusobacteria*). Figure 1B shows the Beta diversity of samples visualized using a non-metric multidimensional scaling (NMDS) plot. There were no clear clusters for the periodontitis and non-periodontitis groups. Statistical analyses showed no significant differences for NMDS2 but did show significant differences with NMDS1 ($p < 0.05$) between the periodontitis and non-periodontitis groups (Figure 1B). At the genus level, the samples belonging to the periodontitis and non-periodontitis groups were ordered as per the NMDS1 ordinates to visualize the differences in the bacterial composition between each sample, as shown in Figure 1B. A boxplot of alpha diversity indices with corresponding *p*-values (Supplementary Figure S2) did not show any significant differences between groups.

When comparing the periodontitis and non-periodontitis groups, the phyla *Firmicutes*, *Bacteroidetes* and *Epsilonbacteraeota* were found to be more abundant in individuals having periodontitis, whereas *Proteobacteria* and *Fusobacteria* were more abundant in the non-periodontitis group. The phylum *Actinobacteria* was found at similar levels in both the periodontitis and non-periodontitis groups (Figure 2A). The abundances of phyla *Firmicutes*, *Bacteroidetes*, *Proteobacteria Fusobacteria* and *Epsilonbacteraeota* were not significantly different between the groups. However, the abundances of *Spirochaetes* and *Synergistetes* were significantly different between the groups ($p = 0.01$ and $p = 0.023$, respectively).

At the family level, the most abundant bacteria were *Actinomycetaceae*, *Cardiobacteriaceae*, *Flavobacteriaceae*, *Neisseriaceae*, *Prevotellaceae*, *Veillonellaceae*, and *Streptococcaceae* (Figure 2B). The abundances of *Pasteurellaceae*, *Spirochaetaceae*, *Synergistaceae*, and *Carnobacteriaceae* were significantly differentially abundant between the two groups ($p = 0.004$, $p = 0.01$, $p = 0.013$, $p = 0.023$, and $p = 0.032$, respectively). The most abundant bacteria at the genus level were *Actinomyces*, *Corynebacterium*, *Neisseria*, *Prevotella*, *Streptococcus*, and *Rothia* (Figures 2C,D). When comparing the periodontitis vs. non-periodontitis samples, the abundances of *Haemophilus*, *Treponema*, *Fretibacterium*, *Granulicatella*, and *Prevotella* were significantly different between the groups ($p = 0.01$, $p = 0.013$, $p = 0.023$, $p = 0.03$, and $p = 0.031$, respectively) (Figure 2C). Of these, *Treponema*, *Fretibacterium*, and *Prevotella* were significantly more abundant in samples of periodontitis patients compared to non-periodontitis individuals.

A PERMANOVA analysis was performed to test for differences between the two groups. When comparing the oral microbial composition between periodontitis and non-periodontitis individuals, at the genus level, the periodontitis microbiome had a high abundance of *Prevotella*, *Campylobacter*, and *Treponema*, whereas the non-periodontitis samples had *Rothia*, *Haemophilus*, and *Capnocytophaga* as the top-three most abundant genera (Figure 3). However, there were no significant differences in overall microbial composition between the two groups ($p = 0.27$).

Subgingival microbiota composition and its association with cancer

Next, we categorized the patients included in this study based on their longitudinal follow-up of 10 years as having cancer (CSC), developed cancer later (DCL), and controls (Figure 4). The most abundant bacteria at the phylum level were *Actinobacteria*, *Proteobacteria*, *Firmicutes*, and *Bacteroidetes*. A comparison of the three groups revealed that *Actinobacteria* were more abundant in the non-cancer control group, as were *Firmicutes* in the CSC group, whereas *Proteobacteria* were enriched in the DCL group (Figure 4A). The abundant phyla *Firmicutes* and *Bacteroidetes* were not statistically significant between among all groups.

At the family level, *Streptococcaceae* and *Corynebacteriaceae* were enriched in the CSC group, whereas *Flavobacteriaceae*, *Micrococcaceae* and *Neisseriaceae* were more abundant in the control group, and *Leptotrichiaceae* were more highly abundant in the DCL group (Figure 4B). Moreover, *Paludibacteraceae* was found to be significantly more enriched in the CSC group compared to both the control and DCL groups ($p = 0.019$ and $p = 0.02$, respectively). At the genus level, *Streptococcus*, *Corynebacterium* and *Fusobacterium* were more abundant in the CSC group; *Neisseria*, *Rothia*, and *Capnocytophaga* were abundant in the control group; and *Prevotella* were more abundant in the DCL group (Figures 4C,D). The genus *Paludibacteraceae-F0058* was significantly ($p = 0.02$) enriched in the CSC group compared to both the controls and the DCL group.

A PERMANOVA analysis was performed to compare the subgingival microbiota between the two different cancer groups (CSC and DCL) and the control group. The results revealed that *Streptococcus*, *Corynebacterium*, and *Fusobacterium* were enriched in

