Heliyon 10 (2024) e33429

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

The effects of sequencing strategies on Metagenomic pathogen detection using bronchoalveolar lavage fluid samples

Ziyang Li [a,b], Zhe Guo [a,b], Weimin Wu [a,b], Li Tan [a,b], Qichen Long [a,b], Han Xia [c,d], Min Hu [a,b,*]

[a] Department of Laboratory Medicine, The Second Xiangya Hospital, Central South University, Changsha, Hunan 410011, China
[b] Center for Clinical Molecular Diagnostics, The Second Xiangya Hospital, Central South University, Changsha, Hunan 410011, China
[c] School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[d] MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Keywords:
mNGS
Pathogen detection
Read length
Dataset size
Bioinformatics analysis

ABSTRACT

Objectives: Metagenomic next-generation sequencing (mNGS) is a powerful tool for pathogen detection. The accuracy depends on both wet lab and dry lab procedures. The objective of our study was to assess the influence of read length and dataset size on pathogen detection.
Methods: In this study, 43 clinical BALF samples, which tested positive via clinical mNGS and were consistent with the diagnosis, were subjected to re-sequencing on the Illumina NovaSeq 6000 platform. The raw re-sequencing data, consisting of 100 million (M) paired-end 150 bp (PE150) reads, were divided into simulated datasets with eight different data sizes (5 M, 10 M, 15 M, 20 M, 30 M, 50 M, 75 M, 100 M) and five different read lengths (single-end 50 bp (SE50), SE75, SE100, PE100, and PE150). Both Kraken2 and IDseq bioinformatics pipelines were employed to analyze the previously diagnosed pathogens in the simulated data. Detection of pathogens was based on read counts ranging from 1 to 10 and RPM values ranging from 0.2 to 2.
Results: Our results revealed that increasing dataset sizes and read lengths can enhance the performance of mNGS in pathogen detection. However, a larger data sizes for mNGS require higher economic costs and longer turnaround time for data analysis. Our findings indicate 20 M reads being sufficient for SE75 mode to achieve high recall rates. Additionally, high nucleic acid loads in samples can lead to increased stability in pathogen detection efficiency, reducing the impact of sequencing strategies. The choice of bioinformatics pipelines had a significant impact on recall rates achieved in pathogen detection.
Conclusions: Increasing dataset sizes and read lengths can enhance the performance of mNGS in pathogen detection but increase the economic and time costs of sequencing and data analysis. Currently, the 20 M reads in SE75 mode may be the best sequencing option.

1. Introduction

Metagenomics is the study of microbial communities by sequencing genomic fragments from biological samples. Emerging as a

* Corresponding author. Department of Laboratory Medicine, The Second Xiangya Hospital, Central South University, Changsha, Hunan 410011, China.
E-mail address: huminjyk@csu.edu.cn (M. Hu).

https://doi.org/10.1016/j.heliyon.2024.e33429
Received 3 December 2023; Received in revised form 17 June 2024; Accepted 21 June 2024
Available online 22 June 2024

powerful tool in characterizing human and environmental microbiome, the high-throughput sequencing technology has largely expanded the scope of metagenomics. In recent years, metagenomic next-generation sequencing (mNGS) has been widely used in clinical laboratories for infectious diseases diagnosis [1,2] and outbreak investigation [3,4]. With higher detection rate [5,6] and shorter turnaround time [7,8] than conventional methods, unbiased mNGS has the potential to identify all pathogens in a single test. It holds the promise of revolutionizing the landscape of clinical microbiology.

The process of mNGS includes experimental operations (wet lab) and computational analysis (dry lab) [2]. The wet lab refers to the experimental procedures involved in preparing and sequencing DNA or RNA samples, such as nucleic acid extraction, library preparation, and sequencing, and the dry lab involves computational analysis of raw sequencing data obtained from microbial samples to identify and quantify the taxonomic composition and functional potential of the microbial community [9–11]. Each step in the process affects the accuracy of pathogen detection, which is crucial for diagnosis and treatment. In addition to ensuring reliable mNGS results, cost and time savings should be considered from a practical perspective [7,12]. The cost mainly depends on sequencing dataset size and platform. Mainstream sequencing platforms generate reads with length ranging from 50 bp to 300 bp (paired-end (PE) or single-end (SE)), and the reads length determines the sequencing run time [13]. For this rapidly evolving technology, comprehensive assessments of performance, cost and turnaround time of different sequencing strategies and bioinformatics pipelines in real time will provide effective information for clinicians and patients in the choice of optimal methods.

In recent years, the field of metagenomics has witnessed significant progress in the development of tools and methods to effectively handle and analyze metagenomic data, such as the rapid evolution of the mNGS bioinformatics workflows [14–16]. These workflows encompass a wide range of tools, each serving a specific purpose [17–20]. Two noteworthy tools are Kraken2 [21] and IDseq [22]. Kraken2 employs exact k-mer matching techniques to efficiently assign taxonomic labels to DNA sequences using reference databases. Its remarkable accuracy and speed make it a suitable choice for processing large-scale datasets. Additionally, Kraken2 offers customizable options for post-processing and visualization, thereby facilitating in-depth analyses [23–25]. IDseq is an open-source platform designed for disease diagnosis and monitoring. It integrates multiple analytical steps and incorporates various algorithms and databases to provide comprehensive analysis of microbial communities present in samples [26–28]. IDseq's distinctive feature lies in its ability to fuse multiple bioinformatics tools into a user-friendly interface, making it easily accessible even to individuals without specialized expertise. These tools contribute significantly to the detection and characterization of microbial communities, ultimately leading to enhanced understanding of complex biological systems.

It is important to note that using the same mNGS procedure may introduce bias in pathogen detection of different infection types or

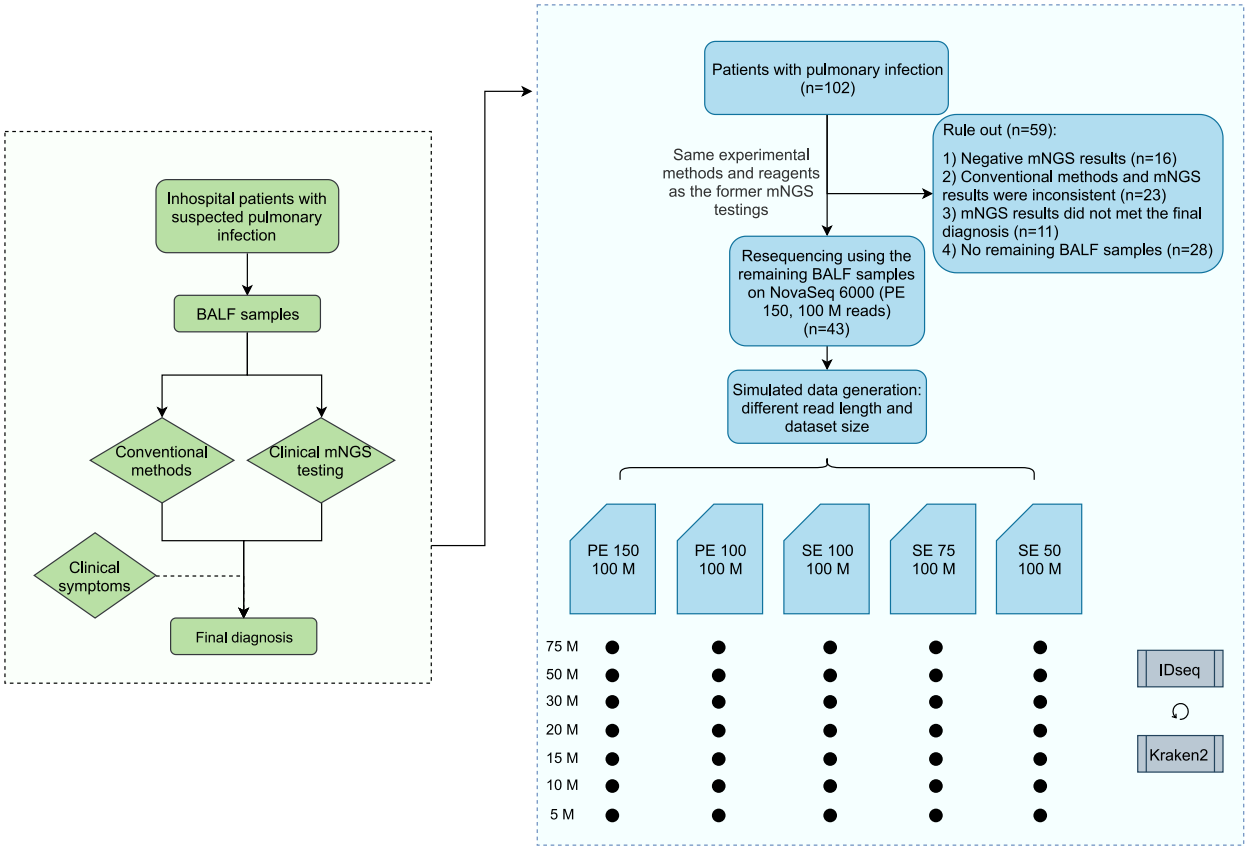

Fig. 1. The overview of study design.

