Heliyon 10 (2024) e30896

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Integrative analyses identify opportunistic pathogens of patients with lower respiratory tract infections based on metagenomic next-generation sequencing

Tingyan Dong [a,b,1], Yueming Liang [c,1], Junting Xie [c], Wentao Fan [b], Haitao Chen [b], Xiaodong Han [a,d,*]

[a] Jiangsu Key Laboratory of Molecular Medicine, Nanjing University, Nanjing, China
[b] Integrated Diagnostic Centre for Infectious Diseases, Guangzhou Huayin Medical Laboratory Center, Guangzhou, China
[c] Department of Respiratory and Critical Care Medicine, The First People Hospital of Foshan, Foshan, China
[d] Immunology and Reproduction Biology Laboratory & State Key Laboratory of Analytical Chemistry for Life Science, Medical School, Nanjing University, Nanjing, China

ARTICLE INFO

Keywords:
Lower respiratory tract infections
Metagenomic next-generation sequencing
Opportunistic pathogens
Identification

ABSTRACT

Lower respiratory tract infections (LRTIs) represent some of the most globally prevalent and detrimental diseases. Metagenomic next-generation sequencing (mNGS) technology has effectively addressed the requirement for the diagnosis of clinical infectious diseases. This study aimed at identifying and classifying opportunistic pathogens from the respiratory tract-colonizing microflora in LRTI patients using data acquired from mNGS analyses. A retrospective study was performed employing the mNGS data pertaining to the respiratory samples derived from 394 LRTIs patients. Linear discriminant analysis effect size (LEfSe) analysis was conducted to discern the discriminant bacteria. Receiver operating characteristic curves (ROC) were established to demonstrate discriminant bacterial behavior to distinguish colonization from infection. A total of 443 discriminant bacteria were identified and segregated into three cohorts contingent upon their correlation profiles, detection frequency, and relative abundance in order to distinguish pathogens from colonizing microflora. Among them, 119 emerging opportunistic pathogens (cohort 2) occupied an average area under the curve (AUC) of 0.976 for exhibiting the most prominent predictability in distinguishing colonization from infection, 39 were colonizing bacteria (cohort 1, 0.961), and 285 were rare opportunistic pathogens (cohort 3, 0.887). The LTRIs patients appeared modular in the form of cohorts depicting complex microbial co-occurrence networks, reduced diversity, and a high degree of antagonistic interactions in the respiratory tract microbiome. The study findings indicate that therapeutic interventions should target interaction networks rather than individual microbes, providing an innovative perspective for comprehending and combating respiratory infections. Conclusively, this study reports a profile of LRTIs-associated bacterial colonization and opportunistic pathogens in a relatively large-scale cohort, which might serve as a reference panel for the interpretation of mNGS results in clinical practice.

* Corresponding author. Jiangsu Key Laboratory of Molecular Medicine, Nanjing University, Nanjing, China.
E-mail address: hanxd@nju.edu.cn (X. Han).
[1] These authors have contributed equally to this study and share first authorship.

https://doi.org/10.1016/j.heliyon.2024.e30896
Received 17 January 2024; Received in revised form 6 May 2024; Accepted 7 May 2024
Available online 8 May 2024

1. Introduction

Lower respiratory tract infections (LRTIs) are one of the predominant causes of morbidity and mortality among all patient populations worldwide [1]. The respiratory system harbors a myriad of bacteria; consequently, bacterial dysbiosis can trigger an escalation in colonization by opportunistic pathogens, which deems the determination of causative pathogens challenging [2]. Despite the prominence of molecular biological techniques in pathogen detection in recent years, accurate determination of the etiology of respiratory infections remains arduous owing to constraints in the sensitivity, speed, and spectrum of available assay targets [3,4]. Additionally, clinical specimens collected from LRTIs patients are easily contaminated by oral colonization flora, which complicates judgment pertaining to whether the detected microorganisms are indicative of infection, colonization, or contamination [5]. Therefore, our study endeavored to explore an approach that discerns potential respiratory pathogens from the colonization microbiome to assist in LRTIs diagnosis through the application of comprehensive statistical analysis to a sizeable sample population.

Metagenomic next-generation sequencing (mNGS) is a highly efficacious method characterized by a brief turnaround time; it can potentially overcome the limitations of current diagnostic tests and facilitate hypothesis-free, culture-independent pathogen detection directly from clinical specimens [4,6]. In recent years, an increasing number of clinical studies and case reports on rare and fastidious pathogen-borne infections have affirmed the auxiliary diagnostic potency of mNGS [7–10]. Nevertheless, the detection of pathogens within respiratory specimens, amidst the intricate presence of commensal microbiota, remains insufficiently investigated. The screening of clinical samples using mNGS enable the detection of normal microbiota, transient colonizers, sample contamination, and/or infection [11]. Moreover, the proficiency and knowledge of interpreters regarding mNGS technology and clinical microbiology may impose limitations on the distinction between pathogens and colonizing bacteria [12]. Consequently, it is necessary to provide a profile of LRTIs-associated bacterial colonization or pathogens that might serve as a reference panel for the clinical diagnosis of infectious diseases. Hence, the current study aimed at addressing the need to distinguish between colonization and infection for efficient LRTIs diagnosis through mNGS-based integrative analyses of the detection frequency and relative abundance of each organism detected in a large number of respiratory samples.

This study adhered to a three-step methodology to discern opportunistic pathogens from colonization or infection. First, a comprehensive analysis of the composition and ecology of the respiratory microbial ecosystem in patients with LRTIs was conducted contingent upon mNGS data. Discriminant bacteria between the patients and healthy individuals were identified and categorized according to their correlation profiles. Second, the opportunistic pathogens and colonizing microflora were distinguished in compliance with the distribution features of the discriminant bacteria based on two principal indicators, namely detection frequency and relative abundance. Receiver operating characteristic (ROC) curves evaluated the bacterial capacity to differentiate between colonization and infection. Finally, the co-abundance analysis was performed to explore the bacterial-bacterial interactions. Collectively, we are aiming to reveal LRTIs-associated pathogens and colonizing microorganisms to assist the clinical infectious diagnose and mNGS reports interpretation.

2. Material and methods

2.1. Patients and sample collection

A total of 394 patients with suspected acute or chronic LRTIs who were admitted to the Huayin Medical Laboratory Center (Guangzhou, China) between March 2020 and March 2023 were recruited for a retrospective review. The attending physician ascertained the presence or absence of respiratory tract infections by evaluating clinical manifestations and imaging examination findings.