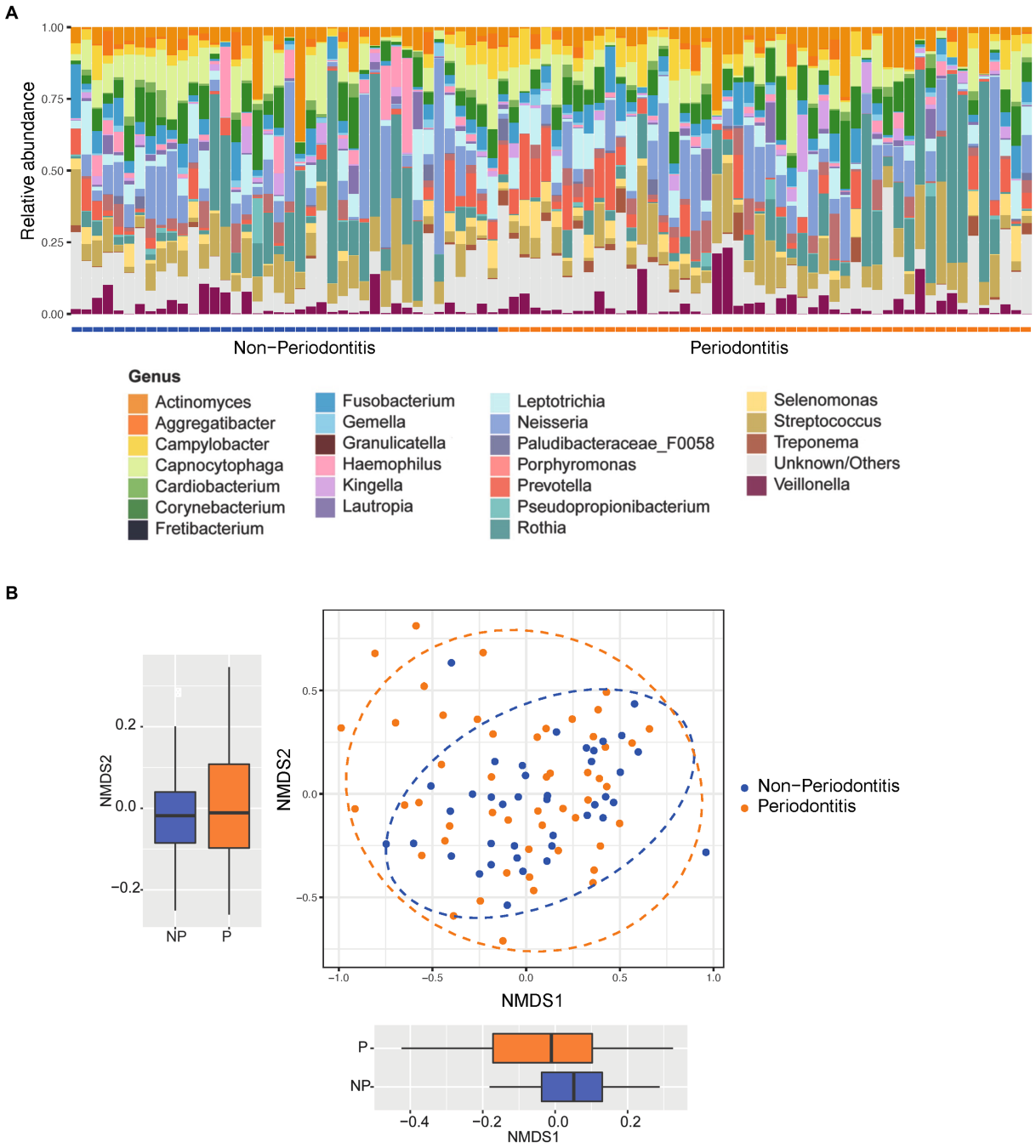


FIGURE 1
Relative abundance and beta diversity between groups. **(A)** Relative abundance of the most abundant organisms of all the 90 samples at the genus level grouped according to disease status (i.e., periodontitis or non-periodontitis). **(B)** Difference in beta diversity between the groups represented as non-metric multidimensional scaling (NMDS) ordination plots using Bray–Curtis distances; the separations between groups at each axis are seen in respective boxplot. The boxplots represent the median (horizontal black line), 25th and 75th quartiles (box edge), and upper and lower ends (whiskers).

the CSC group, whereas *Rothia*, *Neisseria*, and *Actinomyces* were enriched in the control group (Figure 5A). Furthermore, *Leptotrichia*, *Streptococcus*, and *Haemophilus* were the top-three most abundant genera in the DCL group, whereas *Rothia*, *Capnocytophaga*, and *Campylobacter* were more abundant in the control group (Figure 5B). The PERMANOVA analysis was also used to compare the cancer group (CSC and DCL groups combined) with the control group (with no cancer diagnosis). The results showed that the phyla *Firmicutes*, *Fusobacteriota*, *Proteobacteria*, and *Spirochaetes* were more abundant in samples from patients diagnosed with cancer, whereas *Actinobacteria*, *Bacteroidetes*, *Epsilo bacteraeota*, and *Patescibacte* were more abundant in samples from the control group (Figure not shown).

We also analyzed the relationship between periodontitis and cancer using an NMDS plot, and no clear clustering was observed between four groups (Supplementary Figure S4), which were categorized based on having or not having periodontitis and cancer.

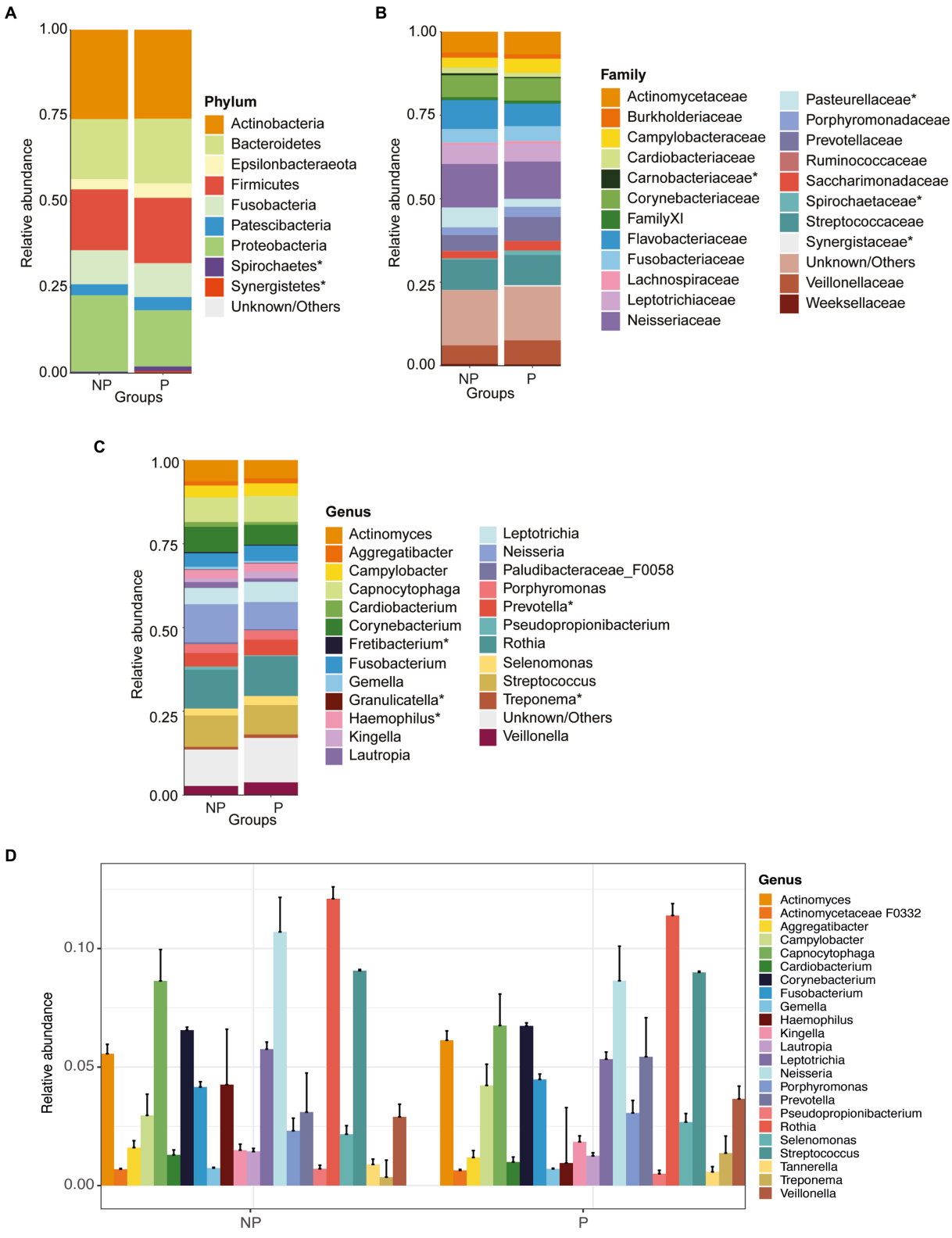


FIGURE 2
Relative abundance plots of microbiome compositions at different taxonomic levels and comparisons of alpha and beta diversity between periodontitis (P) and non-periodontitis (NP) individuals. Distribution of abundant organisms at the **(A)** phylum, **(B)** family, and **(C)** genus levels. **(D)** Extended error barplot showing the most abundant microbiome compositions between periodontitis and non-periodontitis individuals at the genus level. Organisms with a mean relative abundance of at least 1% across all samples are represented in different colors, whereas those with <1% abundance and unclassified are represented as "Unknown/Others". The microbial communities denoted with an asterisk (*) are the significant ones between groups.