sample types [11]. For instance, compared with other sample types, the positive rate of mNGS using bronchoalveolar lavage fluid (BALF) samples was relatively higher [29], which probably due to the presence of colonizers from respiratory tract. Therefore, the assessment of optimal wet lab and dry lab procedures for specific samples is usually not universal. In this study, we re-sequenced 43 previously mNGS-positive samples, with reported pathogen correlated with clinical diagnoses and conventional tests, using higher depth (100 million) and longer read lengths (paired-end 150). We processed these datasets to simulate varying read lengths and data volumes and analyzed the data using different bioinformatics pipelines to explore the effects of data volume, sequencing read length, and bioinformatics pipeline on mNGS detection.

2. Methods and materials

2.1. Study design

BALF samples from patients with pulmonary infection who had undergone mNGS and conventional tests (including culture, PCR, galactomannan antigen test, and other assays) for pathogen detection were collected in this study. The inclusion criteria were: 1) with matching detection results of mNGS and conventional methods; 2) with negative conventional method results, but the mNGS results met the final diagnosis based on clinical symptoms. Patients met the following criteria were excluded: 1) with negative mNGS results; 2) with inconsistent detection results between mNGS and conventional methods; 3) mNGS results did not met the final diagnosis; 4) with no remaining BALF samples. For each sample, causative pathogens were confirmed by clinicians, and gram-positive (G+) bacteria, gram-negative (G-) bacteria, fungi, viruses and intracellular pathogens were expected to be enrolled. The microbes were identified as causative pathogens if they met the following criteria: 1) both mNGS and conventional methods detected the identified pathogen or mNGS alone detected the microbe; 2) Previous literatures reported its pathogenicity; 3) the clinical symptoms of the patients are consistent with the characteristics of infection by the microbe; and 4) the final diagnosis by at least two independent clinicians verified the microbe. The samples were prepared for re-sequencing on Illumina NovaSeq 6000. After bioinformatics analysis, the identified microbes were compared to the previous confirmed pathogens, and the number of reads aligning to each pathogen were counted using IDseq.

To investigate whether pathogen detection would be affected by the sequencing strategies, datasets with different size and read length were simulated based on the raw reads obtained above. Different classification algorithms and reference sequence databases were used to estimate the effect of bioinformatics pipelines on pathogen detection. The specific reads aligning to each detected pathogen were counted, and the recall rates were assessed (Fig. 1).

2.2. Re-sequencing

BALF samples that met the selection criteria were retrieved from the −20 °C refrigerator for nucleic acid extraction. After thawing, 200 μL BALF were used for DNA extraction. The DNA from each sample was extracted and purified using the QIAamp DNA Micro Kit (QIAGEN, Hilden, Germany) following the manufacturer's instructions and the final elution volume was 50 μL. The concentration and quality of extraction were tested through Qubit 4.0 (Thermo Fisher Scientific, MA, USA). DNA libraries were constructed using QIAseq Ultralow Input Library Kit (QIAGEN, Hilden, Germany). Approximately 10–50 μl of DNA samples were utilized for library construction, resulting in a final library volume of 50 μl. Subsequently, the inspected libraries were sequenced on Illumina NovaSeq 6000 (Illumina, San Diego, USA) using PE150 model, with an anticipated data volume of 100 million (M) reads per sample. If the data output for a sample falls below 100 M, the DNA libraries will be reloaded for sequencing.

2.3. qPCR validation

Samples containing mNGS reports of *Acinetobacter baumannii*, *Klebsiella pneumoniae*, and *Staphylococcus aureus* underwent real-time quantitative PCR (qPCR) validation using SYBR green method. The same DNA used for mNGS re-sequencing was utilized for these experiments. Genomic DNAs extracted from *A. baumannii* ATCC 17978, *K. pneumoniae* ATCC 10031, and *S. aureus* ATCC 6538 served as controls. The quantification of genomic DNA was conducted using the Applied Biosystems™ 7500 Fast PCR System (Thermo Fisher Scientific, USA). The designed primer pairs used for detection were as follows: *S. aureus* (Forward: CCCCGCCAACTTGCACATTA, Reverse: GGTGTGGGCCCCAACATAGA), *K. pneumoniae* (Forward: ACTGCGTCTGGTGATCTACG, Reverse: GCGGAATTTCGCCCATGTAG), and *A. baumannii* (Forward: AAGGCCCTGTAGCGATCCATGC, Reverse: AAGCTGCCATCTGTGCCTAGC).

2.4. Bioinformatics analysis

Prior to formal analysis, the acquired sequencing raw data underwent filtration using fastp (v0.21.0) [30] to eliminate adapter sequences, as well as low-quality, low-complexity, and short reads. The parameters used were –cut_mean_quality 15 –length_required 45 –trim_poly_x -e 20. Bowtie2 (v2.4.2) [31] (–sensitive-local) was used to remove human reads by mapping to human reference genome (GRCh38.p13). The remaining reads obtained were used as input to IDseq (https://czid.org/) [22] or Kraken2 v2.1.1 [32] (database: pluspf_20,210,517, default parameters) to detect pathogens. Various specific sequence counts (ranging from 1 to 10 with intervals of 1) and different RPM values (ranging from 0.2 to 2 with intervals of 0.2) were individually established as the criteria for detection.

2.5. Dataset simulation

We utilized Seqtk (v1.2) (https://github.com/lh3/seqtk) to perform random subsampling of the data and employed a local Perl script to manipulate the sequencing read lengths. Based on the raw reads obtained by re-sequencing, we simulated datasets with different size of 100 M, 75 M, 50 M, 30 M, 20 M, 15 M, 10 M, 5 M reads and different read length of PE150, PE100, SE100, SE75, SE50. In the first step, 100 M reads (PE150) were randomly extracted from each sample and repeated 3 times to construct sufficient statistics. Next, the first 100 bp of PE150 reads in each dataset were truncated to generate PE100 reads, and "read-1" of PE100 reads were extracted as SE100 reads. Read length of SE75 and SE50 were simulated by truncating the first 75 bp and 50 bp of SE100 reads. Finally, each of the 100 M reads dataset (different read lengths, 3 repeats) was randomly down-sampled to 75 M, 50 M, 30 M, 20 M, 15 M, 10 M and 5 M. All of these simulated datasets were separately taken as input files to IDseq (NT and NR) and Kraken2 to detect pathogens,