The study included patients presenting with evident manifestations of LRTIs [13], including: (1) pneumonia, fever (>38 °C), suspension or shortness of breath, bradycardia, wheezing, coughing, dry snoring, chest scans exhibiting new or progressive exudation, solid shadows, and cavity or pleural effusion; (2) tracheitis or tracheobronchitis, with two or more symptoms from among cough with increased sputum, dry voice, wheezing, respiratory distress, apnea, or bradycardia, but without clinical symptoms or X-ray evidence of pneumonia; and (3) other LRTIs, such as lung abscess or empyema. In addition, patients who agreed to undergo the mNGS examination were included. The exclusion criteria were as follows: (1) refusal to undergo mNGS examination; (2) failure of the sample detection process to pass quality control for mNGS; and (3) incomplete clinical and laboratory data.

A total of 426 bronchoalveolar lavage fluid (BALF) and 29 sputum samples were extracted from the study participants and used for subsequent extensive and elaborate mNGS-based assays in compliance with a composite reference standard (final clinical diagnosis), including clinical signs and symptoms, microbiological evidence, imaging findings, and clinical adjudication. They were recruited from the hospitalization zone of Guangdong Hospital, China. In total, 455 samples from 394 patients with LRTIs and 187 samples from 171 healthy individuals were analyzed using mNGS in this study. Additionally, raw mNGS data of 179 healthy individuals (90 male, 89 female) downloaded from the National Center for Biotechnology Information (NCBI) database (Bio-project: PRJNA413615 and PRJNA655567), including eight BALF and 179 sputum samples, were analyzed for comparison. The healthy people were recruited according to the following criteria: no diagnosis of asthma or family history of allergy; no history of pneumonia; a lack of wheezing, fever, cough, or other respiratory/allergic symptoms at sampling 1 month prior to the study and 1 week after sampling; no exposure to antibiotics 1 month prior to sampling.

2.2. mNGS and microbial ecological analysis

Samples were collected by following the standards of aseptic processing procedures and high-throughput sequencing: (1) A 1.5–3 mL BALF sample or sputum was collected from each patient. Nucleic acid extraction was performed using a TIANamp Micro DNA Kit (DP316, Tiangen Biotech Co., Beijing, China). (2) DNA libraries were constructed using the VAHTS Universal Plus DNA Library Prep Kit for Illumina (ND617-C2; Vazyme Biotech Co.). Agilent 2100 was used for quality control of the DNA libraries. Qualified libraries were sequenced using Illumina NextSeq CN500. (3) High-quality data were obtained by removing low-quality reads, adapter contamination, duplicate reads, and reads shorter than 50 bp, using Trimmomatic. Low-complexity reads were removed using fastp (v.0.20.1) with the default parameters. The remaining data were aligned to Pathogenic Microbial Genome Databases consisting of bacteria, viruses, fungi, and parasites. The classification reference databases were downloaded from NCBI (ftp://ftp.ncbi.nlm.ni h.gov/genomes/) nucleotide and genome databases. The mapped data were processed for further analysis.

The analysis of the mNGS results included several stages. At least two reads were mapped to pathogens, excluding specific pathogens [14]. The preliminary data contain the relative abundance of microbes detected using mNGS in each microbe's threshold for further test validation, as follows: (1) pathogens with the highest absolute abundance in their genus (Top 5); (2) species with a relative abundance of $\geq$3 % for bacteria were considered dominant species.

$$\textbf{\textit{Absolute Abundance}} = \frac{\textbf{\textit{Read Number}}}{\textbf{\textit{Whole Genome Size}}} \times \mathbf{10^6}$$

$$\textbf{\textit{Relative Abundance}} = \frac{\textbf{\textit{Absolute Abundance}}}{\sum_{\textit{This sample}} \textbf{\textit{Absolute Abundance}}} \times \mathbf{100\%}$$

2.3. Co-abundance analysis of discriminant bacteria

To investigate the associations between the discriminant bacteria, Fast Spar (v1.0.0) was used to construct a compositionality-corrected microbial interaction network capable of estimating correlation values from compositional data. Interactions were calculated using SpaCC refining, after which the statistical significance of each interaction was estimated within 1000 permutations. We then calculated the median magnitude of the interaction partners as the combined association magnitude. Associations with false discovery rate (FDR) $<$ 0.00001 were included in the downstream analysis. The network was visualized using a digraph (v1.4.1).

2.4. Differential abundance bacteria analysis

Linear discriminant analysis effect size analysis (LEfSe, R package MicrobiotaProcess LEfSe-liked) were performed to identify LRTIs-related differential abundance bacteria according to the relative abundance. Relative abundance of a microbe in a sample was calculated as the read count normalised against the total reads in that sample. The statistical analyses were performed using a two-tailed non-parametric Kruskal–Wallis H-test to evaluate the significance of differences in microbial taxa between groups. The Wilcoxon rank sum test was used to compare the crucial taxa between groups. Species with false discovery rate (FDR) $<$ 0.01 and LDA score $\leq$2 was determined as discriminant bacteria. Complete linkage clustering based on the species composition and abundance of bacterial communities that define community clusters I to V. Three cohorts were generated according to the bacterial mean abundance

Table 1
Basic clinical characteristics of 394 patients with LRTIs.

Characteristics	Patients with LRTIs (n = 394, %)
Demographics	
Age, years, n (%)	Mean age (60.1)
≤60	160 (41.6 %)
>60	234 (59.4 %)
Gender, n (%)	
Male	226 (57.4 %)
Female	168 (43.6 %)
Clinical symptoms, n (%)	266 (67.5 %)
Fever	67 (25.2 %)
Cough	160 (60.1 %)
Breathing difficulty	39 (14.7 %)
Comorbidities, n (%)	184 (46.7 %)
Cardiovascular disease	9 (4.9 %)
History of lung disease	40 (21.7 %)
Diabetes	35 (19.1 %)
Chronic obstructive pulmonary disease	21 (11.4 %)
Hypertension	31 (16.8 %)
Renal disease	10 (5.4 %)
Bronchiectasis	38 (20.7 %)

Abbreviations: LRTIs, lower respiratory tract infections.

and detection frequency in each cluster. Receiver operating characteristics (ROC) curve (R version 3.6.0, pROC package version 1.18.0) analysis was performed to predict distinguish capability based on SDSMRN from mNGS results. The area under the parametric curve (AUC) was computed by numerical integration using the R software pROC package to represent the ROC effect according to the protocols established in previous study [15] The calculated AUC as a quantitative measure of the discrimination power between patients and health groups. The results were presented with 95 % confidence interval (CI).