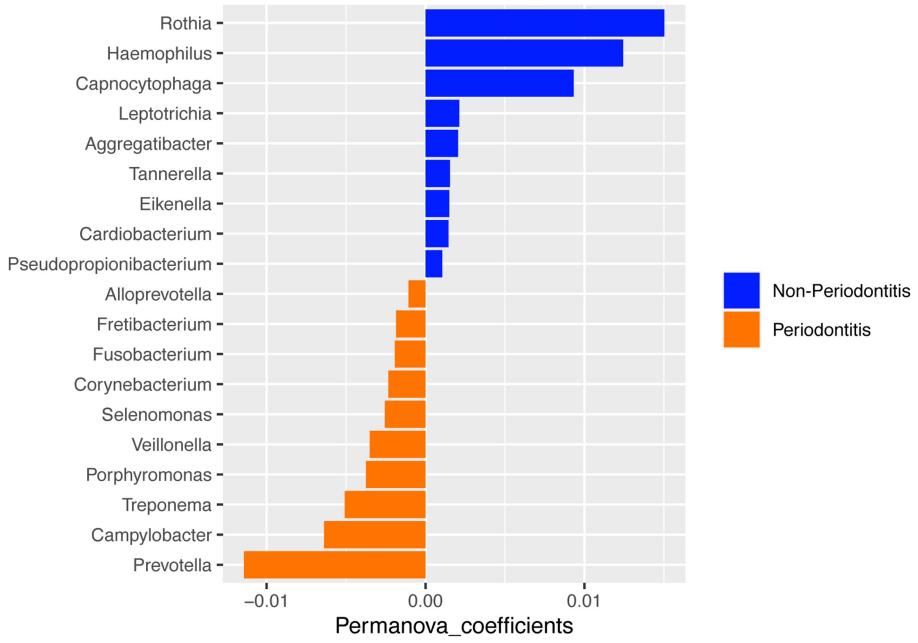


FIGURE 3
PERMANOVA analysis of microbial composition between the periodontitis and non-periodontitis groups at the genus level.

The four groups were individuals with non-periodontitis with cancer, NPC ($n = 15$); individuals with both periodontitis and cancer, PC ($n = 19$); individuals with periodontitis but no cancer, PNC ($n = 31$); and individuals with non-periodontitis and without cancer, NPNC ($n = 25$). When comparing these groups, *Pseudopropionibacterium* was differentiated between PC and PNC; as was *Granulicatella*, *Lautropia*, *Haemophilus*, and *Desulfobulbus* between NPC and PNC; and *Eubacterium saphenum*, *Filifactor*, and *Desulfobulbus* between NPC and PC.

Correlations between microbiota and periodontal clinical variables

A correlation analysis (for $p < 0.05$) was also performed to investigate the relationship between microbiota and periodontal disease. A matrix of the correlations between the periodontal clinical parameters AL, BOP, GI, PD, and PLI and microbial taxa (species level) for the three different groups, CSC, DCL and controls, is illustrated in Figure 6. In the CSC group, strong positive correlations were observed between the periodontal parameters AL, BOP, and GI and the species *Prevotella pleuritidis* and *Treponema parvum* (coefficients ranging from 0.6 to 0.75). In addition, PLI strongly correlated with the bacteria *Eubacterium nodatum*, *Eubacterium saphenum*, *Mycoplasma salivarium*, *Porphyromonas asaccharolytica*, *Prevotella dentalis*, *Prevotella pleuritidis*, and *Treponema parvum*. Negative correlations were found between AL, BOP, GI, PD, and PLI and *Actinomyces massiliensis;* as well as between *Capnocytophaga* sp. *oral taxon* and BOP and GI values (Figure 6A). In the DCL group, the strongest correlations (coefficients ranging from 0.50 to 0.59) were shown between all the periodontal parameters and the bacterium *Mitsuokella sp. oral taxon*. In this group, *Prevotella scopos JCM* was negatively correlated with AL and PD scores (Figure 6B). In contrast to the two cancer groups, CSC and DCL, no strong correlations (ranging between 0.27 to 0.45) were observed in the control group between the periodontal variables and the significantly abundant species (Figure 6C).

Discussion

Periodontal infection causes chronic inflammation in the oral cavity and is considered an important statistical risk factor for several types of cancer (Hajishengallis, 2014; Hajishengallis, 2015; Flemer et al., 2018). Some have proposed that the association between periodontitis and the risk of different types of cancer is due to the chronic inflammation caused by periodontitis, which drives cancer development by infiltration of leukocytes in the tumor microenvironment (Hanke et al., 1990; Garrett, 2015). Numerous studies have indeed reported a relationship between periodontal disease and various types of cancer (Aas et al., 2005; Keijser et al., 2008; Bik et al., 2010; Segata et al., 2011; Soder et al., 2011, 2021; Norder Grusell et al., 2013; Dong et al., 2018; Bai et al., 2022). In addition, studies have also shown that the oral microbiota may contribute to carcinogenesis by altering the homeostasis/cellular metabolism, the immune responses creating a proinflammatory microenvironment, cell migration and production of carcinogenic metabolites (Garrett, 2015; Bai et al., 2022; Lamont et al., 2022). In the current study, we aimed to investigate the subgingival microbial composition in periodontal health and disease and its relationship with the diagnosis and development of cancer in a longitudinal 10-year follow-up study. Our 16S rRNA results identified several significant genera differentiating individuals with periodontitis from those without periodontitis (*Haemophilus*, *Treponema*,