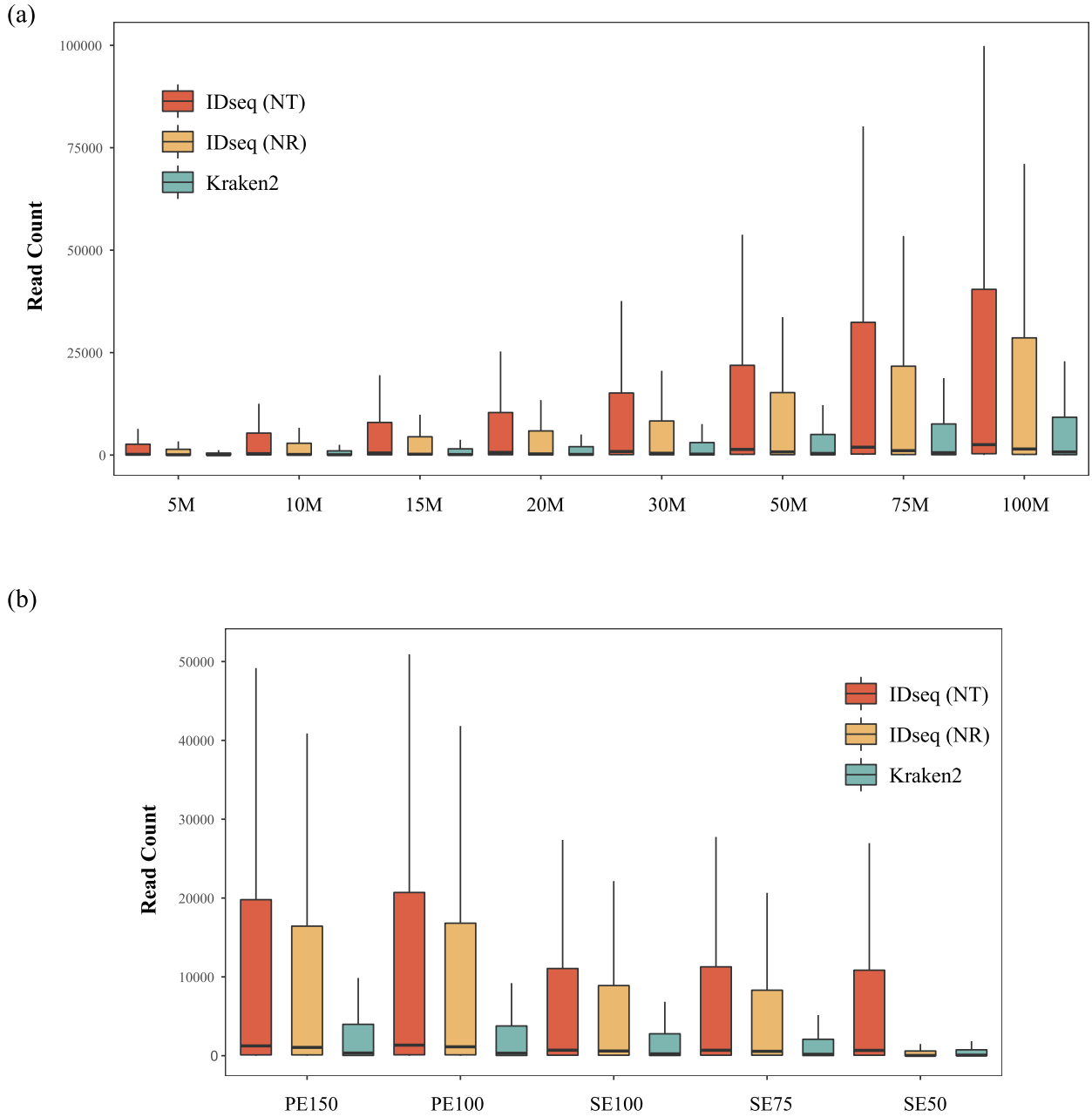

Fig. 2. Box plots illustrating the correlation between the number of pathogenic reads and both dataset size and read length. (a) The black horizontal lines within the box represent the median, while the top and bottom edges of the box represent the upper and lower quartiles. The number of pathogenic reads varied with dataset size; (b) The number of pathogenic reads varied with read length.

respectively. For each dataset, the number of reads aligning to the detected pathogens were counted. The recall performance of pathogen detection was assessed at each dataset size and each read length (3 repeats) by each bioinformatics pipeline (Fig. 1).

2.6. Statistical analysis

The chi-square test and paired samples *t*-test employed to examine the statistical variance of single or group paried data. Analysis was performed utilizing RStudio, leveraging R version 4.1.2. A p-value below 0.05 was deemed to indicate statistical significance. P-values below 0.05, 0.01, and 0.001 were denoted by "*", "**", and "***" respectively.

3. Results

3.1. Re-sequencing statistics and overview

Of the initial 102 BALF samples collected from patients with pulmonary infections, 59 were excluded for reasons including negative mNGS results (16), inconsistencies between mNGS results and conventional diagnostic methods (23), inconsistencies between mNGS results and final diagnosis (11), and insufficient samples volume (28). At last, 43 clinical BALF samples that had been tested for mNGS and conventional methods were collected (Fig. 1). These samples contained 89 causative pathogens, including 57 bacteria (including 8 *Mycobacterium tuberculosis* complex (MTBC)), 22 fungi, and 10 DNA viruses (Table S1). We re-sequenced these samples using Illumina NovaSeq 6000, and obtained 1533.4 giga (G) raw bases with an average of 113.6 million reads per sample. The Q30 ratio of the samples ranged from 84.93 % to 93.79 % (Table S1). All of the previously confirmed causative pathogens were again identified, and the number of reads aligning to each pathogen, as classified by IDseq, were counted (Table S1).

3.2. The number of pathogen reads depend on read length and dataset size

We simulated 120 datasets (8 dataset sizes, 5 read lengths, 3 repeats) for each sample, with different sizes and read lengths based on the raw reads. The number of reads classified by Kraken2 or IDseq (NT and NR) aligning to each pathogen were counted. We found that the detected pathogen read counts increased with the size of dataset, regardless of the classifier or read length. IDseq (especially IDseq NT) identified a higher number of reads for the same pathogen compared to Kraken2, which is consistent with previous reports regarding the utility of IDseq [22,33] (Fig. 2a).

Given the same dataset size, the read counts of each pathogen increased with read length and sequencing mode. The PE150 and PE100 groups identified higher number of reads than the SE100, SE75, and SE50 groups. However, we did not observe any significant difference in detected reads number in two pipelines between PE150 and PE100 (IDseq: $P = 0.62$, Kraken2: $P = 0.82$) (Table S2). Longer reads and paired-end sequencing mode were associated with higher mapped read counts, and a read length of 100 bp is

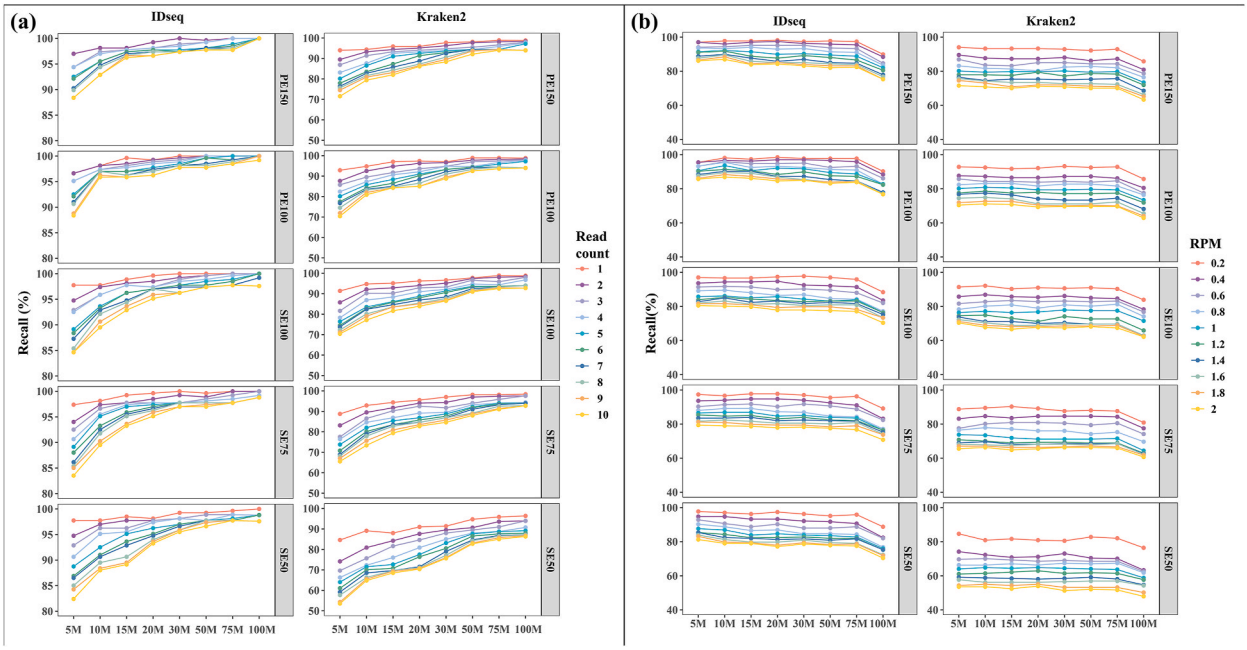

Fig. 3. Assessment of recall performance for pathogen detection using IDseq and Kraken2, across different dataset sizes and read lengths. The different colored lines represent the recall rates (%) for pathogen detection based on different positive criteria, with read counts ranging from 1 to 10 (a) and RPM ranging from 0.2 to 2 (b).

sufficient in pair-end sequencing. It is worth noting that the number of pathogen reads obtained using IDseq NR and Kraken2 in SE50 mode was extremely low, which could result in some pathogens not being identified due to the number of reads falling below the cutoff for detection (Fig. 2b). Consistent results were observed when bacteria, DNA viruses, fungi and *M. tuberculosis* complex (MTBCs) were analyzed separately (Figs. S1–S4).