2.5. Statistical analysis

SPSS 21.0 (IBM Corp.: Armonk, NY, US) was used for clinical data analysis. Statistical significance was established for a value of $p < 0.05$, using two-tailed tests of hypotheses. Two groups comparison were analyzed by Wilcoxon rank sum. All statistical analyses were performed using the R version 4.1.2 software.

3. Results

3.1. Basic characteristics of patients with LRTIs

In total, 394 patients (57.4 % male, 43.6 % female) with LRTIs were recruited for mNGS-based sample analyses (Table 1, Table S1). The mean age was 60.1 years, with patient age predominantly exceeding 60 years (59.4 %). Among them, 184 (46.7 %) patients possessed comorbidities, primarily including 35 cases of diabetes (19.1 %), 40 (21.7 %) cases with a history of lung disease, and 38 (20.7 %) cases of bronchiectasis. We focused on samples obtained using Illumina shotgun sequencing to accurately profile bacteria at the species level by reanalysis of the METAnnotatorX2 platform. To reduce technical bias in the bioinformatic analysis, all raw mNGS data were processed.

3.2. Community composition of LRTIs-associated differential bacteria

To explore the differences in the respiratory microflora, we assessed the relative proportions of bacterial composition in patients using mNGS. The bubble plot provides a detailed overview of all 1208 bacterial species detected in patients with LRTIs (Fig. 1) based on the relative abundance, frequency, and standard deviation (SD). The top 10 most frequently detected bacteria were *Pseudomonas aeruginosa* (94.07 %), *Cutibacterium acnes* (92.31 %), *Staphylococcus aureus* (90.33 %), *Pseudomonas fluorescens* (89.67 %), *Klebsiella pneumoniae* (87.25 %), *Stenotrophomonas maltophilia* (86.59 %), *Pseudomonas stutzeri* (86.59 %), *Acinetobacter baumannii* (79.78 %), *Moraxella osloensis* (78.68 %), and *Achromobacter xylosoxidans* (74.95 %). The top 10 most abundant bacteria were *Mycoplasma pneumoniae* (7.20 %), *Corynebacterium striatum* (2.25 %), *Chlamydia psittaci* (2.23 %), *Tropheryma whipplei* (1.81 %), *Acinetobacter baumannii* (1.81 %), *Nocardia cyriacigeorgica* (1.15 %), *Pseudomonas aeruginosa* (0.93 %), *Bacillus thuringiensis* (0.83 %), *Pseudomonas fluorescens* (0.81 %), and *Mycobacteroides abscessus* (0.71 %).

To distinguish between pathogens and colonization, we performed LEfSe analysis to compare the differential abundance bacteria. Furthermore, for comparison, we assessed the relative proportions of bacterial composition in samples from healthy people using mNGS. In total, 443 differential bacteria exhibited substantially differential abundances between patients (Fig. 1A) and controls

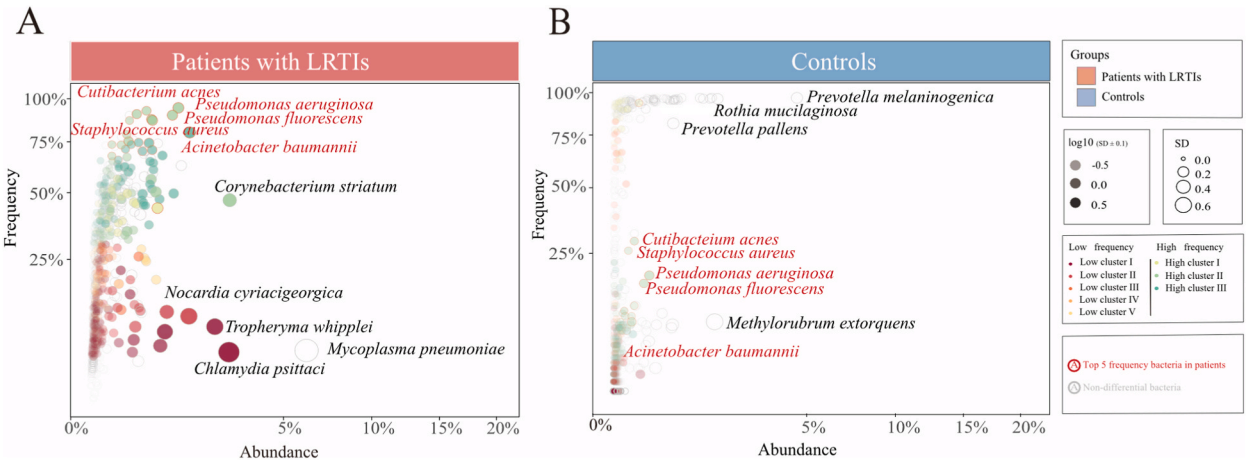


Fig. 1. Frequency and abundance of bacteria in patients with LRTIs. Linear discriminant analysis effect size (LEfSe) based on the relative abundance was used to differentiate between the bacteria of (**A**) patients with LRTIs and (**B**) controls. The dots in the double logarithmic graphs symbolize the identification frequency of the bacteria and their mean relative abundance. The distinct colors of the nodes indicate the differential abundance bacteria in each cluster (High cluster I, II, III; Low cluster I–V)from the high- or low-frequency groups. The taxon is specified for the top 5 abundant bacteria (black font) or most frequent species (red font) in patient group. Node size depicts the standard deviation (SD) of each bacteria. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

(Fig. 1B) according to the FDR <0.01 and LDA score ≥2. To further explore the suitable conditions for correctly identifying pathogens amid abundant and heterogeneous populations of commensals, we divided the discriminant bacteria into the high-frequency (frequency ≥30 %) and low-frequency (frequency <30 %) groups.

The bacteria were clustered based on their similar correlation profiles. By clustering taxa, the three clusters in the high-frequency group (Fig. 1A, green dot) were designated as high clusters I, II, and III, containing 142 discriminant bacteria (Table S2). Of these discriminatory taxa, we analyzed the mean abundance and mean frequency of discriminatory bacteria in each cluster (Table 2). Bacteria in high cluster I primarily belonged to the colonizing species and demonstrated a lower frequency in the patients than in the controls (42.0 % < 89.7 %). Conversely, bacteria in high clusters II and III were more likely to be emerging opportunistic pathogens, and the detection frequency was considerably higher in the patients than that in the healthy group, including widely reported pathogens such as *P. aeruginosa*, *S. maltophilia*, *S. aureus*, and *K. pneumoniae*.