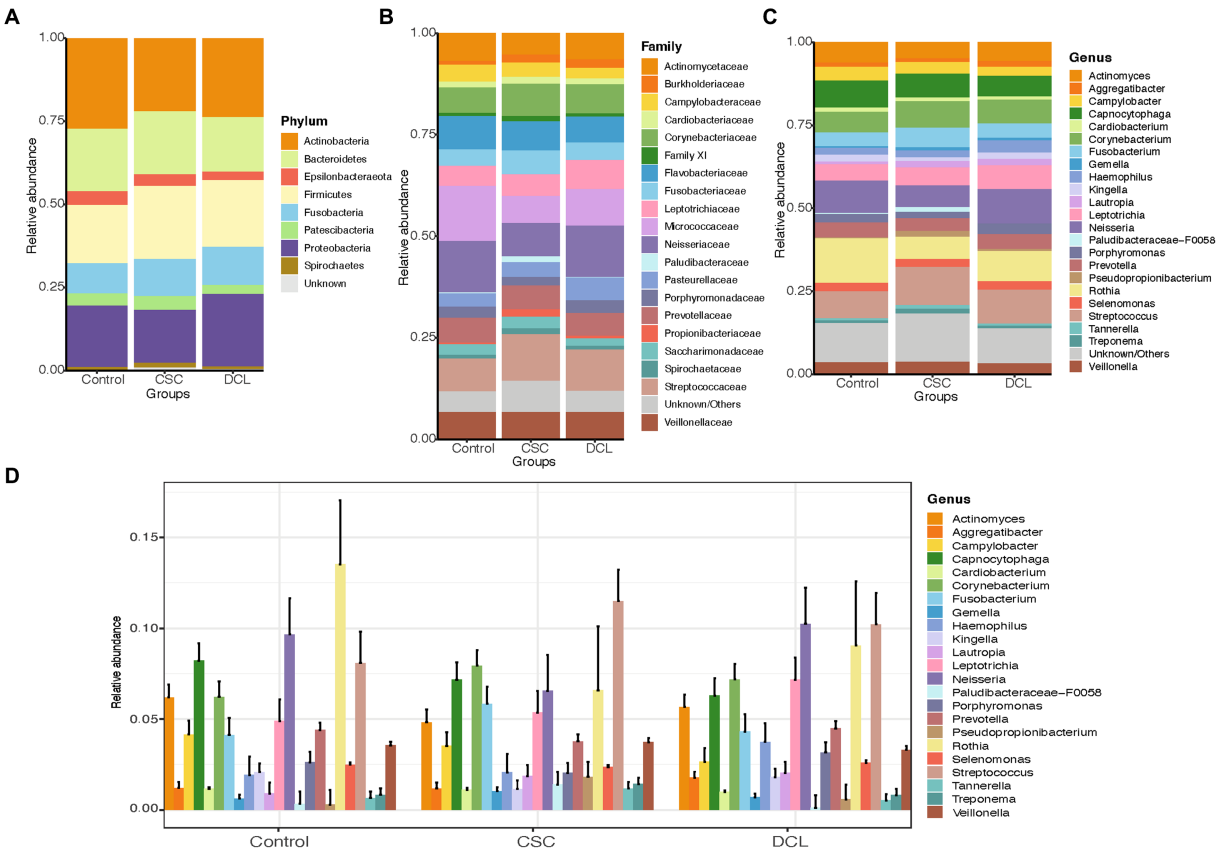


FIGURE 4
Distribution of the most abundant microbiome compositions between control, cancer at sample collection (CSC), and developed cancer later (DCL) groups at the **(A)** phylum, **(B)** family, and **(C)** genus levels. **(D)** Extended error barplot showing the distribution of the most abundant microbiome compositions between control, CSC and DCL groups at the genus level. Organisms with a mean relative abundance of at least 1% across all samples are represented in different colors, whereas those with <1% abundance and unclassified are represented as "Unknown/Others".

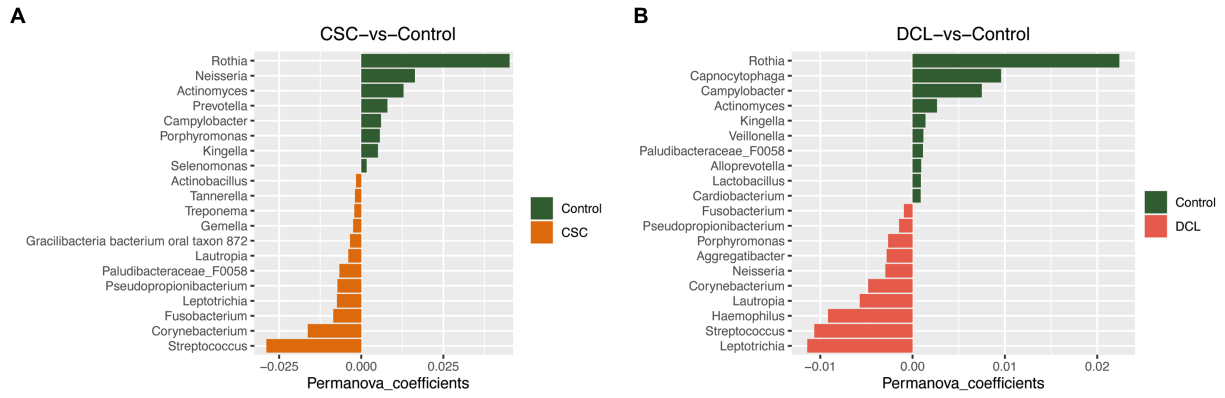


FIGURE 5
PERMANOVA analysis of microbial composition between control and cancer groups at the genus level: **(A)** cancer at sample collection (CSC) vs. control; **(B)** developed cancer later (DCL) vs. control.

Fretibacterium, *Granulicatella*, *Prevotella and Defluviitaleaceae UCG-011*). However, only one genus (*Pseudopropionibacterium*) appeared to be significantly different between the cancer patient samples and controls with no cancer.

Our bacterial abundance study was carried out at different taxonomic levels, including phylum, family, and genus. First, we analyzed the subgingival microbial composition in periodontally healthy individuals and patients with periodontitis. In agreement with

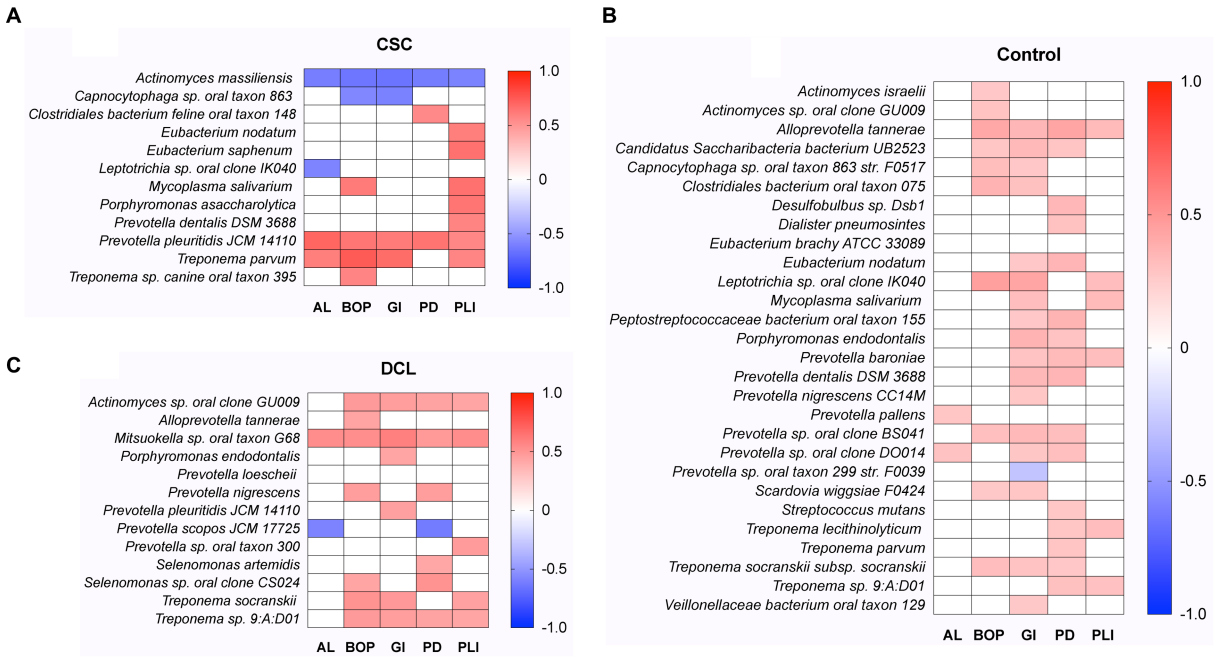


FIGURE 6
Spearman rank correlation analysis between subgingival microbiota and periodontal clinical parameters (AL, BOP, GI, PD, PLI). The heatmap shows statistically significant (p <0.05) correlations between microbial taxa (at the species level) and periodontal parameters for the **(A)** cancer at sample collection (CSC), **(B)** developed cancer later (DCL), and **(C)** control groups. Positive correlations are displayed in red and negative correlations in blue color.