Additionally, we compared the time required for data quality control (QC) and host data removal processes during data analysis across different data sizes and read lengths. The time cost ranged from over 3 h for 100 M PE150 to less than 5 min for 5 M SE50 (Fig. S5). As the fragment length increased and the data size became larger, the turnaround time for data analysis also increased.

3.3. The influence of multiple factors on assay recall

We assessed the recall performance of pathogen detection at each dataset size and each read length using both IDseq and Kraken2. The cutoff for detection was set as read count of 1–10, RPM of 0.2–2, respectively. Our results showed that both IDseq and Kraken2 had higher recall rates with lower cutoff for detection, and increased recall rates with larger dataset sizes and longer read lengths. However, IDseq outperformed Kraken2, particularly with smaller dataset sizes and read lengths (Fig. 3a–b). In nearly every dataset (95.5 % based on read count standard, 87.5 % based on RPM standard), there is a notable difference in the detection rates between the two methods (Fig. 4). At a dataset size of 100 M reads, the recall rate of IDseq was over 97.59 % under all read count cutoffs and read lengths. This indicates that IDseq is less affected by other factors if the dataset size is sufficient. In contrast, the recall rate of Kraken2 decreased even with larger dataset sizes when the cutoff for detection was stricter or the read length decreased. This underscores the importance of dataset size and bioinformatics pipeline selection for pathogen detection.

3.4. Comparison of SE75 20 M and SE50 50 M

The long-read sequencing results in long turnaround time and high sequencing costs, so we compared the two most commonly used read length, SE75 and SE50. Regardless of the cutoff for detection, the recall rate using IDseq requires 50 M reads in SE50 mode to achieve the same performance as 20 M SE75 strategies. This observation was in line with that of Kraken2, but the recall rate under SE75 20 M using Kraken2 was even higher than SE50 50 M when the cutoff for detection was low (<5). In addition, we found that the recall rate of SE50 using Kraken2 was much lower than that of other read lengths in each dataset, and this trend was also evident in

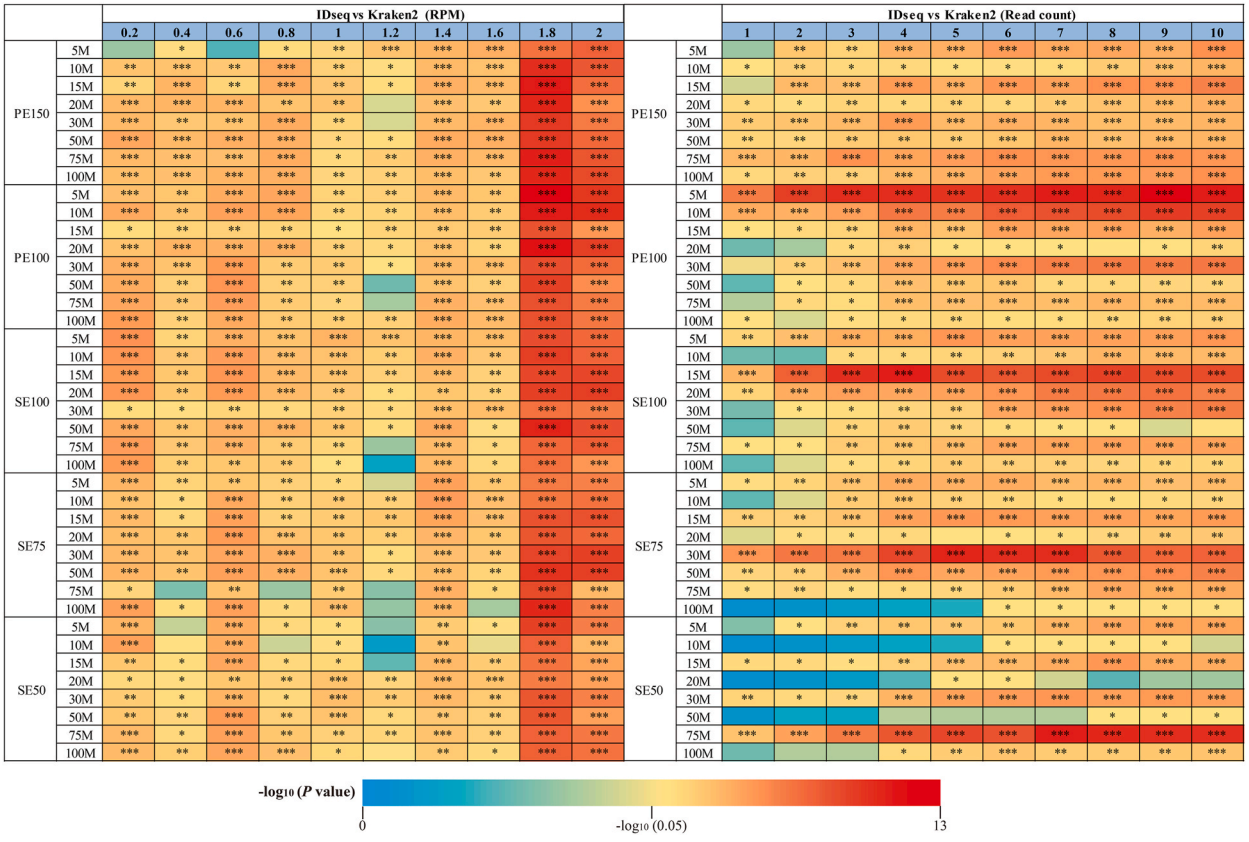

Fig. 4. Comparison of detection rates between different detection pipelines on the same dataset. Chi-square test was used to compare the recall rate of same dataset. "*", "**" and "***" stands for a *P* value less than 0.05, 0.01 and 0.001.

IDseq results with strict cutoffs for detection and small dataset sizes (Fig. 5a–b, Figs. S6–S7). Besides the similar recall rate for pathogen detection, the cost for 50 M SE50 on the same sequencing platform is 2.5 times that of 20 M SE75, and the turnaround time for data analysis is twice as long as 20 M SE75. Therefore, we proposed that the sequencing strategies of 20 M SE75 was superior to 50 M SE50.

(a)

	Pipeline	Read count	SE50 20M	SE50 30M	SE50 50M	SE50 75M	SE50 100M	SE50 20M	SE50 30M	SE50 50M	SE50 75M	SE50 100M
SE75 20 M	IDseq	1	2.2E-01	1.0E+00	1.0E+00	1.0E+00	1.9E-04					***
		2	7.5E-01	1.0E+00	1.0E+00	1.0E+00	1.0E-03					**
		3	1.0E+00	1.0E+00	5.0E-01	5.0E-01	5.7E-03					**
		4	1.0E+00	1.0E+00	1.0E+00	5.0E-01	5.7E-03					**
		5	6.2E-01	1.0E+00	1.0E+00	7.7E-01	1.2E-02					*
		6	3.7E-01	1.0E+00	7.9E-01	7.9E-01	2.2E-02					*
		7	3.9E-01	1.0E+00	6.0E-01	6.0E-01	1.2E-02					*
		8	3.1E-01	1.0E+00	4.5E-01	4.5E-01	2.1E-02					*
		9	3.3E-01	1.0E+00	4.7E-01	3.2E-01	3.6E-02					*
		10	4.6E-01	1.0E+00	5.1E-01	1.6E-01	8.8E-02					
	Kraken2	1	5.8E-02	8.0E-02	8.4E-01	1.0E+00	2.0E-02					*
		2	1.6E-02	8.3E-02	1.9E-01	1.0E+00	1.6E-02	*				*
		3	6.5E-03	1.1E-01	2.9E-01	6.4E-01	8.2E-02	**				
		4	1.1E-02	2.0E-01	7.9E-01	1.0E+00	1.6E-01	*				
		5	6.6E-03	2.8E-01	9.0E-01	6.0E-01	2.8E-01	**				
		6	8.0E-03	1.3E-01	9.0E-01	6.1E-01	2.9E-01	**				
		7	3.8E-04	1.2E-01	1.0E+00	5.4E-01	3.6E-01	***				
		8	2.6E-04	3.6E-02	1.0E+00	6.2E-01	3.6E-01	***	*			
		9	6.7E-04	4.1E-02	1.0E+00	5.5E-01	5.0E-01	***	*			
		10	1.1E-03	5.5E-02	1.0E+00	5.6E-01	5.8E-01	**				