The clusters in the low-frequency group (Fig. 1A, red dot) were designated as low clusters I–V, containing 301 bacteria (Table S2). Bacteria in the low-frequency group were mostly barely reported and were easily ignored because of their uncommon or unclear clinical importance, unless the most abundant species were identified. Notably, 16 bacteria in low cluster III presented a considerably higher detection frequency in the healthy participants than in the patients (65.8 % vs. 23.4 %) and may have been colonizing species. Conversely, bacteria in the low clusters I, II, IV, and V were likely to be rare pathogens, with the detection frequency being considerably higher in patients than in controls, including the most abundant rare pathogens such as *C. psittaci*, *T. whipplei*, and *L. pneumophila*.

3.3. Distinguishing opportunistic pathogens from the respiratory tract microbiome based on the mNGS results

Based on the distribution characteristics of opportunistic pathogens and airway microbiomes, detection frequency, and relative abundance, we attempted to convert the differential species of the clusters into three cohorts that were relatively easy to interpret (Fig. 2A, Table S3). Cohort 1 (high cluster I and low cluster III) was defined as primary colonization bacteria by the co-occurrence of *Streptococcus*, *Actinomyces*, and *Porphyromonas*, and the mean detection frequency of 39 bacteria was decreased in patients with LRTIs compared to controls (34.4 % vs. 79.9 %). The mean abundance was low in both the groups (patients: 0.08 %, controls: 0.01 %; Fig. 2B).

Cohort 2 (high cluster II, III) was characterized by emerging opportunistic pathogens by the co-occurrence of *Corynebacterium*, *Pseudomonas*, *Staphylococcus*, *Stenotrophomonas*, and *Klebsiella*, which are conditionally pathogenic in healthy individuals but potentially important pathogens in patients with LRTIs. The mean detection frequency of 119 bacteria was elevated in patients with LRTIs compared to controls (51.7 % vs. 6.8 %, $P < 0.001$), and the mean abundance was substantially higher in LRTIs patients than that in controls (0.19 % vs. 0.01 %, $P < 0.001$).

Cohort 3 (low clusters I, II, IV, and V) was marked by the co-occurrence of *Chlamydia*, *Tropheryma*, and *Legionella*, which are mostly conditionally rare pathogens that barely exist in controls. The detection frequency of 285 bacteria was higher in patients with LRTIs than that in controls (11.4 % vs. 2.1 %, $P < 0.001$). In addition, the mean abundance was higher in LRTIs patients than that in controls (0.05 %vs. 0.00 %, $P < 0.001$).

As noted above, these cohorts may be readily distinguished based on detection frequency and relative abundance. ROC curves, as described previously [15], were used to identify the predictive performance of potential pathogens from airway microbiomes in the human respiratory tract. From the ROC curves of the three cohorts, the AUC value was the largest in cohort 2 (0.976), followed by

Table 2
Mean abundance and mean frequency of differentially expressed bacteria between the two groups.

Clusters	Numbers	Mean abundance (Minima to Maxima)			Mean frequency % (Minima to Maxima)		
High frequency group		Patients with LRTIs	Controls	*P* value	Patients with LRTIs	Controls	*P* value
High cluster I	23	0.105 (0.012–0.552)	0.015 (0.003–0.096)	*P < 0.001*	42.0 (30.3–56.0)	89.7 (65.0–98.4)	*P < 0.01*
High cluster II	54	0.185 (0.007–2.252)	0.02 (0.002–0.175)	*P < 0.001*	51.9 (30.1–94.1)	8.5 (1.0–29.9)	*P < 0.001*
High cluster III	65	0.186 (0.004–1.167)	0.008 (0.002–0.061)	*P < 0.001*	51.6 (30.3–79.8)	5.0 (0.5–16.0)	*P < 0.001*
Low frequency group							
Low cluster I	174	0.045 (0.002–2.227)	0.005 (0.002–0.106)	*P < 0.001*	8.9 (3.1–29.7)	1.68 (0.5–14.4)	*P < 0.001*
Low cluster II	60	0.074 (0.005–1.149)	0.005 (0.002–0.024)	*P < 0.001*	12.9 (1.0–29.0)	2.3 (0.5–8.6)	*P < 0.001*
Low cluster III	16	0.048 (0.008–0.297)	0.007 (0.002–0.031)	*P < 0.001*	23.4 (11.6–29.9)	65.8 (34.4–85.6)	*P < 0.001*
Low cluster IV **symptoms, n (%)**	30	0.021 (0.007–0.113)	0.003 (0.002–0.005)	*P < 0.001*	18.0 (8.6–28.1)	3.0 (0.5–12.3)	*P < 0.001*
Low cluster V	21	0.105 (0.004–0.531)	0.003 (0.002–0.007)	*P < 0.001*	19.1 (7.7–28.1)	1.0 (0.5–2.1)	*P < 0.001*