previous findings, the most abundant bacteria at the phylum level were *Actinobacteria*, *Proteobacteria*, *Firmicutes,* and *Bacteroidetes* (Griffen et al., 2012; Cai et al., 2021; Sedghi et al., 2021). *Firmicutes* and *Bacteroidetes* were more abundant in the periodontitis group compared to the non-periodontitis group, also confirming previous findings reporting an increased abundance of *Bacteroidetes* and *Firmicutes* in periodontitis (Segata et al., 2012; Sedghi et al., 2021). The most abundant microbial communities at the genus level were *Actinomyces*, *Corynebacterium*, *Neisseria*, *Prevotella*, *Streptococcus,* and *Rothia*, whereas *Haemophilus*, *Treponema*, *Fretibacterium*, *Granulicatella*, *Prevotella*, and *Defluviitaleaceae UCG-011* genera were significantly different when comparing individuals with and without periodontitis. Notably, samples from periodontitis patients had significantly higher levels of *Treponema*, *Fretibacterium*, and *Prevotella* compared to non-periodontitis individuals. Thus, our results confirm earlier research as *Treponema*, *Fretibacterium*, and *Prevotella* (belonging to the phyla *Spirochetes*, *Synergistetes*, and *Bacteroidetes*, respectively) are reported to play essential roles in the pathogenesis of periodontitis (Perez-Chaparro et al., 2014; Hajishengallis, 2015; Lundmark et al., 2019). Similarly, the genus *Defluviitaleaceae UCG-011* (belonging to phylum *Firmicutes*) is more abundant in supragingival plaque samples of periodontitis than healthy controls (Kawamoto et al., 2021). *Treponema*, on the other hand, is a diverse bacterial genus and a constituent of healthy oral flora; however, with a vital role in the etiology and pathogenesis of periodontal disease, its reduction prompts the dysbiosis of microbiota (Buyuktimkin et al., 2019; Velusamy et al., 2019; Li et al., 2021).

Second, we analyzed the distribution of subgingival microbial composition and its relationship with occurrence of cancer in the CSC and DCL groups. At the phylum level, similar microbial composition was observed between the cancer and control groups. Indeed, *Actinomyces*, *Corynebacterium*, *Fusobacterium*, *Neisseria*, *Prevotella*, *Rothia*, and *Streptococcus* were the differentially abundant genera found in the periodontitis as well as in the cancer groups. In agreement with our findings, microbes including *Fusobacterium*, *Streptococcus*, and *Prevotella* have been detected in high abundance in cancerous periodontal tissues (Dong et al., 2018). Consistent with these findings, genomic analysis, 16S rDNA sequence analysis, and quantitative PCR have revealed that *Fusobacterium* sequences are enriched in colorectal carcinoma (Kostic et al., 2012). When comparing the three groups (CSC, DCL and controls) at the genus level, *Neisseria*, *Rothia*, and *Capnocytophaga* were more abundant in the control group, whereas *Corynebacterium* and *Streptococcus* were more abundant in the CSC group, and *Prevotella* were more abundant in the DCL group. Indeed, the genus *Corynebacterium*, has been shown to be enriched in saliva samples of gastric cancer patients (Oliveira et al., 2017). Similarly, strains of *Streptococcus* have been reported to be involved in numerous types of cancer including colorectal adenocarcinomas and gastric cancer (Abdulamir et al., 2011). On the other hand, a higher abundance of *Corynebacterium* has also been associated with reduced risk of head and neck squamous cell cancer, as well as good oral health (Meuric et al., 2017; Hayes et al., 2018).

Several subgingival genera were differentially enriched among the study groups. For example, *Pseudopropionibacterium* (phylum *Actinobacteria*) was found to be significantly enriched in the periodontitis group with cancer, whereas *Eubacterium saphenum*, *Filifactor*, and *Desulfobulbus* were found to be enriched in the non-periodontitis group with cancer, suggesting that these genera may be involved in cancer development. Notably, *Pseudopropionibacterium* has been found to be more prevalent in cases of apical periodontitis

and to have a significant difference in abundance in esophageal cancer cases compared to controls (Liu et al., 2020; Perez-Carrasco et al., 2023).

Our PERMANOVA results showed that *Rothia* were the most prevalent genus in the oral microbiota in both the non-cancer and non-periodontitis groups, which suggests that this genus may have a protective effect on periodontitis and cancer development. This hypothesis is supported by a previous study demonstrating that the genus *Rothia* were more prevalent in healthy controls compared to subjects with oral squamous cell carcinoma (Zhao et al., 2017).

Finally, we investigated the inter-relationship and correlations between the microbial communities and periodontitis/periodontal parameters and cancer (in the CSC and DCL samples). The differential abundance of microbial composition was confirmed using a PERMANOVA analysis, which compared the bacterial compositions between two groups and characterized the top discriminative taxa between them. The alpha diversity results showed that non-periodontitis individuals had lower diversity compared to periodontitis patients. However, there were no significant differences in the diversity between periodontitis and non-periodontitis groups (Supplementary Figure S2). Similarly, there were no significant differences in the diversity between the CSC, DCL, and control groups (Supplementary Figure S3). Notably, the correlation analysis, which was performed to evaluate the relationship between microbiota and different clinical periodontal parameters, revealed strong correlations between BOP, GI, PLI, and several species belonging to genera *Prevotella*, *Treponema*, and *Mycoplasma* in the CSC group. Both BOP and GI are well-known indicators of gingival inflammation, which may thus contribute to the development of cancer.

The strength of our study was the homogeneity of the subject material, as the cohort ($n = 1,676$) was followed-up together for over 30 years, with cumulated disease data from the national population registers of Sweden. The current study is a clinically examined sample of the large cohort that was followed-up for 10 years. However, the study was limited by the relatively small 10-year sample that was available for the present investigation, as well as the relatively short follow-up period considering the development of new cancer cases. Indeed, the development of cancer is a slow process. With regard to the cross-sectional design of the study, providing a snapshot of the microbiota collected at one point in time, longitudinal studies that follow participants over time are needed to gain deeper insights into the relationship between oral microbiota and various types of cancer. Another limitation is that the study did not consider confounding factors such as smoking, medication use or diet into consideration, which may have influenced the composition of the subgingival microbiota. Furthetmore, in the current study, we did not differentiate between healthy and diseased sites when pooling subgingival samples from each participant, which may have contributed to the lack of clear beta diversity clusters between the periodontitis and non-periodontitis groups. This is consistent with previous findings indicating that the subgingival microbiota can differ between healthy and diseased sites in patients with periodontitis, suggesting a site-specific presence of periodontal pathogens in plaque samples (Belstrøm et al., 2017).