(b)

	Pipeline	RPM	SE50 20M	SE50 30M	SE50 50M	SE50 75M	SE50 100M	SE50 20M	SE50 30M	SE50 50M	SE50 75M	SE50 100M
SE75 20 M	IDseq	0.2	1.0E+00	4.5E-01	1.6E-01	3.2E-01	7.2E-05					***
		0.4	5.8E-01	2.9E-01	2.3E-01	9.6E-02	1.3E-05					***
		0.6	1.0E+00	4.9E-01	4.9E-01	6.7E-01	8.5E-03					**
		0.8	1.0E+00	4.5E-01	4.5E-01	3.9E-01	2.3E-03					**
		1	1.0E+00	9.1E-01	8.1E-01	6.4E-01	1.7E-02					*
		1.2	9.1E-01	1.0E+00	8.2E-01	8.2E-01	3.3E-02					*
		1.4	9.1E-01	1.0E+00	8.2E-01	1.0E+00	7.3E-02					
		1.6	9.1E-01	1.0E+00	8.3E-01	7.5E-01	3.2E-02					*
		1.8	7.5E-01	1.0E+00	8.3E-01	9.2E-01	5.5E-02					
		2	9.2E-01	9.2E-01	1.0E+00	1.0E+00	6.0E-02					
	Kraken2	0.2	1.1E-02	8.0E-03	4.6E-02	2.7E-02	1.6E-04	*	**	*	*	***
		0.4	2.6E-04	1.5E-03	1.3E-04	8.6E-05	3.3E-08	***	**	***	***	***
		0.6	1.4E-03	2.0E-03	1.4E-03	1.4E-03	4.0E-06	**	**	**	**	***
		0.8	1.7E-02	3.5E-02	2.7E-02	3.5E-02	5.4E-04	*	*	*	*	***
		1	1.4E-01	1.2E-01	9.6E-02	7.9E-02	3.7E-03					**
		1.2	1.4E-01	6.9E-02	8.4E-02	6.9E-02	7.0E-03					**
		1.4	2.0E-02	2.5E-02	3.9E-02	2.0E-02	1.9E-03	*	*	*	*	**
		1.6	5.7E-03	7.4E-03	9.5E-03	9.5E-03	1.4E-03	**	**	**	**	**
		1.8	1.0E-02	2.7E-03	2.7E-03	2.7E-03	2.3E-04	*	**	**	**	***
		2	8.1E-03	1.2E-03	2.1E-03	1.6E-03	5.9E-05	**	**	**	**	***

Fig. 5. Comparison of detection rates difference between SE75 (20 M) and SE50 (20M–100 M).
Chi-square test was used to compare the difference between different datasets with same bioinformatic pipeline and detection standard. The left numbers show the p value and the right symbols stand for different significance. "*", "**" and "***" stands for a *P* value less than 0.05, 0.01 and 0.001.

Table 1
The causative pathogens identified in 43 clinical BALF samples.

Samples	Pathogens	Type	[a]Specific reads	Genome Size (Mb)	Reads/Genome size (Depth)	qPCR (Ct value)	Group
GZ2200228	*Klebsiella pneumoniae*	Bacteria	75,503	5.61	13,458.65	28	High
GZ2200228	*Stenotrophomonas maltophilia*	Bacteria	26,518	4.60	5764.78		High
GZ2200228	*Burkholderia cenocepacia*	Bacteria	16,260	7.96	2042.71		High
GZ2200228	*Pseudomonas aeruginosa*	Bacteria	524	6.62	79.15		Middle
GZ2200470	*Klebsiella pneumoniae*	Bacteria	17,645	5.61	3145.28	26	High
GZ2200470	*Simplexvirus humanalpha*1	Virus	21,002	0.15	140,013.33		High
GZ2202808	*Mycoplasma pneumoniae*	Bacteria	24	0.82	29.27		Middle
GZ2203190	*Haemophilus influenzae*	Bacteria	11	1.85	5.95		Low
GZ2203190	*Mycoplasma hominis*	Bacteria	16	0.70	22.86		Middle
GZ2203708	*Aspergillus fumigatus*	Fungi	11	28.54	0.39		Low
GZ2203708	*Enterococcus faecium*	Bacteria	2130	2.91	731.96		High
GZ2203708	*Pneumocystis jirovecii*	Fungi	63	8.18	7.70		Low
GZ2203708	*Ureaplasma urealyticum*	Bacteria	66	0.84	78.57		Middle
GZ2204613	*Pneumocystis jirovecii*	Fungi	11	8.18	1.34		Low
GZ2204613	*Ureaplasma urealyticum*	Bacteria	18	0.84	21.43		Middle
GZ2204613	*Enterococcus faecium*	Bacteria	698	2.91	239.86		Middle
HC2200029	*Simplexvirus humanalpha*1	Virus	610	0.15	4066.67		High
HC2200029	*Pneumocystis jirovecii*	Fungi	6	8.18	0.73		Low
HC2200029	*Klebsiella pneumoniae*	Bacteria	100	5.61	17.83	32	Low
HC2200211	*Streptococcus pneumoniae*	Bacteria	834	2.12	393.40		High
HC2200211	*Mycobacterium tuberculosis complex*	MTBC/ Bacteria	8	4.41	1.81		Low
HC2200211	*Staphylococcus aureus*	Bacteria	59	2.83	20.85	34	Middle
GZ2206727	*Mycobacterium tuberculosis complex*	MTBC/ Bacteria	5	4.41	1.13		Low
GZ2206727	*Pseudomonas aeruginosa*	Bacteria	134	6.62	20.24		Low
GZ2104496	*Staphylococcus aureus*	Bacteria	110	2.83	38.87	29	Middle
GZ2104496	*Klebsiella aerogenes*	Bacteria	495	5.27	93.93		Middle
GZ2105620	*Mycobacterium tuberculosis complex*	MTBC/ Bacteria	5	4.41	1.13		Low
GD11970	*Staphylococcus aureus*	Bacteria	111	2.83	39.22	31	Middle
GD11970	*Mycobacterium tuberculosis complex*	MTBC/ Bacteria	19	4.41	4.31		Low
GD11640	*Simplexvirus humanalpha*1	Virus	26,887	0.15	179,246.67		High
GD11640	*Pneumocystis jirovecii*	Fungi	6227	8.18	761.25		High
GD11640	*Klebsiella pneumoniae*	Bacteria	645	5.61	114.97	32	Middle
GD11640	*Candida tropicalis*	Fungi	1266	14.75	85.83		Middle
GD10886	*Acinetobacter baumannii*	Bacteria	3165	3.99	793.23	24	High
GD10886	*Chlamydia psittaci*	Bacteria	62	1.18	52.54		Middle
GD09444	*Acinetobacter baumannii*	Bacteria	1189	3.99	297.99	31	High
GD09444	*Aspergillus fumigatus*	Fungi	30	28.54	1.05		Low
GD09444	*Simplexvirus humanalpha*1	Virus	4425	0.15	29,500.00		High
GD08920	*Escherichia coli*	Bacteria	148	5.12	28.91		Middle
GD08920	*Simplexvirus humanalpha*1	Virus	32	0.15	213.33		Middle
DM-4192	*Mycobacteroides abscessus*	Bacteria	293	5.10	57.45		Middle
D2-0531	*Mycobacterium tuberculosis complex*	MTBC/ Bacteria	90	4.41	20.40		Low
D1-0530	*Cryptococcus neoformans*	Fungi	83	18.59	4.46		Low
D1-0527	*Moraxella catarrhalis*	Bacteria	5534	1.91	2897.38	22	High
D1-0527	*Staphylococcus aureus*	Bacteria	95	2.83	33.57	33	Middle
D1-0517	*Aspergillus terreus*	Fungi	42	30.03	1.40		Low
DM-4385	*Streptococcus pneumoniae*	Bacteria	55,504	2.12	26,181.13		High
DM-4385	*Staphylococcus aureus*	Bacteria	17,250	2.83	6095.41	29	High
DM-4385	*Candida albicans*	Fungi	6361	14.70	432.72		High
DM-4531	*Pneumocystis jirovecii*	Fungi	13,147	8.18	1607.21		High
DM-4531	*Talaromyces marneffei*	Fungi	606	28.31	21.41		Middle
DM-4395	*Acinetobacter baumannii*	Bacteria	4434	3.99	1111.28	25	High
DM-4395	*Klebsiella pneumoniae*	Bacteria	793	5.61	141.35	34	Middle
DM-4395	*Mycobacteroides abscessus*	Bacteria	25	5.10	4.90		Low
DM-4395	*Simplexvirus humanalpha*1	Virus	14,629	0.15	97,526.67		High
GZ2100214	*Aspergillus fumigatus*	Fungi	24	28.54	0.84		Low
GZ2100214	*Klebsiella pneumoniae*	Bacteria	116	5.61	20.68	36	Middle
GZ2100214	*Pseudomonas aeruginosa*	Bacteria	1189	6.62	179.61		Middle
GZ2100214	*Streptococcus mitis*	Bacteria	1044	1.97	529.95		High
GZ2100214	*Streptococcus pneumoniae*	Bacteria	162	2.12	76.42		Middle
GZ2100442	*Candida tropicalis*	Fungi	30	14.75	2.03		Low
GZ2100442	*Escherichia coli*	Bacteria	1730	5.12	337.89		High