Wilcoxon rank sum test. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

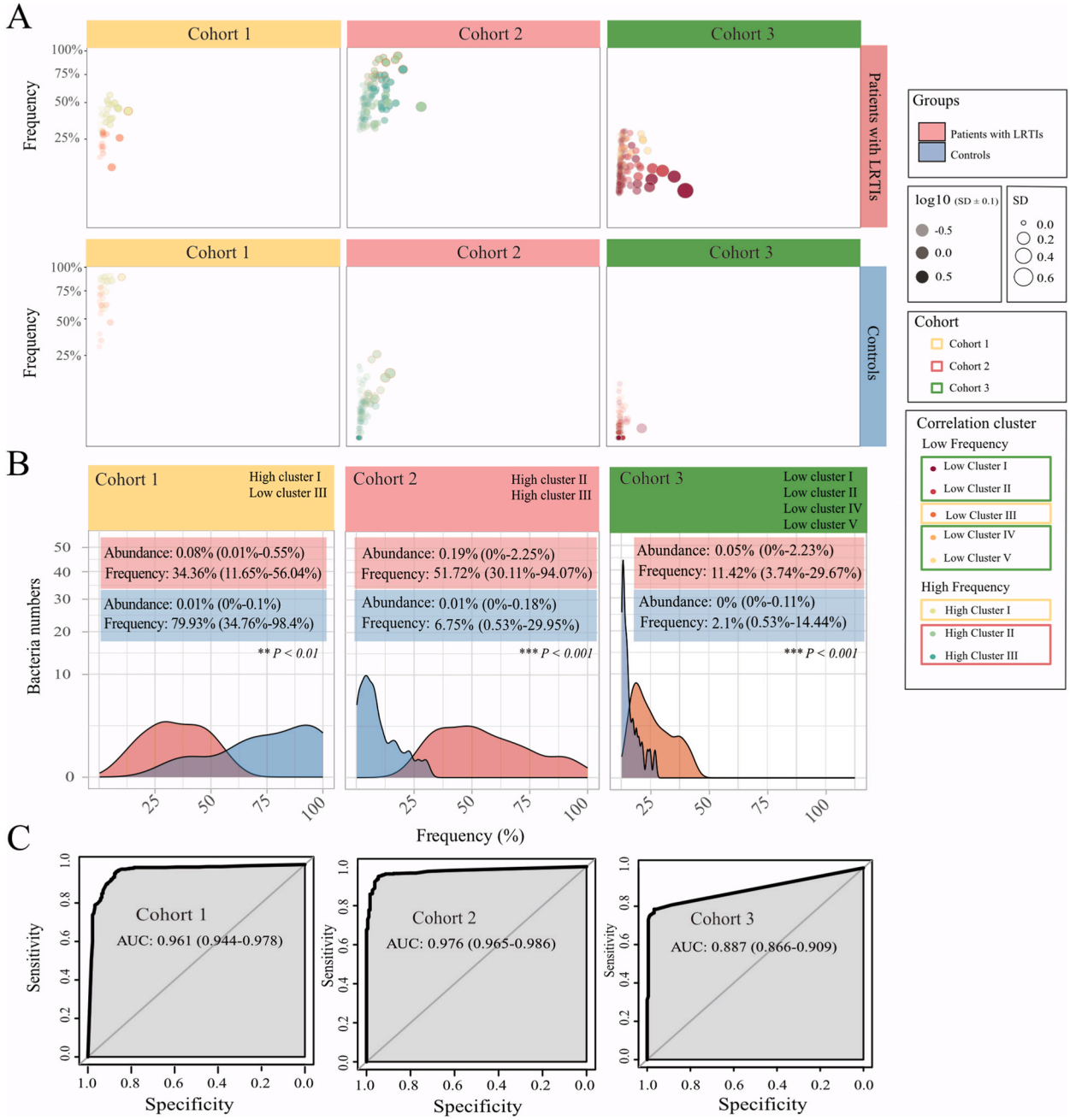
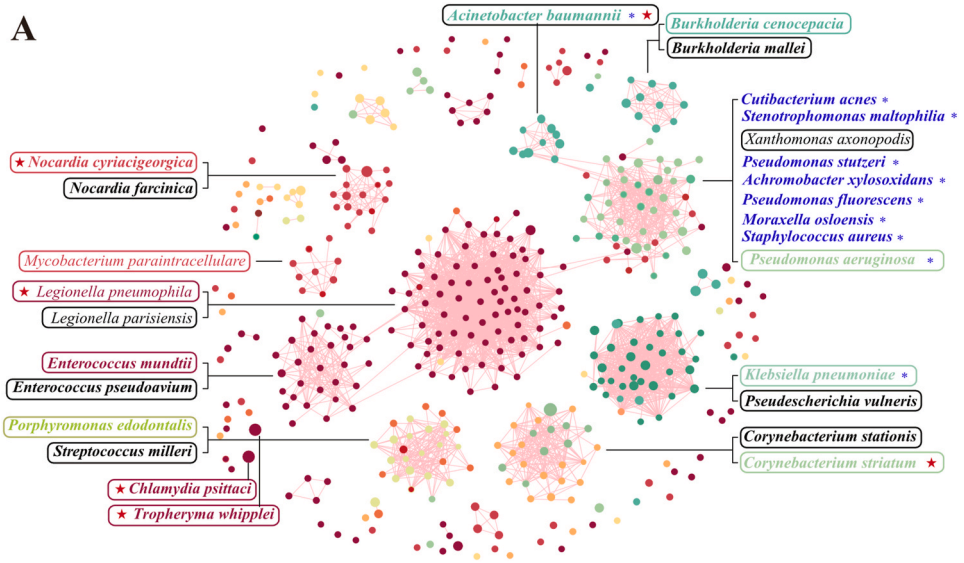
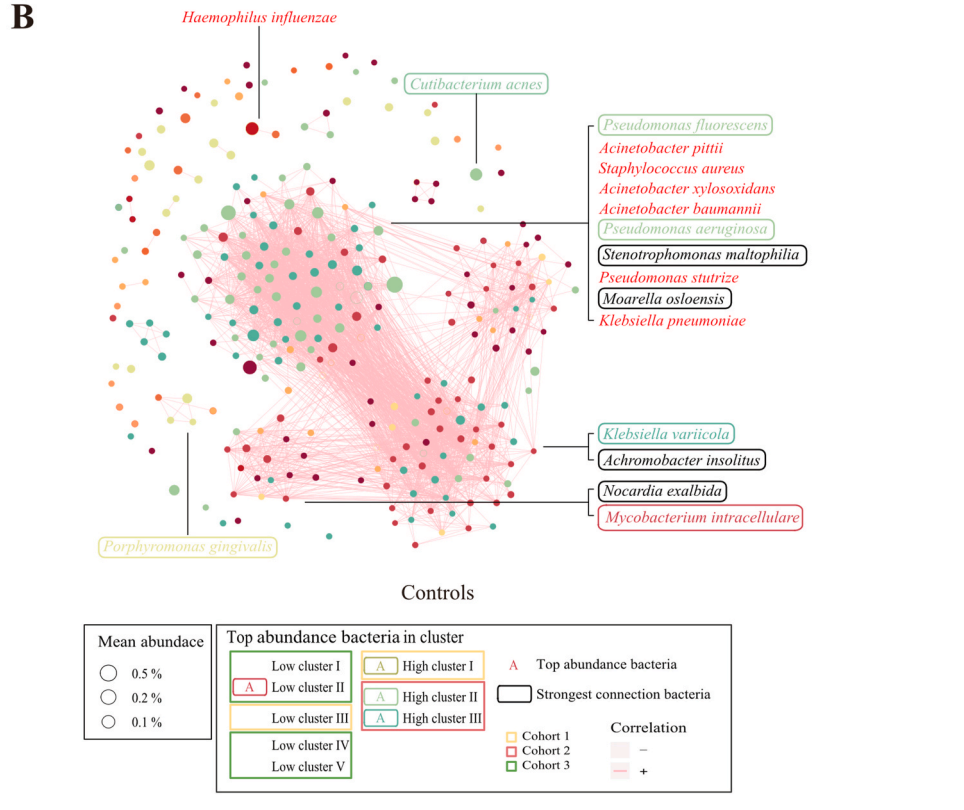


Fig. 2. Distinguishing opportunistic pathogens identified in the airway microbiomes in patients with LRTIs. (**A**) The distribution characteristics of differential bacteria. The dots in the graphs symbolize the identification frequency of the bacteria and their mean relative abundance. The distinct colors of the nodes indicate the differential abundance bacteria in each cluster. Bacterial taxa in each cluster belonging to the same cohort were indexed by matching color. The size of the circles roughly represents the value of standard deviation (SD). (**B**) Comparison between the distribution numbers of differential bacteria in cohorts 1, 2, and 3. (**C**) Receiver operating characteristic (ROC) curve analysis of bacteria in the three cohorts to demonstrate their diagnostic performance in detecting pathogens in patients. ROC plot for cohorts 1, area under the parametric curve (AUC) value = 0.961. ROC plot for cohorts 2, AUC value = 0.976. ROC plot for cohorts 3, AUC value = 0.887. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

cohorts 1 (0.961) and 3 (0.887) (Fig. 2C). As the models constructed with these three cohorts were all effective, as a guideline, our analysis might be helpful for mNGS reports and clinicians to make a judgment between colonization or infection when detecting rarely reported bacteria.





(caption on next page)

Fig. 3. Co-abundance correlations among differential bacteria in patients with LRTIs. Co-abundance networks involving combined differential species from 443 discriminatory bacteria in (**A**) patients with LRTIs and (**B**) controls. Only significant (FDR <0.00001, two-sided tests of 1000 permutations) absolute correlations exceeding 0.3 (Spearman), which are considered as fair have been illustrated. The most abundant bacteria in each unit are depicted by coloured box. The strongest connection bacteria in each unit are depicted by black box. The blue fonts with"*" indicate the top 10 frequency bacteria in all the patients' samples. The node colors and sizes indicate bacteria from different clusters (high cluster I, II, III; low cluster I–V) and the mean abundance of each bacteria, respectively. The red lines symbolize positive species interactions. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.4. Alterations in the differential bacterial co-abundance network in patients with LRTIs

Based on the above classified colonizers and opportunistic pathogens in the cohorts, we further explored the potential the interaction of microorganisms in the development and progression of LRITs. To gain insights into the potential interplay between microbial interactions and their role in LRITs pathogenesis, we performed a co-abundance association analysis based on the 443 differential abundance bacteria (Fig. 3). Notably, differential bacteria exhibited modular patterns with numerous unique units in the patients' group (Fig. 3A), which differed from the uniform distribution observed in healthy group with a stable unit (Fig. 3B). Our findings revealed that the ecological network of LRITs patients were more complex than those of healthy individuals. It revealed the dysbiotic respiratory microbiomes in LRITs patients, and accompanied with the microbial community structure changes and decreased diversity.

Significantly different high-frequency or high-abundance bacteria may correlate with changes in human health or disease risk. We further marked out the most abundant bacteria (coloured box) and the strongest correlation bacteria (black box) in each unique unit, as well as the top 10 frequency bacteria (blue font *), as illustrated in Fig. 3. It was observed that bacteria with a higher abundance or frequency does not necessarily imply the stronger correlation species within each community, making it challenging to accurately define the core pathogen. However, some highly abundant bacteria, such as the top 6 high abundance species (*, Fig. 3A) including *C. striatum*, *C. psittaci*, *T. whipplei*, *A. baumannii*, *N. cyriacigeorgica* and *P. aeruginosa*, are clinically considered pathogens, which should be concerned. Furthermore, our study demonstrated that in cohort 2, there was a strong correlation between the top 10 frequency bacteria, including pathogens *P. aeruginos* , *S. aureus* , *S. maltophilia*, *A. baumannii* and *Achromobacter xylosoxidans*, suggesting possible co-infections with one or a few pathogens. These results indicate that the microbial interactions, especially within the same community, probably play a crucial role in microbiota homeostasis and LRITs pathogenesis.

4. Discussion

Microorganisms in the human respiratory system are relatively complex because the respiratory tract is exposed to the ambient atmosphere [16,17]. Concerns for clinicians remain relatively poorly investigated, including the identification of pathogens from colonization, standardization of the mNGS method, and identification of responsible pathogens [18–20]. In this study, we explored approaches to assess the respiratory colonization microbiome and opportunistic pathogens according to clustering taxa, detection frequency, and relative abundance of species based on a large sample statistical analysis. Our findings provide insights into potential pathogen evaluation scopes that contribute to the interpretation of mNGS report results and assist in improving diagnosis of LRTIs due to unknown pathogens.

Microbiome diversity in patients with LRTIs is an active area of research. Previous studies mostly focused on microbiome differences and only the most frequent or most abundant bacteria in patients with LRTIs [21]. There has been little research on distinguishing pathogens based on a comprehensive analysis of detection frequency, abundance, and standard deviation. Discriminating respiratory pathogens from background commensal microbiota is a key challenge in LRTIs diagnostics and is particularly relevant for sensitive molecular assays [22]. A previous study showed that species variance analysis [23] and ROC curve analysis [21] could effectively distinguish significantly altered species. Our study identified 443 significantly abundant bacterial species relative to the LRTIs by summarizing the distribution regulars of microbiota in clusters according to detection frequency, relative abundance, and species corrections. Owing to the lack of complete public metadata, such as BALF samples from healthy older adults, few studies have performed in-depth multiple correlations with the composition of the LRTIs microbiome. In this context, the predominance of data derived from lavage samples associated with LRTIs, specific studies assessing bacterial differences between LRTIs and healthy individuals using mNGS approaches are expected to contribute to the identification of bacteria that are involved in or associated with the onset of human respiratory tract pathologies.

A prior study employed mNGS to screen tracheal aspirates from 92 adults presenting with acute respiratory failure [15]. They developed an approach based on pathogens, microbiome diversity, and host gene expression metrics to identify LRTIs-positive patients and differentiate them from critically ill controls with non-infectious acute respiratory illnesses. An ROC curve was used to assess the pathogens, airway microbiome, and host transcriptome. Similarly, in our study, we discovered that combining the bacterial detection frequency, abundance, and SD value could help distinguish pathogens from the airway microbiome between patients with LRTIs and healthy individuals. If a bacterium was significantly more abundant in the patients than in healthy individuals, it was defined as an opportunistic pathogen; depending on its detection frequency, it is regarded as a common or rare pathogen. If the abundance of the bacterium does not differ between patients and healthy individuals, but the detection frequency is higher in healthy individuals than that in patients, it is considered a colonization bacterium. Notably, the species in cohort 1 were mostly consistent with the airway microbiota in previous reports [15,16,24]. Opportunistic pathogens in cohort 2 with high abundance and frequency included many widely reported pathogens from previous research on infectious diseases [2,20,25]. Interestingly, the species in cohort 3 were almost only present in patients and at a low frequency, including rarely reported definitive pathogens such as *C. psittaci*, *N. cyriacigeorgica*, and

T. whipplei. We comprehensively provided detailed metadata on all the differential species in these three cohorts (Table S3). Thus, our study provides physicians with a good understanding of pathogens and colonization according to the abundance and frequency of respiratory microbiota. Our findings could serve as a reference for host and LRTIs microbial abundance thresholds, assisting clinical diagnosis and mNGS report interpretation.

The most abundant and most frequent microbe in patients, potentially were considered as pathogens [15]. Co-abundance analysis substantiated that the high-abundance pathogens did not necessarily represent the core pathogens. Intriguingly, models based on differential bacterial species have depicted various individual core units in patients. Microorganisms do not exist in isolation and invariably exhibit direct or indirect communication among themselves [26]. Prior research has revealed that bacteria such as *M. pneumoniae* are capable of depleting other bacterial populations through direct competition and activating bacterial clearance factors during host responses [27]. The cooperative and competitive interactions in microbiome may impact the resulting community structure and composition, which could lead to microbial community changes and decreased diversity [28]. High-frequency pathogens in this study interact with other members of the respiratory microbiome in ways that may facilitate infection or exacerbate its severity. A comparable mechanism was conjectured to take effect in the high-frequency pathogens, including *C. striatum*, *K. pneumoniae*, *A. baumannii*, and *P. aeruginosa,* in each unit. Although the precise mechanisms underlying sets in the respiratory system remain ambiguous, the current study findings provide innovative insights into other pathogens exhibiting similar interaction models in LRTIs. Consequently, exploring species associations within each cohort throughout the development and progression of LRTIs would be a fascinating endeavor. Understanding bacterial-bacterial interactions and the polymicrobial community in the respiratory tract could be informative in understanding how the microbiome contributes to disease progression and how to preserve microbial diversity. Targeted interventions, such as viral/bacterial vaccination, antibiotics, high-frequency/high-abundance pathogens, prebiotics, or other microbiome-modulating agents, could potentially to be explored to maintain the respiratory microbial balance and prevent LRTIs in various population groups.

Our study had several limitations. The retrospective study design did not allow the application of PCR, complement-fixation testing, or micro-immunofluorescence to confirm the mNGS results. Due to lacking participants' living environment information, this study did not consider the relationship of air pollution with dysbiosis of respiratory microbiome. Future research could collect patients' samples and environmental samples in a certain region or community to explore the impact of air pollution on LRTIs. The healthy group comprised only eight BALF samples. Consequently, a constraint was imposed by the difficulty of acquiring additional BALF samples from older, healthy individuals. Nevertheless, the present study primarily focused on delineating the detection frequency and abundance of pathogens in patients with LRTIs. Future study could expand larger cohorts from diverse geographic regions based on the prospective multi-center cohort study or the retrospective analysis of publicly available shotgun metagenomic datasets. The burgeoning comprehension of the detection threshold of the aforementioned crucial microbiomes in patients with LRTIs could facilitate the formulation of guidelines for LRTIs diagnosis, interpret mNGS results, and inspire investigations into potential clinical applications of mNGS. The pathogen characterization and colonization in our cohort classification warrant subsequent investigations into specific respiratory infectious diseases for verification. In future study, we could combine the epigenetic factors including air pollution, vaccine strategy and standardized mNGS analysis, to further explore the respiratory microbiome differences and main pathogens among different populations in different geographic regions. By adopting these strategies, it is expect to improve the diagnosis, treatment strategies, and prevention of respiratory diseases.

5. Conclusions

The current study documents a potential approach based on discriminant bacterial correlation profiles, detection frequency, and relative abundance to distinguish pathogens from the respiratory tract microbiome and assist in LRTIs diagnosis. A profile of LRTIs-associated bacterial colonization or opportunistic pathogens from a relatively large-scale cohort may serve as a reference panel for the interpretation of mNGS results in clinical practice.

Ethical statement

This study was approved by the Ethics Committee of Huayin Medical Laboratory Center (number: 2022-004-02). The participants provided written informed consent to participate in this study.

Data availability statement

The raw data are available online on the National Center of Biotechnology Information database under accession number PRJNA993600 and are available at the following link: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA993600.

Funding

This work was supported by the Key Research and Development Program of Jiangsu Province of China (BE2021707).

CRediT authorship contribution statement

Tingyan Dong: Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Conceptualization. **Yueming**

Liang: Resources, Investigation, Data curation. **Junting Xie:** Visualization, Software, Methodology, Formal analysis, Data curation. **Wentao Fan:** Validation, Software, Methodology. **Haitao Chen:** Software, Resources, Methodology, Investigation. **Xiaodong Han:** Validation, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank all patients and their families who participated in this study. The authors thank Dr. Yuhua Ye for helpful suggestions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e30896.

References

[1] H. Raghav, P. Tayal, R. Das, D.K. Mehta, Appropriate use of antibiotics for the management of respiratory tract infections, Infect. Disord.: Drug Targets 22 (5) (2022) e180122200335, https://doi.org/10.2174/1871526522666220118122516.
[2] Z. Zhao, J. Song, C. Yang, et al., Prevalence of fungal and bacterial Co-infection in pulmonary fungal infections: a metagenomic next generation sequencing-based study, Front. Cell. Infect. Microbiol. 11 (2021) 749905, https://doi.org/10.3389/fcimb.2021.749905.
[3] C.A. Glaser, S. Honarmand, L.J. Anderson, et al., Beyond viruses: clinical profiles and etiologies associated with encephalitis, Clin. Infect. Dis. 43 (12) (2006) 1565–1577, https://doi.org/10.1086/509330.
[4] R. Schlaberg, C.Y. Chiu, S. Miller, et al., Validation of metagenomic next-generation sequencing tests for universal pathogen detection, Arch. Pathol. Lab Med. 141 (6) (2017) 776–786, https://doi.org/10.5858/arpa.2016-0539-RA.
[5] D.W. Eyre, Infection prevention and control insights from a decade of pathogen whole-genome sequencing, J. Hosp. Infect. 122 (2022) 180–186, https://doi.org/10.1016/j.jhin.2022.01.024.
[6] P. Zhang, Y. Chen, S. Li, et al., Metagenomic next-generation sequencing for the clinical diagnosis and prognosis of acute respiratory distress syndrome caused by severe pneumonia: a retrospective study, PeerJ 8 (2020) e9623, https://doi.org/10.7717/peerj.9623.
[7] L.Y. Guo, W.Y. Feng, X. Guo, B. Liu, G. Liu, J. Dong, The advantages of next-generation sequencing technology in the detection of different sources of abscess, J. Infect. 78 (1) (2019) 75–86, https://doi.org/10.1016/j.jinf.2018.08.002.
[8] C.A. Hogan, S. Yang, O.B. Garner, et al., Clinical impact of metagenomic next-generation sequencing of plasma cell-free DNA for the diagnosis of infectious diseases: a multicenter retrospective cohort study, Clin. Infect. Dis. 72 (2) (2021) 239–245, https://doi.org/10.1093/cid/ciaa035.
[9] H. Chen, Y. Yin, H. Gao, et al., Clinical utility of in-house metagenomic next-generation sequencing for the diagnosis of lower respiratory tract infections and analysis of the host immune response, Clin. Infect. Dis. 71 (Suppl 4) (2020) S416–S426, https://doi.org/10.1093/cid/ciaa1516.
[10] N. Li, Q. Cai, Q. Miao, Z. Song, Y. Fang, B. Hu, High-throughput metagenomics for identification of pathogens in the clinical settings, Small Methods 5 (1) (2021) 2000792, https://doi.org/10.1002/smtd.202000792.
[11] P.J. Simner, S. Miller, K.C. Carroll, Understanding the promises and hurdles of metagenomic next-generation sequencing as a diagnostic tool for infectious diseases, Clin. Infect. Dis. 66 (5) (2018) 778–788, https://doi.org/10.1093/cid/cix881.
[12] H. Liu, Y. Zhang, J. Yang, Y. Liu, J. Chen, Application of mNGS in the etiological analysis of lower respiratory tract infections and the prediction of drug resistance, Microbiol. Spectr. 10 (1) (2022) e0250221, https://doi.org/10.1128/spectrum.02502-21.
[13] L. Chao, J. Li, Y. Zhang, H. Pu, X. Yan, Application of next generation sequencing-based rapid detection platform for microbiological diagnosis and drug resistance prediction in acute lower respiratory infection, Ann. Transl. Med. 8 (24) (2020) 1644, https://doi.org/10.21037/atm-20-7081.
[14] Y. Liang, T. Dong, M. Li, et al., Clinical diagnosis and etiology of patients with Chlamydia psittaci pneumonia based on metagenomic next-generation sequencing, Front. Cell. Infect. Microbiol. 12 (2022) 1006117, https://doi.org/10.3389/fcimb.2022.1006117.
[15] C. Langelier, K.L. Kalantar, F. Moazed, et al., Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults, Proc. Natl. Acad. Sci. U.S.A. 115 (52) (2018) E12353–E12362, https://doi.org/10.1073/pnas.1809700115.
[16] L. Mancabelli, C. Milani, F. Fontana, et al., Mapping bacterial diversity and metabolic functionality of the human respiratory tract microbiome, J. Oral Microbiol. 14 (1) (2022) 2051336, https://doi.org/10.1080/20002297.2022.2051336.
[17] M. Sommariva, V. Le Noci, F. Bianchi, et al., The lung microbiota: role in maintaining pulmonary immune homeostasis and its implications in cancer development and therapy, Cell. Mol. Life Sci. 77 (14) (2020) 2739–2749, https://doi.org/10.1007/s00018-020-03452-8.
[18] Y. Zheng, X. Qiu, T. Wang, J. Zhang, The diagnostic value of metagenomic next-generation sequencing in lower respiratory tract infection, Front. Cell. Infect. Microbiol. 11 (2021) 694756, https://doi.org/10.3389/fcimb.2021.694756.
[19] X. Wu, Y. Li, M. Zhang, et al., Etiology of severe community-acquired pneumonia in adults based on metagenomic next-generation sequencing: a prospective multicenter study, Infect. Dis. Ther. 9 (4) (2020) 1003–1015, https://doi.org/10.1007/s40121-020-00353-y.
[20] Z. Peng, J. Zhou, L. Tian, Pathogenic characteristics of sputum and bronchoalveolar lavage fluid samples from patients with lower respiratory tract infection in a large teaching hospital in China: a retrospective study, BMC Pulm. Med. 20 (1) (2020) 233, https://doi.org/10.1186/s12890-020-01275-8.
[21] C. Langelier, M.S. Zinter, K. Kalantar, et al., Metagenomic sequencing detects respiratory pathogens in hematopoietic cellular transplant patients, Am. J. Respir. Crit. Care Med. 197 (4) (2018) 524–528, https://doi.org/10.1164/rccm.201706-1097LE.
[22] J.M. Walter, R.G. Wunderink, Severe respiratory viral infections: new evidence and changing paradigms, Infect. Dis. Clin. 31 (3) (2017) 455–474, https://doi.org/10.1016/j.idc.2017.05.004.
[23] N.N. Liu, N. Jiao, J.C. Tan, et al., Multi-kingdom microbiota analyses identify bacterial-fungal interactions and biomarkers of colorectal cancer across cohorts, Nat Microbiol 7 (2) (2022) 238–250, https://doi.org/10.1038/s41564-021-01030-7.
[24] R.P. Dickson, J.R. Erb-Downward, C.M. Freeman, et al., Bacterial topography of the healthy human lower respiratory tract, mBio 8 (1) (2017), https://doi.org/10.1128/mBio.02287-16.
[25] T. Doan, M.R. Wilson, E.D. Crawford, et al., Illuminating uveitis: metagenomic deep sequencing identifies common and rare pathogens, Genome Med. 8 (1) (2016) 90, https://doi.org/10.1186/s13073-016-0344-6.

[26] E. Kolwijck, F.L. van de Veerdonk, The potential impact of the pulmonary microbiome on immunopathogenesis of Aspergillus-related lung disease, Eur. J. Immunol. 44 (11) (2014) 3156–3165, https://doi.org/10.1002/eji.201344404.
[27] W. Dai, H. Wang, Q. Zhou, et al., An integrated respiratory microbial gene catalogue to better understand the microbial aetiology of Mycoplasma pneumoniae pneumonia, GigaScience 8 (8) (2019), https://doi.org/10.1093/gigascience/giz093.
[28] A.L. Welp, J.M. Bomberger, Bacterial community interactions during chronic respiratory disease, Front. Cell. Infect. Microbiol. 10 (2020) 213, https://doi.org/10.3389/fcimb.2020.00213.