In conclusion, in the present study, we identified several genera differentiating periodontitis from non-periodontitis groups of subjects (*Haemophilus*, *Treponema*, *Fretibacterium*, *Granulicatella, Prevotella* and *Defluviitaleaceae UCG-011*). Only one genus, *Pseudopropionibacterium*, differentiated individuals with periodontitis having cancer (PC) from periodontitis patients without any cancer (PNC). Additionally, in the CSC group, we also found that the presence of periodontal inflammation (as reflected in BOP, GI, and PLI scores) strongly correlated with species belonging to the genera *Prevotella*, *Treponema*, and *Mycoplasma*. Collectively, our findings revealed significant differences in the subgingival microbiota among the studied groups, underscoring the need for further investigation into the potential role of oral pathogens in the development of cancer.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found at: NCBI BioProject, accession number: PRJNA985445 [https://www.ncbi.nlm.nih.gov/bioproject/985445].

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee of the Karolinska University Hospital at Huddinge (Dnr 2007/1669-31; 2012/590-32; 2017/2204-32). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

BS, TY-L, JM, and UN contributed to the conception and design of the study. BS contributed to the clinical dental examinations and sample collection. AL and YH performed all the experiments. AN and UN performed the bioinformatical and statistical analyses. TY-L, AN, and UN wrote the first draft of the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

The study was supported by grants from the Swedish Research Council (Grant No. 2017-02084), the Patent Revenue Fund for Research in Preventive Odontology; King Gustaf V's and Queen Victoria's Freemasons Foundation, Stockholm; the Finnish Society of Sciences and Letters; and the Finnish Medical Society, Helsinki, Finland. UN acknowledges support from the Swedish Research Council (Grant Nos 2018-06156 and 2021-01756). Additionally, the authors would like to acknowledge support from the National Board of Health and Welfare and Science for Life Laboratory, the National Genomics Infrastructure, NGI, and

Uppmax for providing assistance in sequencing and computational infrastructure. The data-handling was enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at [SNIC CENTRE] partially funded by the Swedish Research Council through grant agreement (Grant No. 2018-05973). This research work is dedicated to Professor Per-Östen Söder in recognition of his initial contribution to the project.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1172340/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Rarefaction curves generated according to the different alpha diversity indices between periodontitis and non-periodontitis individuals, aiming to determine how many microbial communities can be detected with increasing numbers of sequencing reads. The curves do not converge, indicating that increasing the number of sequencing depths may result in the identification of additional species.

SUPPLEMENTARY FIGURE S2
Boxplots comparing alpha diversity indices (Observed, Chao1, ACE, Shannon, Simpson, and Inverse simpson) between the periodontitis and non-periodontitis groups.

SUPPLEMENTARY FIGURE S3
Boxplots comparing alpha diversity indices (Observed, Chao1, ACE, Shannon, Simpson, and Inverse simpson) between the control, cancer at sample collection (CSC), and developed cancer later (DCL) groups.

SUPPLEMENTARY FIGURE S4
Relative abundance and beta diversity between non-periodontitis and cancer (NPC), periodontitis and cancer (PC), periodontitis and no cancer (PNC), and non-periodontitis and no cancer (NPNC) groups. The significant bacteria ($P<0.05$) at the genus level, comparing the different groups with the periodontitis vs. cancer groups, is presented.

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

Abdulamir, A. S., Hafidh, R. R., and Abu Bakar, F. (2011). The association of *Streptococcus bovis*/gallolyticus with colorectal tumors: the nature and the underlying mechanisms of its etiological role. *J. Exp. Clin. Cancer Res.* 30:11. doi: 10.1186/1756-9966-30-11

Ahn, J., Chen, C. Y., and Hayes, R. B. (2012). Oral microbiome and oral and gastrointestinal cancer risk. *Cancer Causes Control* 23, 399–404. doi: 10.1007/s10552-011-9892-7

Anderson, M. J. (2017). Permutational multivariate analysis of variance (PERMANOVA). *Wiley StatsRef: Statistics Reference Online*, 1–15. doi: 10.1002/9781118445112.stat07841

Andrews, S. (2015). FASTQC a quality control tool for high throughput sequence data. Babraham Institute. Available at: https://www.bibsonomy.org/bibtex/f230a919c34360709aa298734d63dca3

Asikainen, S., and Chen, C. (1999, 2000). Oral ecology and person-to-person transmission of Actinobacillus actinomycetemcomitans and *Porphyromonas gingivalis*. *Periodontology* , 65–81. doi: 10.1111/j.1600-0757.1999.tb00158.x

Båge, T., Kats, A., Lopez, B. S., Morgan, G., Nilsson, G., Burt, I., et al. (2011). Expression of prostaglandin E synthases in periodontitis immunolocalization and cellular regulation. *Am. J. Pathol.* 178, 1676–1688. doi: 10.1016/j.ajpath.2010.12.048

Bai, H., Yang, J., Meng, S., and Liu, C. (2022). Oral microbiota-driven cell migration in carcinogenesis and metastasis. *Front. Cell. Infect. Microbiol.* 12:864479. doi: 10.3389/fcimb.2022.864479

Bascones, A., Noronha, S., Gomez, M., Mota, P., Gonzalez Moles, M. A., and Villarroel Dorrego, M. (2005). Tissue destruction in periodontitis: bacteria or cytokines fault? *Quintessence Int.* 36, 299–306.

Belstrøm, D., Sembler-Møller, M. L., Grande, M. A., Kirkby, N., Cotton, S. L., Paster, B. J., et al. (2017). Microbial profile comparisons of saliva, pooled and site-specific subgingival samples in periodontitis patients. *PLoS One* 12:e0182992. doi: 10.1371/journal.pone.0182992

Bik, E. M., Long, C. D., Armitage, G. C., Loomer, P., Emerson, J., Mongodin, E. F., et al. (2010). Bacterial diversity in the oral cavity of 10 healthy individuals. *ISME J.* 4, 962–974. doi: 10.1038/ismej.2010.30

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

Buyuktimkin, B., Zafar, H., and Saier, M. H. Jr. (2019). Comparative genomics of the transportome of ten Treponema species. *Microb. Pathog.* 132, 87–99. doi: 10.1016/j.micpath.2019.04.034

Cai, Z., Lin, S., Hu, S., and Zhao, L. (2021). Structure and function of Oral microbial Community in Periodontitis Based on integrated data. *Front. Cell. Infect. Microbiol.* 11:663756. doi: 10.3389/fcimb.2021.663756

Cavalla, F., Osorio, C., Paredes, R., Valenzuela, M. A., García-Sesnich, J., Sorsa, T., et al. (2015). Matrix metalloproteinases regulate extracellular levels of SDF-1/CXCL12, IL-6 and VEGF in hydrogen peroxide-stimulated human periodontal ligament fibroblasts. *Cytokine* 73, 114–121. doi: 10.1016/j.cyto.2015.02.001