(*continued on next page*)

Table 1 (*continued*)

Samples	Pathogens	Type	[a]Specific reads	Genome Size (Mb)	Reads/Genome size (Depth)	qPCR (Ct value)	Group
GZ2100627	*Fusobacterium nucleatum*	Bacteria	2265	2.31	980.52		High
GZ2100702	*Aspergillus fumigatus*	Fungi	6	28.54	0.21		Low
GZ2100775	*Klebsiella aerogenes*	Bacteria	11,139	5.27	2113.66		High
GZ2101042	*Aspergillus flavus*	Fungi	15	37.75	0.40		Low
GZ2101856	*Aspergillus flavus*	Fungi	1	37.75	0.03		Low
GZ2101856	*Cytomegalovirus humanbeta5*	Virus	46	0.23	200.00	33	Middle
GZ2101856	*Lymphocryptovirus humangamma4*	Virus	4	0.17	23.53	36	Middle
GZ2101856	*Pneumocystis jirovecii*	Fungi	7	8.18	0.86		Low
GZ2102120	*Acinetobacter baumannii*	Bacteria	4	3.99	1.00	38	Low
GZ2102120	*Orientia tsutsugamushi*	Bacteria	275	1.98	138.89		Middle
GZ2102518	*Acinetobacter baumannii*	Bacteria	1638	3.99	410.53	32	High
GZ2102518	*Klebsiella pneumoniae*	Bacteria	4805	5.61	856.51	33	High
GZ2103473	*Aspergillus fumigatus*	Fungi	30	28.54	1.05		Low
GZ2103993	*Pneumocystis jirovecii*	Fungi	1356	8.18	165.77		Middle
GZ2103993	*Ureaplasma urealyticum*	Bacteria	30	0.84	35.71		Middle
GZ2205173	*Aspergillus fumigatus*	Fungi	20	28.54	0.70		Low
GZ2205173	*Enterococcus faecium*	Bacteria	310	2.91	106.53		Middle
GZ2205173	*Stenotrophomonas maltophilia*	Bacteria	1370	4.60	297.83		High
GZ2205256	*Mycobacterium tuberculosis complex*	MTBC/ Bacteria	177	4.41	40.12		Middle
GZ2205648	*Mycobacterium tuberculosis complex*	MTBC/ Bacteria	6	4.41	1.36		Low
GZ2206201	*Klebsiella aerogenes*	Bacteria	9120	5.27	1730.55		High
GZ2206505	*Mycobacterium tuberculosis complex*	MTBC/ Bacteria	42	4.41	9.52		Low
GZ2206509	*Enterococcus faecium*	Bacteria	98	2.91	33.68		Middle
GZ2206801	*Aspergillus fumigatus*	Fungi	151	28.54	5.29		Low
GZ2206801	*Cytomegalovirus humanbeta5*	Virus	210	0.23	913.04	27	High
GZ2206801	*Lymphocryptovirus humangamma4*	Virus	1	0.17	5.88	Negative	Low
GZ2206962	*Klebsiella pneumoniae*	Bacteria	46,425	5.61	8275.40	23	High

[a] Note: "Specific reads" means the reads can only be mapped to a specific species.

3.5. Nucleic acid load affect pathogen detection

We also conducted analysis of recall for different types of pathogens, including bacteria, fungi, DNA viruses, and MTBCs (Figs. S8–S9). The results for bacteria and fungi were generally consistent with those of all pathogens. However, the small number of enrolled viruses and MTBCs resulted in some variation. The recall rate for DNA viruses was consistently high (almost all above 90 %) across all datasets, regardless of read length, dataset size, or the cutoff for detection. Upon checking the initial number of reads aligning to these 10 DNA viruses calculated from the raw data, we found that most of them (8/10) were higher than the strictest cutoff for detection (10 reads). For MTBC, half of the read numbers (4/8) calculated from the raw data were below 10, resulting in variable recall rates that depended on read length, dataset size and the cutoff for detection (Table 1). When using either Kraken2 or IDseq, the levels of reads number for each pathogen obtained from the raw data and from datasets with large size (>50 M) were basically consistent, suggesting that the number of reads could serve as an indicator of nucleic acid load in samples.

To further investigate the impact of nucleic acid load on pathogen detection, we categorized the pathogens into three groups based on the number of reads obtained from the raw data. Considering the influence of genome size on sequencing read counts, we employed a parameter of reads/genome-size (depth) to differentiate between high, medium, and low pathogen load data. The High group included pathogens with a depth above the third quartile, the Low group included those with a depth below the first quartile, and the Middle group included all others. We conducted correlation analysis between the cycle-threshold (Ct) values of qPCR and both the number of specific reads and depth separately. The results revealed a negative correlation between depth and Ct values, with an R^2 value of 0.6121, surpassing the R^2 value between Ct values and the number of reads, which was 0.5344 (Fig. S10). These results suggested that higher depth values indicate higher pathogen loads and the three groups of pathogens should have different loads in the original samples.

The pathogens in Middle and High groups were nearly all identified using IDseq, and the recall rate of those in Low group varied with dataset size and read length (Fig. 6). While the performance of Kraken was not as superior as that of IDseq, the results were similar, suggesting that selecting appropriate sequencing strategies and bioinformatics pipelines can improve the detection rate of pathogens with low nucleic acid load in samples.

4. Discussion

In this study, we collected and re-sequenced 43 BALF samples in which 89 causative pathogens had been previously confirmed.

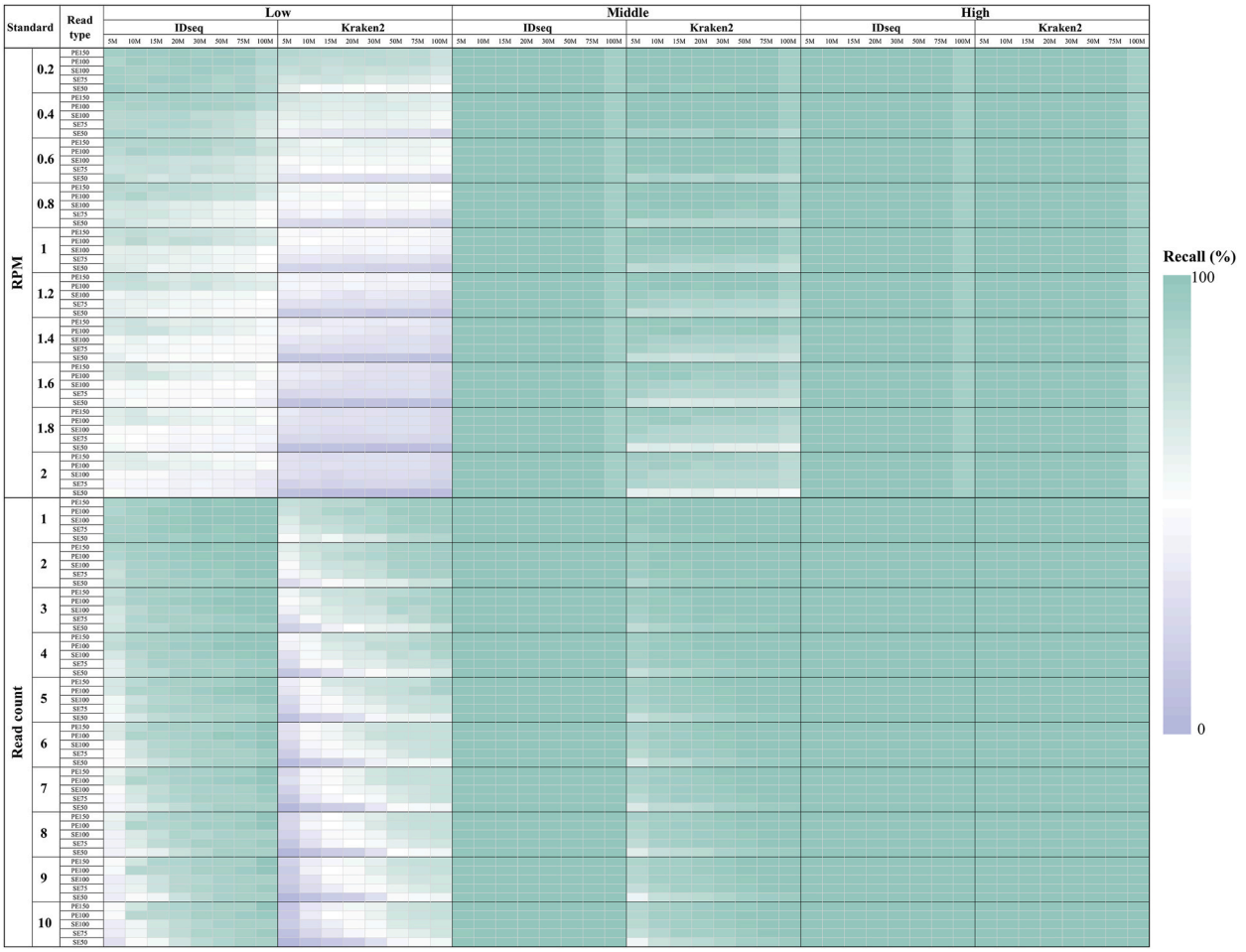

Fig. 6. The recall rate of pathogen with high, middle and low nucleic acid load in samples.
The Low and High group included pathogens with a read-number/genome-size below or above the first and third quartile respectively, while the Middle group encompassed all others. Samples with lower nucleic acid load have a lower recall rate in pathogen detection by mNGS. IDseq exhibits a better pathogen recall rate compared to Kraken2, especially in samples with low nucleic acid load.
Since data extraction does not affect the pathogen's RPM, under a consistent RPM detection standard, the differences in detection rates among varying data volumes with the same sequencing read length are minimal, aligning with expectations.

Based on the raw reads obtained from re-sequencing, datasets with different sizes and read lengths were simulated to investigate the effect of sequencing strategies on pathogen detection. The number of reads aligning to detected pathogens increased with dataset size and read length, and was influenced by bioinformatics pipelines. The recall rate of pathogens with low nucleic acid load in samples varied with dataset size and read length. In general, the sequencing strategy of 20 M SE75 was sufficient to achieve high performance. The recall rate under SE50 mode was significantly lower than other read lengths, and at least 50 M reads was required to achieve high detection rates. However, considering the cost and turnaround time for data analysis, we proposed that the sequencing strategies of 20 M SE75 were superior to 50 M SE50.

The performance of mNGS in pathogen detection can be influenced by various factors in both wet and dry labs, including sample processing, nucleic acid extraction, library construction, sequencing platforms and bioinformatics workflows [34,35]. By using simulation data based on sequencing reads, we were able to eliminate the effects of experimental processes, sequencing platforms, and focus solely on the impact of sequencing strategies and bioinformatics pipelines on pathogen detection. Our study found that increasing dataset sizes and read lengths improved assay performance, and that 20 M reads of data is sufficient for SE75 mode to achieve a high recall rate, consistent with previous research [34]. In contrast, the shorter sequences in SE50 may result in loss of critical sequence information, which presents a challenge that cannot be fully addressed by increasing dataset size.

Previous studies have shown that there exists a correlation between the microbial load present in samples and the quantitative reads per million reads (RPM) values identified by mNGS [35,36]. In the current study, using both IDseq and Kraken2, we obtained almost identical levels of read counts for the same pathogen from both the raw data and large-sized datasets. This leads us to believe that the reads number of pathogen can serve as an indicator of the nucleic acid load of the corresponding pathogen present in the sample. In this study, we conducted qPCR experiments to validate the correlations between read numbers and pathogen load. The

results of the qPCR also indicate that samples with higher pathogen loads yield higher quantities of specific reads. Hence, when the nucleic acid load of pathogens is unknown, the read count of pathogens identified through mNGS can be utilized to assess the nucleic acid load for analysis.

By classifying pathogens into high, medium, and low nucleic acid load groups and analyzing their recall rates, our findings suggest that high nucleic acid loads in the sample can result in more stable pathogen detection efficiency that is less affected by sequencing strategies. This highlights the importance of collecting samples with adequate pathogen nucleic acid loads, such as by directly sampling at the site of infection during the initial acute presentation and before the administration of antibiotics, and ensuring proper storage during transport and storage [11]. On the other hand, when nucleic acid loads are insufficient, detection efficiency can be improved by increasing sequencing data and choosing an appropriate bioinformatics analysis process.

This study also has some limitations. Firstly, that most patients received empirical antibiotic treatment before clinical sampling prevented the complete identification of all microorganisms present in the samples, thus limiting the assessment of pathogen detection precision. Secondly, due to the limited availability of clinical samples and storage conditions, RNA viruses were not included in the study. Furthermore, the inclusion of a limited number of pathogens may have introduced some bias into the findings. While qPCR can offer more reliable quantitative results, only a small subset of pathogens was tested using qPCR in a limited number of samples in this study. Besides, spiked-in controls were not included in the experimental process, and the potential impact of DNA extraction and library construction on final detection was not taken into account.

The bioinformatics analysis of mNGS is a crucial step in pathogen detection, which involves several key processes such as sequence quality control, sequence alignment, species classification, and data visualization. Nevertheless, benchmarking mNGS tools remains a challenge in this field, as the choice of parameters, databases, and datasets can all affect the outcomes. Currently, some commercial or open resources have been developed for researchers' convenience [18,22,37–39]. Our study focused on IDseq and Kraken2 for performance testing and confirmed that the bioinformatics pipelines significantly impact pathogen detection. Therefore, developing more efficient and accurate sample pretreatment and analysis processes to enhance mNGS tool performance should remain a key priority for researchers.

Ethics approval

This study was reviewed and approved by the Clinical Research Ethics Committee of The Second Xiangya Hospital of Central South University, with the approval number: LYF2022229.

Informed consent

All participants/patients (or their proxies/legal guardians) provided informed consent to participate in the study.

Funding

This study was supported by National Natural Science Foundation of China (No. 82102499; ZL), Hunan Natural ScienceFoundation (No. 2021JJ40840; ZL), the Scientific Research Launch Project for new employees of the Second Xiangya Hospital of Central South University.

Data availability statement

Sequencing data were deposited at the National Genomics Data Center under accession numbers PRJCA019784. We declare that the main data supporting the findings are available within this article and the supplemental material. The other data generated and analyzed for this study are available from the corresponding author upon reasonable request.

CRediT authorship contribution statement

Ziyang Li: Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Zhe Guo:** Writing – original draft, Methodology, Investigation. **Weimin Wu:** Writing – original draft, Visualization, Formal analysis. **Li Tan:** Writing – original draft, Methodology, Formal analysis. **Qichen Long:** Writing – original draft, Data curation. **Han Xia:** Writing – original draft, Project administration, Data curation, Conceptualization. **Min Hu:** Writing – review & editing, Validation, Supervision, Conceptualization.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e33429.

References

[1] 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 (2018) 778–788.
[2] C.Y. Chiu, S.A. Miller, Clinical metagenomics, Nat. Rev. Genet. 20 (2019) 341–355.
[3] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, et al., A novel coronavirus from patients with pneumonia in China, 2019, N. Engl. J. Med. 382 (2020) 727–733.
[4] S.K. Gire, A. Goba, K.G. Andersen, R.S. Sealfon, D.J. Park, L. Kanneh, et al., Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak, Science 345 (2014) 1369–1372.
[5] Y. Zhang, P. Cui, H.C. Zhang, H.L. Wu, M.Z. Ye, Y.M. Zhu, et al., Clinical application and evaluation of metagenomic next-generation sequencing in suspected adult central nervous system infection, J. Transl. Med. 18 (2020) 199.
[6] L. Chen, Y. Zhao, J. Wei, W. Huang, Y. Ma, X. Yang, et al., Metagenomic next-generation sequencing for the diagnosis of neonatal infectious diseases, Microbiol. Spectr. 10 (2022) e0119522.
[7] D. Han, Z. Li, R. Li, P. Tan, R. Zhang, J. Li, mNGS in clinical microbiology laboratories: on the road to maturity, Crit. Rev. Microbiol. 45 (2019) 668–685.
[8] P.S. Ramachandran, M.R. Wilson, Metagenomics for neurological infections - expanding our imagination, Nat. Rev. Neurol. 16 (2020) 547–556.
[9] F.X. Lopez-Labrador, J.R. Brown, N. Fischer, H. Harvala, S. Van Boheemen, O. Cinek, et al., Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: wet lab procedure, J. Clin. Virol. 134 (2021) 104691.
[10] J.J.C. de Vries, J.R. Brown, N. Couto, M. Beer, P. Le Mercier, I. Sidorov, et al., Recommendations for the introduction of metagenomic next-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting, J. Clin. Virol. 138 (2021) 104812.
[11] S. Miller, C. Chiu, The role of metagenomics and next-generation sequencing in infectious disease diagnosis, Clin. Chem. 68 (2021) 115–124.
[12] L. Schuele, H. Cassidy, N. Peker, J.W.A. Rossen, N. Couto, Future potential of metagenomics in microbiology laboratories, Expert Rev. Mol. Diagn 21 (2021) 1273–1285.
[13] 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 (2021) 2000792.
[14] S. Miller, S.N. Naccache, E. Samayoa, K. Messacar, S. Arevalo, S. Federman, et al., Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid, Genome Res. 29 (2019) 831–842.
[15] D. Paez-Espino, G.A. Pavlopoulos, N.N. Ivanova, N.C. Kyrpides, Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data, Nat. Protoc. 12 (2017) 1673–1682.
[16] S. Nooij, D. Schmitz, H. Vennema, A. Kroneman, M.P.G. Koopmans, Overview of virus metagenomic classification methods and their biological applications, Front. Microbiol. 9 (2018) 749.
[17] S.H. Ye, K.J. Siddle, D.J. Park, P.C. Sabeti, Benchmarking metagenomics tools for taxonomic classification, Cell 178 (2019) 779–794.
[18] S.N. Naccache, S. Federman, N. Veeraraghavan, M. Zaharia, D. Lee, E. Samayoa, et al., A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples, Genome Res. 24 (2014) 1180–1192.
[19] J.J.C. de Vries, J.R. Brown, N. Fischer, I.A. Sidorov, S. Morfopoulou, J. Huang, et al., Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples, J. Clin. Virol. 141 (2021) 104908.
[20] D.C. Gaston, H.B. Miller, J.A. Fissel, E. Jacobs, E. Gough, J. Wu, et al., Evaluation of metagenomic and targeted next-generation sequencing workflows for detection of respiratory pathogens from bronchoalveolar lavage fluid specimens, J. Clin. Microbiol. 60 (2022) e0052622.
[21] D.E. Wood, S.L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments, Genome Biol. 15 (2014) R46.
[22] K.L. Kalantar, T. Carvalho, C.F.A. de Bourcy, B. Dimitrov, G. Dingle, R. Egger, et al., IDseq-An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring, GigaScience 9 (2020).
[23] H. Chen, Y. Yin, H. Gao, Y. Guo, Z. Dong, X. Wang, 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 (2020) S416–S426.
[24] Z. Diao, Y. Zhang, Y. Chen, Y. Han, L. Chang, Y. Ma, et al., Assessing the quality of metagenomic next-generation sequencing for pathogen detection in lower respiratory infections, Clin. Chem. 69 (2023) 1038–1049.
[25] S. Chen, C. He, Y. Li, Z. Li, C.E. Melancon, A computational toolset for rapid identification of SARS-CoV-2, other viruses and microorganisms from sequencing data, Briefings Bioinf. 22 (2021) 924–935.
[26] E. Crawford, J. Kamm, S. Miller, L.M. Li, S. Caldera, A. Lyden, et al., Investigating transfusion-related sepsis using culture-independent metagenomic sequencing, Clin. Infect. Dis. 71 (2020) 1179–1185.
[27] S. Saha, A. Ramesh, K. Kalantar, R. Malaker, M. Hasanuzzaman, L.M. Khan, et al., Unbiased metagenomic sequencing for pediatric meningitis in Bangladesh reveals neuroinvasive chikungunya virus outbreak and other unrealized pathogens, mBio 10 (2019).
[28] E. Mick, J. Kamm, A.O. Pisco, K. Ratnasiri, J.M. Babik, G. Castaneda, et al., Upper airway gene expression reveals suppressed immune responses to SARS-CoV-2 compared with other respiratory viruses, Nat. Commun. 11 (2020) 5854.
[29] H. Duan, X. Li, A. Mei, P. Li, Y. Liu, X. Li, et al., The diagnostic value of metagenomic next rectanglegeneration sequencing in infectious diseases, BMC Infect. Dis. 21 (2021) 62.
[30] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (2018) i884–i890.
[31] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–359.
[32] D.E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2, Genome Biol. 20 (2019) 257.
[33] J.G. Chappell, T. Byaruhanga, T. Tsoleridis, J.K. Ball, C.P. McClure, Identification of infectious agents in high-throughput sequencing data sets is easily achievable using free, cloud-based bioinformatics platforms, J. Clin. Microbiol. 57 (2019).
[34] D. Liu, H. Zhou, T. Xu, Q. Yang, X. Mo, D. Shi, et al., Multicenter assessment of shotgun metagenomics for pathogen detection, EBioMedicine 74 (2021) 103649.
[35] D. Han, Z. Diao, H. Lai, Y. Han, J. Xie, R. Zhang, et al., Multilaboratory assessment of metagenomic next-generation sequencing for unbiased microbe detection, J. Adv. Res. 38 (2022) 213–222.
[36] A. Babiker, H.L. Bradley, V.D. Stittleburg, J.M. Ingersoll, A. Key, C.S. Kraft, et al., Metagenomic sequencing to detect respiratory viruses in persons under investigation for COVID-19, J. Clin. Microbiol. 59 (2020).
[37] M. Alawi, L. Burkhardt, D. Indenbirken, K. Reumann, M. Christopeit, N. Kroger, et al., DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples, Sci. Rep. 9 (2019) 16841.
[38] D. Kim, L. Song, F.P. Breitwieser, S.L. Salzberg, Centrifuge: rapid and sensitive classification of metagenomic sequences, Genome Res. 26 (2016) 1721–1729.
[39] M. Vilsker, Y. Moosa, S. Nooij, V. Fonseca, Y. Ghysens, K. Dumon, et al., Genome Detective: an automated system for virus identification from high-throughput sequencing data, Bioinformatics 35 (2019) 871–873.