Coussens, L. M., and Werb, Z. (2002). Inflammation and cancer. *Nature* 420, 860–867. doi: 10.1038/nature01322

Davanian, H., Båge, T., Lindberg, J., Lundeberg, J., Concha, H. Q., Sällberg Chen, M., et al. (2012). Signaling pathways involved in the regulation of TNFalpha-induced toll-like receptor 2 expression in human gingival fibroblasts. *Cytokine* 57, 406–416. doi: 10.1016/j.cyto.2011.12.008

de Martel, C., Ferlay, J., Franceschi, S., Vignat, J., Bray, F., Forman, D., et al. (2012). Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol.* 13, 607–615. doi: 10.1016/S1470-2045(12)70137-7

Dong, L., Yin, J., Zhao, J., Ma, S. R., Wang, H. R., Wang, M., et al. (2018). Microbial similarity and preference for specific sites in healthy Oral cavity and esophagus. *Front. Microbiol.* 9:1603. doi: 10.3389/fmicb.2018.01603

Dye, B. A. (2012). Global periodontal disease epidemiology. *Periodontology* 2000, 10–25. doi: 10.1111/j.1600-0757.2011.00413.x

Eke, P. I., Dye, B. A., Wei, L., Slade, G. D., Thornton-Evans, G. O., Borgnakke, W. S., et al. (2015). Update on prevalence of periodontitis in adults in the United States: NHANES 2009 to 2012. *J. Periodontol.* 86, 611–622. doi: 10.1902/jop.2015.140520

Flemer, B., Warren, R. D., Barrett, M. P., Cisek, K., das, A., Jeffery, I. B., et al. (2018). The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* 67, 1454–1463. doi: 10.1136/gutjnl-2017-314814

Gallimidi, A. B., Fischman, S., Revach, B., Bulvik, R., Maliutina, A., Rubinstein, A. M., et al. (2015). Periodontal pathogens Porphyromonas gingivalis and *Fusobacterium nucleatum* promote tumor progression in an oral-specific chemical carcinogenesis model. *Oncotarget* 6, 22613–22623. doi: 10.18632/oncotarget.4209

Garrett, W. S. (2015). Cancer and the microbiota. *Science* 348, 80–86. doi: 10.1126/science.aaa4972

Griffen, A. L., Beall, C. J., Campbell, J. H., Firestone, N. D., Kumar, P. S., Yang, Z. K., et al. (2012). Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *ISME J.* 6, 1176–1185. doi: 10.1038/ismej.2011.191

Groeger, S., Domann, E., Gonzales, J. R., Chakraborty, T., and Meyle, J. (2011). B7-H1 and B7-DC receptors of oral squamous carcinoma cells are upregulated by *Porphyromonas gingivalis*. *Immunobiology* 216, 1302–1310. doi: 10.1016/j.imbio.2011.05.005

Hajishengallis, G. (2014). Immunomicrobial pathogenesis of periodontitis: keystones, pathobionts, and host response. *Trends Immunol.* 35, 3–11. doi: 10.1016/j.it.2013.09.001

Hajishengallis, G. (2015). Periodontitis: from microbial immune subversion to systemic inflammation. *Nat. Rev. Immunol.* 15, 30–44. doi: 10.1038/nri3785

Hajishengallis, G., and Chavakis, T. (2021). Local and systemic mechanisms linking periodontal disease and inflammatory comorbidities. *Nat. Rev. Immunol.* 21, 426–440. doi: 10.1038/s41577-020-00488-6

Hanke, C. W., Arndt, K. A., Dobson, R. L., Dzubow, L. M., Parish, L. C., and Taylor, J. S. (1990). Dual publication and manipulation of the editorial process. *Arch. Dermatol.* 126, 1625–1626. doi: 10.1111/j.1525-1470.1990.tb01035.x

Hayes, R. B., Ahn, J., Fan, X., Peters, B. A., Ma, Y., Yang, L., et al. (2018). Association of Oral Microbiome with Risk for incident head and neck squamous cell Cancer. *JAMA Oncol.* 4, 358–365. doi: 10.1001/jamaoncol.2017.4777

Herlemann, D. P., Labrenz, M., Jurgens, K., Bertilsson, S., Waniek, J. J., and Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J.* 5, 1571–1579. doi: 10.1038/ismej.2011.41

Irfan, M., Delgado, R. Z. R., and Frias-Lopez, J. (2020). The Oral microbiome and Cancer. *Front. Immunol.* 11:591088. doi: 10.3389/fimmu.2020.591088

Kawamoto, D., Borges, R., Ribeiro, R. A., de Souza, R. F., Amado, P. P. P., Saraiva, L., et al. (2021). Oral Dysbiosis in severe forms of periodontitis is associated with gut Dysbiosis and correlated with salivary inflammatory mediators: a preliminary study. *Front Oral Health* 2:722495. doi: 10.3389/froh.2021.722495

Kechin, A., Boyarskikh, U., Kel, A., and Filipenko, M. (2017). cutPrimers: a new tool for accurate cutting of primers from reads of targeted next generation sequencing. *J. Comput. Biol.* 24, 1138–1143. doi: 10.1089/cmb.2017.0096

Keijser, B. J., Zaura, E., Huse, S. M., van der Vossen, J. M., Schuren, F. H., Montijn, R. C., et al. (2008). Pyrosequencing analysis of the oral microflora of healthy adults. *J. Dent. Res.* 87, 1016–1020. doi: 10.1177/154405910808701104

Kim, E. H., Nam, S., Park, C. H., Kim, Y., Lee, M., Ahn, J. B., et al. (2022). Periodontal disease and cancer risk: a nationwide population-based cohort study. *Front. Oncol.* 12:901098. doi: 10.3389/fonc.2022.901098

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., et al. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res.* 22, 292–298. doi: 10.1101/gr.126573.111

Lamont, R. J., Fitzsimonds, Z. R., Wang, H., and Gao, S. (2022). Role of *Porphyromonas gingivalis* in oral and orodigestive squamous cell carcinoma. *Periodontol.* 89, 154–165. doi: 10.1111/prd.12425

Li, X., Liu, Y., Yang, X., Li, C., and Song, Z. (2022). The Oral microbiota: community composition, influencing factors, pathogenesis, and interventions. *Front. Microbiol.* 13:895537. doi: 10.3389/fmicb.2022.895537

Li, W., Xu, J., Zhang, R., Li, Y., Wang, J., Zhang, X., et al. (2021). Is periodontal disease a risk indicator for colorectal cancer? A systematic review and meta-analysis. *J. Clin. Periodontol.* 48, 336–347. doi: 10.1111/jcpe.13402

Liu, F., Liu, M., Liu, Y., Guo, C., Zhou, Y., Li, F., et al. (2020). Oral microbiome and risk of malignant esophageal lesions in a high-risk area of China: a nested case-control study. *Chin. J. Cancer Res.* 32, 742–754. doi: 10.21147/j.issn.1000-9604.2020.06.07

Lundin, S., Stranneheim, H., Pettersson, E., Klevebring, D., and Lundeberg, J. (2010). Increased throughput by parallelization of library preparation for massive sequencing. *PLoS One* 5:e10029. doi: 10.1371/journal.pone.0010029

Lundmark, A., Hu, Y. O. O., Huss, M., Johannsen, G., Andersson, A. F., and Yucel-Lindberg, T. (2019). Identification of salivary microbiota and its association with host inflammatory mediators in periodontitis. *Front. Cell. Infect. Microbiol.* 9:216. doi: 10.3389/fcimb.2019.00216

Meuric, V., Le Gall-David, S., Boyer, E., Acuna-Amador, L., Martin, B., Fong, S. B., et al. (2017). Signature of microbial Dysbiosis in periodontitis. *Appl. Environ. Microbiol.* 83:e00462-17. doi: 10.1128/AEM.00462-17

Michaud, D. S. (2013). Role of bacterial infections in pancreatic cancer. *Carcinogenesis* 34, 2193–2197. doi: 10.1093/carcin/bgt249

Michaud, D. S., Fu, Z., Shi, J., and Chung, M. (2017). Periodontal disease, tooth loss, and Cancer risk. *Epidemiol. Rev.* 39, 49–58. doi: 10.1093/epirev/mxx006

Norder Grusell, E., Dahlen, G., Ruth, M., Ny, L., Quiding-Jarbrink, M., Bergquist, H., et al. (2013). Bacterial flora of the human oral cavity, and the upper and lower esophagus. *Dis. Esophagus* 26, 84–90. doi: 10.1111/j.1442-2050.2012.01328.x

Nwizu, N. N., Marshall, J. R., Moysich, K., Genco, R. J., Hovey, K. M., Mai, X., et al. (2017). Periodontal disease and incident Cancer risk among postmenopausal women: results from the Women's Health Initiative observational cohort. *Cancer Epidemiol. Biomark. Prev.* 26, 1255–1265. doi: 10.1158/1055-9965.EPI-17-0212

Oliveira, A., Oliveira, L. C., Aburjaile, F., Benevides, L., Tiwari, S., Jamal, S. B., et al. (2017). Insight of genus Corynebacterium: ascertaining the role of pathogenic and non-pathogenic species. *Front. Microbiol.* 8:1937. doi: 10.3389/fmicb.2017.01937

O'Toole, G., Kaplan, H. B., and Kolter, R. (2000). Biofilm formation as microbial development. *Annu. Rev. Microbiol.* 54, 49–79. doi: 10.1146/annurev.micro.54.1.49

Page, R. C., and Eke, P. I. (2007). Case definitions for use in population-based surveillance of periodontitis. *J. Periodontol.* 78, 1387–1399. doi: 10.1902/jop.2007.060264

Page, R. C., and Kornman, K. S. (1997). The pathogenesis of human periodontitis: an introduction. *Periodontol.* 14, 9–11. doi: 10.1111/j.1600-0757.1997.tb00189.x

Perez-Carrasco, V., Uroz-Torres, D., Soriano, M., Solana, C., Ruiz-Linares, M., Garcia-Salcedo, J. A., et al. (2023). Microbiome in paired root apices and periapical lesions and its association with clinical signs in persistent apical periodontitis using next-generation sequencing. *Int. Endod. J.* 56, 622–636. doi: 10.1111/iej.13893

Perez-Chaparro, P. J., Goncalves, C., Figueiredo, L. C., Faveri, M., Lobao, E., Tamashiro, N., et al. (2014). Newly identified pathogens associated with periodontitis: a systematic review. *J. Dent. Res.* 93, 846–858. doi: 10.1177/0022034514542468

Pollanen, M. T., Laine, M. A., Ihalin, R., and Uitto, V. J. (2012). Host-bacteria crosstalk at the dentogingival junction. *Int J Dent* 2012:821383. doi: 10.1155/2012/821383

Romandini, M., Baima, G., Antonoglou, G., Bueno, J., Figuero, E., and Sanz, M. (2021). Periodontitis, Edentulism, and risk of mortality: a systematic review with Meta-analyses. *J. Dent. Res.* 100, 37–49. doi: 10.1177/0022034520952401

Sedghi, L. M., Bacino, M., and Kapila, Y. L. (2021). Periodontal disease: the good, the bad, and the unknown. *Front. Cell. Infect. Microbiol.* 11:766944. doi: 10.3389/fcimb.2021.766944

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

Serhan, C. N. (2014). Pro-resolving lipid mediators are leads for resolution physiology. *Nature* 510, 92–101. doi: 10.1038/nature13479

Soder, B., Jin, L. J., Klinge, B., and Soder, P. O. (2007). Periodontitis and premature death: a 16-year longitudinal study in a Swedish urban population. *J. Periodontal Res.* 42, 361–366. doi: 10.1111/j.1600-0765.2006.00957.x

Soder, B., Kallmen, H., Yucel-Lindberg, T., and Meurman, J. H. (2021). Periodontal microorganisms and diagnosis of malignancy: a cross-sectional study. *Tumour Biol.* 43, 1–9. doi: 10.3233/TUB-200066

Soder, B., Yakob, M., Meurman, J. H., Andersson, L. C., Klinge, B., and Soder, P. O. (2011). Periodontal disease may associate with breast cancer. *Breast Cancer Res. Treat.* 127, 497–502. doi: 10.1007/s10549-010-1221-4

Van Dyke, T. E. (2009). Resolution of inflammation-unraveling mechanistic links between periodontitis and cardiovascular disease. *J. Dent.* 37, S582–S583. doi: 10.1016/j.jdent.2009.05.013

Velusamy, S. K., Sampathkumar, V., Ramasubbu, N., Paster, B. J., and Fine, D. H. (2019). *Aggregatibacter actinomycetemcomitans* colonization and persistence in a primate model. *Proc. Natl. Acad. Sci. U. S. A.* 116, 22307–22313. doi: 10.1073/pnas.1905238116

Virtanen, E., Soder, B., Andersson, L. C., Meurman, J. H., and Soder, P. O. (2014). History of dental infections associates with cancer in periodontally healthy subjects: a 24-year follow-up study from Sweden. *J. Cancer* 5, 79–85. doi: 10.7150/jca.7402

Yakob, M., Kari, K., Tervahartiala, T., Sorsa, T., Soder, P. O., Meurman, J. H., et al. (2012). Associations of periodontal microorganisms with salivary proteins and MMP-8 in gingival crevicular fluid. *J. Clin. Periodontol.* 39, 256–263. doi: 10.1111/j.1600-051X.2011.01813.x

Yucel-Lindberg, T., and Bage, T. (2013). Inflammatory mediators in the pathogenesis of periodontitis. *Expert Rev. Mol. Med.* 15:e7. doi: 10.1017/erm.2013.8

Zhao, H., Chu, M., Huang, Z., Yang, X., Ran, S., Hu, B., et al. (2017). Variations in oral microbiota associated with oral cancer. *Sci. Rep.* 7:11773. doi: 10.1038/s41598-017-11779-9

Frontiers in Microbiology

Explores the habitable world and the potential of microbial life

The largest and most cited microbiology journal which advances our understanding of the role microbes play in addressing global challenges such as healthcare, food security, and climate change.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact