Optimized Sequencing Adaptors Enable Rapid and Real-Time Metagenomic Identification of Pathogens during Runtime of Sequencing

Dong Zhang,[a,†] Jingjia Zhang,[a,†] Juan Du,[a,†] Yiwen Zhou,[b,†] Pengfei Wu,[b] Zidan Liu,[b] Zhunzhun Sun,[b] Jianghao Wang,[b] Wenchao Ding,[b] Junjie Chen,[b] Jun Wang,[b] Yingchun Xu,[a] Chuan Ouyang ,[b,*] and Qiwen Yang [a,*]

BACKGROUND: Metagenomic next-generation sequencing (mNGS) offers the promise of unbiased detection of emerging pathogens. However, in indexed sequencing, the sequential paradigm of data acquisition, demultiplexing, and analysis restrain read assignment in advance and real-time analysis, resulting in lengthy turnaround time for clinical metagenomic detection.

METHODS: We described the utility of internal-index adaptors with different lengths of barcode in multiplex sequencing. The base composition for each position within these adaptors was well-balanced to ensure nucleotide diversity and optimal sequencing performance and to achieve the early assignment of reads by first sequencing the barcodes. Combined with an automated library preparation device, we delivered a rapid and real-time bioinformatics pathogen identification solution for the Illumina NextSeq platform. The diagnostic performance was evaluated by testing 153 lower respiratory tract specimens using mNGS in comparison to culture, 16S/internal transcribed spacer amplicon sequencing, and additional PCR-based tests.

RESULTS: By calculating the average F1 scores of all read lengths under different threshold values, we established the optimal threshold for pathogens identification, and found that 36 bp was the optimal shortest read length for rapid mNGS analysis. Rapid detection had a negative percentage agreement and positive percentage agreement of 100% and 85.1% for bacteria and 97.4% and 80.3% for fungi, when compared to a composite standard. The rapid mNGS solution enabled accurate pathogen identification in about 9.1 to 10.1 h sample-to-answer turnaround time.

CONCLUSIONS: Optimized internal index adaptors combined with a real-time analysis pipeline provide a potential tool for a first-line test in critically ill patients.

Introduction

Metagenomic next-generation sequencing (mNGS) has the potential to detect emerging pathogens without bias and to facilitate their characterization without a priori knowledge of their genomic sequences. Although mNGS is one of the few state-of-the-art methods available at the earliest stages of an epidemic, 1 of the fundamental drawbacks of NGS-based pathogen identification approaches is that its turnaround time (TAT) is longer than that of other targeted molecular methods. However, microbial culture, the basic standard for etiological agent identification, is also time-consuming. Other rapid molecular detection approaches such as PCR are hypothesis driven and thus require a priori suspicion of the causative pathogen(s). Real-time Oxford nanopore sequencing technology provides an alternative metagenomic approach to identify pathogens with an unprecedented TAT of <6 h (1–5). Albeit rather fast, nanopore sequencing and analysis has its drawbacks, including low sequencing depth/coverage and reads output, high-cost of sequencing, and higher error rates compared to Illumina sequencing. Lack of accurate and rapid detection methods for the majority of potential pathogens poses considerable challenges for diagnosis of infectious disease, leading to increased mortality and morbidity and high risk of antimicrobial drug resistance due to empiric antimicrobial treatment (6).

[a]Department of Clinical Laboratory, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China; [b]Hangzhou Matridx Biotechnology Co., Ltd., Hangzhou, Zhejiang, China.
*Address correspondence to: Q.Y. at Peking Union Medical College Hospital, No. 1 Shuaifuyuan Wangfujing, Dongcheng District, Beijing 100730, China. E-mail yangqiwen81@vip.163.com. C.O. at Hangzhou Matridx Biotechnology Co., Ltd., Bd 2-3, 2073 Jinchang Rd, Liangzhu St, Yuhang District, Hangzhou, Zhejiang 310000, China. E-mail ouyangchuan@matridx.com.
†These authors contributed equally.
Received August 17, 2021; accepted December 28, 2021.
https://doi.org/10.1093/clinchem/hvac024

The sequential paradigm of Illumina data acquisition and analysis is 1 of the main bottlenecks leading to high TAT. Consequently, several bioinformatics tools or pipelines have been developed to speed up sequencing data acquisition and analysis (7–10). However, live-mapping approaches have limitations when dealing with multiplexed sequencing data, because such approaches are unlikely to distinguish between different samples before the indexes have been sequenced. The position of each cluster in a tile is defined in the template generation step during cycles 1 to 7 of Read 1 for Illumina real-time analysis software. Template generation is a critical step because it serves as a reference for registration and intensity extraction in subsequent sequencing cycles. Furthermore, during cycles 1 to 25 of Read 1, a real-time analysis filter removes unreliable clusters from the image extraction results (11). High percentage values for cluster passing filter (%PF) are 1 of the key factors in high yield and sequencing quality. A balanced and random base composition in each cycle of the first couple of Read 1 sequencing cycles guarantees the accurate calculation of the position of each cluster and normal cluster %PF values (12). Therefore, even though sequencing the barcodes first before Read 1 would accomplish demultiplexing first and improve analysis speed, it would negatively influence the initial template generation and cluster %PF values because the barcodes do not offer a wide variety of nucleotide sequences, especially in case of a small sample size (12). To overcome this obstacle and provide methods to accelerate mNGS analysis in clinical application, we hereby describe a new sequencing adaptor design strategy as well as a simple, rapid, and real-time metagenomic pathogen identification method based on the Illumina NextSeq sequencing platform.

Materials and Methods

BIOINFORMATICS ANALYSIS

Internal indexing was used to demultiplex the sequencing data at the 22nd cycle to obtain a mapping of the reads and the samples stored in memory as a hash table. The first 9 nucleotides containing the internal index were trimmed in follow-up analysis even if the internal index were 6 bp or 7 bp long. Taxonomic classification began at the 37th cycle, where raw reads were aligned to a human-specific database constructed from *Homo sapiens* sequences in the National Center for Biotechnology Information nucleotide database using Kraken2 (version 2.1.2) (13) and Bowtie2 (version 2.3.5.1) (14) in sequence. Subsequent real-time analysis was automatically triggered via sequencing and analysis progress (every 2 sequencing cycles). Nonhuman reads were processed via adaptor filtering, where those containing more than 80% of the adaptor sequence were discarded. Kraken2 was then used to rapidly classify the remaining reads in the nucleotide database. Bowtie2 and a microbial database based on RefSeq were then used for candidate pathogen identification. Finally, BLAST (version 2.9.0+) (15) alignment to the nucleotide database was conducted to validate candidate reads, where Kraken2 and Bowtie2 were inconsistent.

Further information on research materials and methods is available in the online Supplemental Methods.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The human research ethics committee of the Institutional Review Board of the Peking Union Medical College Hospital approved this study. This project did not affect the normal diagnosis and treatment of patients. Formal ethical approval was reviewed and waived, and written patient consent was not required after consultation with the Institutional Review Board (ethics approval number S-K1186).

DATA AVAILABILITY

Metagenomic sequencing data (FASTQ files) after removing human genomic reads have been deposited in the National Center for Biotechnology Information Sequence Read Archive as BioProject PRJNA742139.

Results

ADAPTOR DESIGN

Single-end sequencing is sufficient for k-mer based taxonomic classification and pathogen identification. Generally, using the traditional Illumina TruSeq adaptor, the first read for DNA insert is sequenced for 25 to 150 base pairs (bp), followed by barcoding reads for Index 1 and Index 2. A balanced and random base composition of DNA inserts ensures normal cluster %PF and sequencing yield (Fig. 1, A, upper panel). Sequencing adaptors can be designed to accommodate 8 bp internal indexes downstream from the Read 1/Read 2 sequencing primer site (Fig. 1, A, middle panel). Demultiplexing by internal index can be achieved using these adaptors. However, the approach is flawed by unbalanced base composition of combinations of limited barcodes, which lead to low cluster %PF value and low yield, especially when the sequencing result of the ninth cycle is undoubtedly T nucleotide for almost every single molecule of DNA libraries. Here we present a series of refined internal index adaptors, featured by the combination of 3 types of barcode with length of 6, 7, and 8 bp, respectively. Additionally, the base composition for each position within these adaptors was adjusted and well-balanced. Our pilot test using no less than 6 of

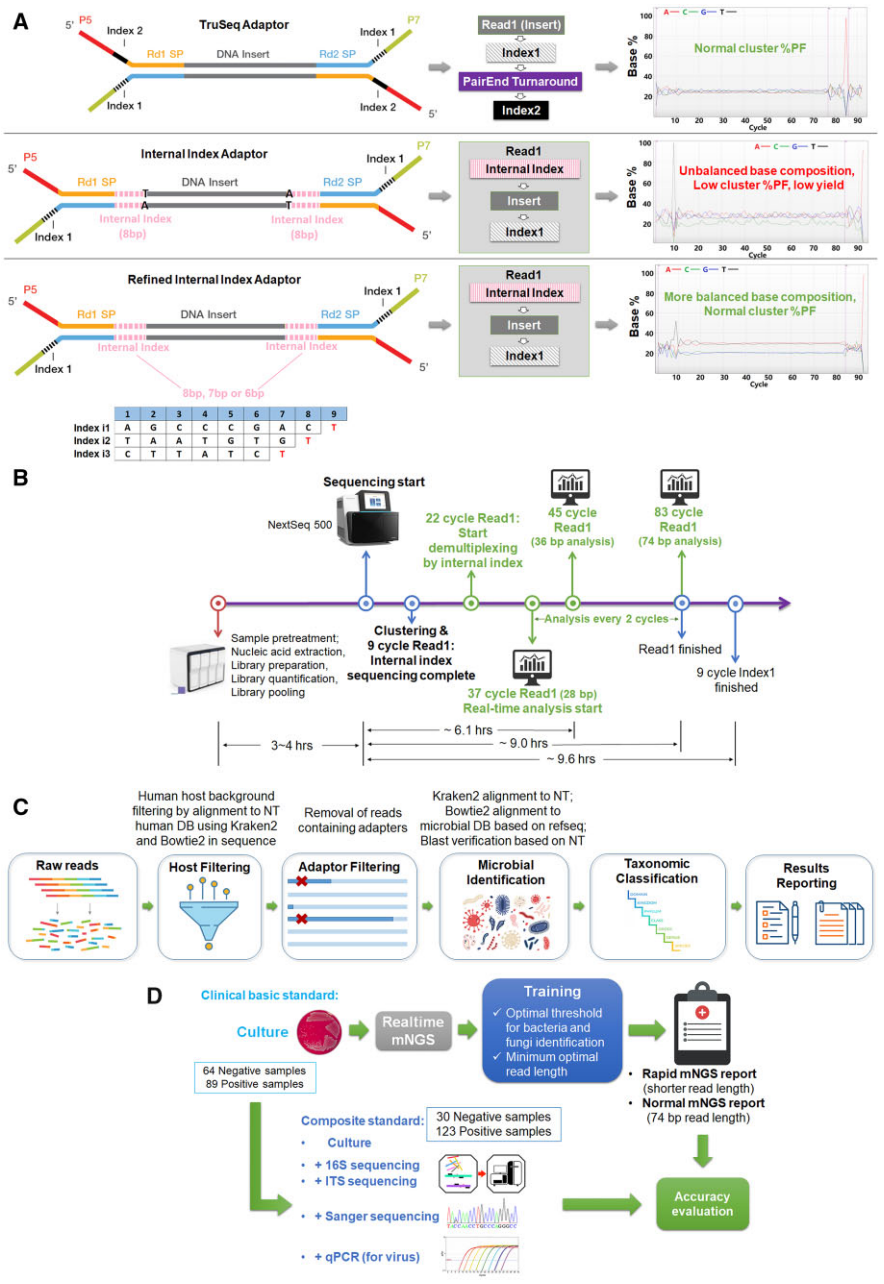

Fig. 1. Internal index and study workflow. (A), Structures of conventional TruSeq adaptor (upper panel), internal index adaptor (middle panel), and refined internal index adaptor (lower panel). The sequencing paradigm for each type of adaptor and %Base by cycle plot from Illumina sequencing analysis viewer are also shown; (B), Schematic and timeline of real-time mNGS workflow. The approximate turnaround time for a rapid real-time mNGS pipeline (45 cycle Read 1) was 9.1 to 10.1 h (3–4 h of wet-lab work and 6.1 h of Illumina sequencing and analysis); (C), mNGS analysis of the pipeline. Microbial reads were aligned to the reference database using Kraken2, Bowtie2, and BLAST in sequence after human host background filtering and adaptor removal, followed by taxonomic classification and report generation; (D), Schematic overview of accuracy evaluation workflow. Culture results were used as the clinical basic standard. The composite standard included additional results from 16S sequencing, internal transcribed spacer sequencing, PCR + Sanger sequencing, and virus qPCR verification. Abbreviations: Rd1 SP, Read 1 sequencing primer; Rd2 SP, Read 2 sequencing primer; %PF: percent cluster passing filter. Color figure available online at clinchem.org.

these optimized adaptors demonstrated that more-balanced internal barcode sequences enabled normal cluster %PF and high-quality data generation (Fig. 1, A, lower panel; Supplemental Table 1).

MNGS DETECTION AND ANALYSIS WORKFLOW

We developed a rapid mNGS test workflow consisting of automated library preparation, sequencing on Illumina NextSeq, and real-time bioinformatics analysis (Fig. 1, B; Supplemental Methods). Generally, it took about 3 to 3.5 h to prepare cell-free DNA libraries for a variety of body fluids, while preparation time for genomic DNA libraries was longer, at 3 to 4 h. We took advantage of a time-saving single-end run (83 cycles for Read 1 and 9 cycles for Index 1) on Illumina NextSeq 500/550 system, which took 9.6 h for sequencing. The mNGS analysis strategy was as follows: clustering and internal index sequencing were completed at the 9th cycle; sequencing data were demultiplexed by internal index at the 22nd cycle; real-time analysis started at the 37th cycle (read length was 28 bp) and was repeated every 2 cycles until the 83rd cycle (final read length was 74 bp); sequencing was completed after another 9 cycles of sequencing Index 1. Previous studies have demonstrated that a dual-index demultiplex strategy (internal index + Index 1), which is optional, eliminates index switching (index crosstalk) and increases multiplex sequencing accuracy (16–18). We also developed an inhouse mNGS pipeline that can run in parallel to ensure real-time analysis for each round (Fig. 1, C, Supplemental Methods and Movie).

SAMPLE COLLECTION, PATIENT DEMOGRAPHICS, AND CLINICAL STUDY DESIGN

To evaluate the accuracy of the previously mentioned mNGS detection and analysis approach, a total of 153 lower respiratory tract specimens from different patients in the hospital, including 139 sputum, 11 tracheobronchial aspirate fluid, and 3 bronchoalveolar lavage fluid were collected as residual samples after routine clinical testing in the microbiology laboratory (Table 1; Supplemental Table 2). A blinded study was carried out as follows: 89 culture positive specimens [with pathogen(s) identified to genus or species level] and 64 culture-negative specimens were sequenced and analyzed by real-time mNGS test (Fig. 1, D); each run was performed on Illumina NextSeq 500 system with 17 to 22 libraries (8 runs in total), including 1 external positive control and 1 negative control (Supplemental Methods, Supplemental Table 3). Besides, a constant amount of internal control DNA was incorporated into each specimen to monitor the level of host DNA background and microbial abundance. The internal control DNA had an average reads count per million total reads (RPM) of 69.8 (median 20.94, range 0.07–1052.9) (Supplemental Fig. 1, A). The optimal thresholds for bacteria and fungi identification, as well as the minimum optimal read length of our mNGS test were calculated compared to culture results, which were considered the clinical basic standard. Rapid mNGS report and normal mNGS report for each sample were generated at the minimum optimal read length and at 74 bp, respectively (Fig. 1, D).

Besides comparison between culture and mNGS analysis, a strategic multilevel evaluation was performed using a composite standard in which culture results were combined with additional results from various sources including (i) 16S/internal transcribed spacer sequencing for bacteria/fungi identification in all samples, (ii) PCR + Sanger sequencing for verification of some discordant cases, and (iii) confirmatory research-based real-time quantitative PCR (qPCR) for viruses detection in all specimens (Fig. 1, D). The traditional test and mNGS results of all cases are shown in Supplemental Table 4.

SEQUENCING PERFORMANCE OF THE INTERNAL INDEX ADAPTORS

To compare the internal index adaptors to traditional Illumina adapters, we also used TruSeq adaptors and the remaining DNA samples extracted from 153 lower respiratory tract specimens to prepare and sequence the libraries. The base% by cycle for each type of adaptor is shown in Fig. 2, A. Even though the sequencing result of the ninth cycle was T nucleotide dominantly (>40%) for internal index adaptor libraries, well-balanced base composition ensured normal cluster passing filter ratio (Fig. 2, B) and Q30 score (Fig. 2, C), compared to TruSeq adaptor libraries. The average read depths were 18.3M for libraries demultiplexed by internal index only and 17.5M for libraries demultiplexed by internal index and index1, compared to 16.2M for TruSeq adaptor libraries demultiplexed by index1 and index2 (Fig. 2, D).

The total sequencing time for internal index libraries in 8 runs was 540 min on average. Real-time analysis began at 332 min after the sequencing run started (at the 37th cycle, target read length was 28 bp) and took an average of 12.1 min per round (Fig. 2, E, left panel; Supplemental Fig. 1, B). Therefore, for testing lower respiratory tract specimens, the theoretical TAT of the real-time mNGS method might be reduced to any time between 9.7 h (37-cycle Read 1) and 13.2 h (83-cycle Read 1). However, the elapsed time for TruSeq adaptor libraries was 700 min on average (Fig. 2, E, right panel). Besides, the average delay from receiving a sample in the laboratory to getting a positive culture result was much longer at about 81.1 h (median 51, range 18–245) (Supplemental Table 2).

Table 1. Patient and sample characteristics.

Characteristics	Value
Patient demographics (n = 153)	
Age, years, mean (range)	60.4 (14–94)
Sex, male, n (%)	98 (64.1)
Hospitalization, n (%)	
ICU[a]	61 (39.9)
Non-ICU	92 (60.1)
Days hospitalized, mean (range)	35.5 (1–473)
Antibiotic use	139 (90.8)
Principal diagnosis, n (%)	
Infectious diseases	90 (58.8)
Pulmonary infection	68 (44.4)
Severe pneumonia	14 (9.2)
Lung abscess	2 (1.3)
Others	6 (3.9)
Noninfectious diseases	63 (41.2)
Cancer	8 (5.2)
Pulmonary emphysema	5 (3.3)
Pleural effusion	4 (2.6)
Pulmonary hypertension	4 (2.6)
Pulmonary nodule	4 (2.6)
Valvular heart disease	4 (2.6)
Others	34 (22.2)
Laboratory findings of patient	
WBC[b] count of blood, mean (range) ($\times 10^9$/L)	10.8 (1.15–38.74)
Percentage of lymphocytes, mean (range) (%)	12.6 (0.8–36.0)
Percentage of neutrophils, % mean (range)	79.0 (7.2–90.5)
Sample information (n = 153)	
Sample type, n (%)	
Sputum	139 (90.8)
Tracheobronchial aspirate fluid	11 (7.2)
Bronchoalveolar lavage fluid	3 (2.0)
Organism cultured, n	
Candida albicans	25
Pseudomonas aeruginosa	23
Acinetobacter baumannii	18
Klebsiella pneumoniae	11
Staphylococcus aureus	9
Others	27
Negative	64

[a]Intensive care unit.
[b]White blood cell.

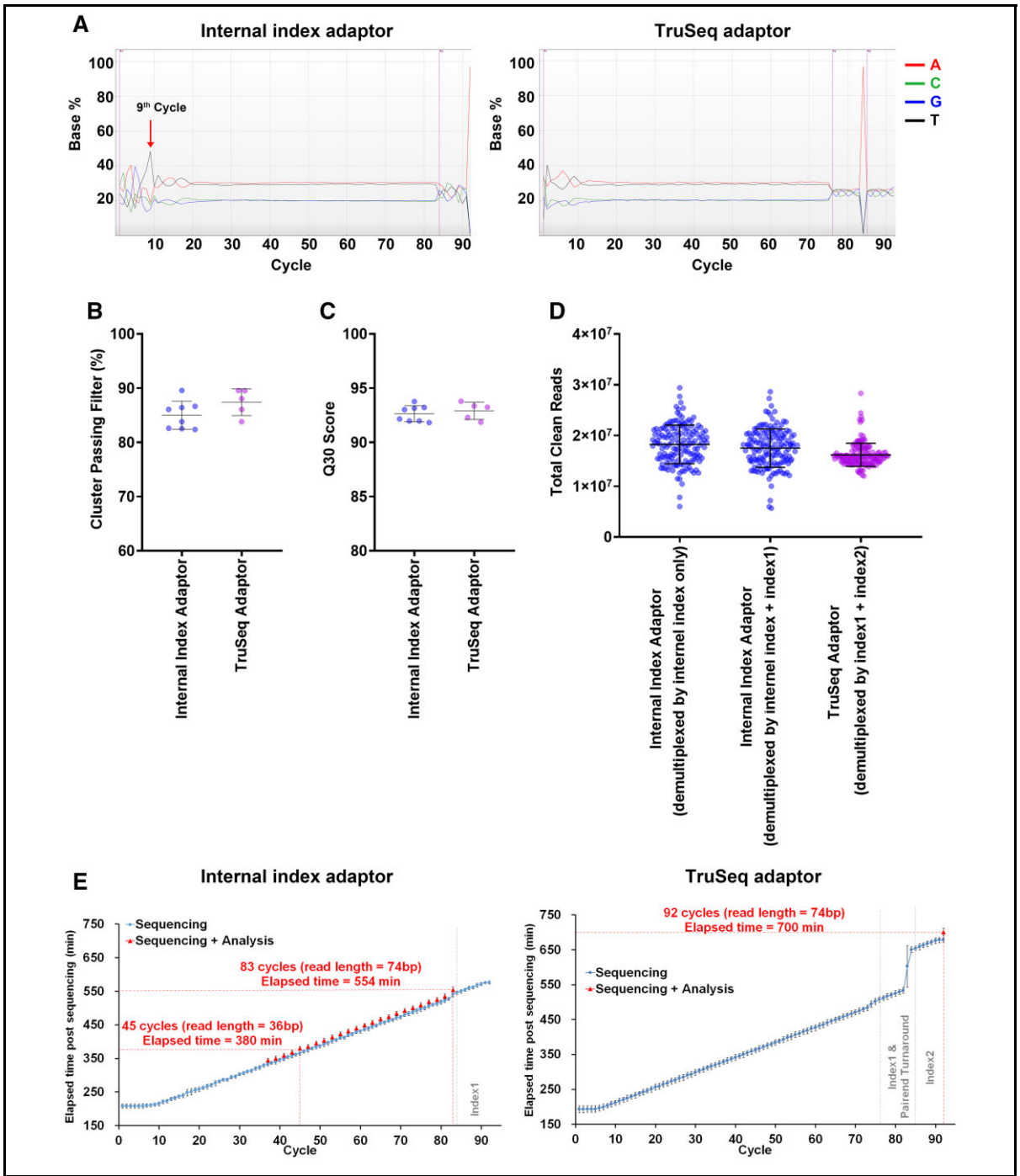

Fig. 2. Sequencing performance of the internal index adaptors. (A), %Base by cycle plot for internal index adaptors (left panel) and Illumina TruSeq adaptors (right panel). Only 1 run statistics for each adaptor type was shown as there was no obvious difference between each sequencing run; (B–D), Scatter dot plot showing the cluster passing filter ratio (B), Q30 score (C), and the count of total clean reads for 153 libraries (D). The solid line in the center represents the mean, and error bars represent the SD; (E) Elapsed time of sequencing and analysis for internal index adaptors (left panel) and Illumina TruSeq adaptors (right panel). Error bars represent SD. Color figure available online at clinchem.org.

OPTIMAL THRESHOLD FOR PATHOGEN IDENTIFICATION

Shorter reads are more likely to map to microorganisms with high sequence similarity, resulting in false positives (FP) or false negatives (FN) (19). We indeed observed that shorter reads (28–32 bp) resulted in a higher total species count than longer reads (Fig. 3, A, left). We first established criteria to preliminarily filter the results (Supplemental Methods). After filtering, the number of species detected in the sample decreased markedly (Fig. 3, A, right). Next, we compared the performance of the 4 threshold metrics, including reads count, RPM, RPM ratio (20), and normalized RPM (3) (Supplemental Methods). We calculated the average F1 scores of all read lengths for each metric under different threshold values to find the optimal metric and threshold combination for identification of bacteria and fungi (Fig. 3, B and C). The results indicated that the 4 types of threshold metrics were comparable. The optimal threshold was chosen to maximize the average F1 score across all read lengths. The RPM threshold of 64 was optimal for bacteria, whereas normalized RPM was best fit, and the threshold was 0.05 for fungi. For virus pathogen detection, reads count $\geq$3 was used as an empirical threshold for virus detection, as described by previous studies (20, 21).

To determine the minimal optimal read length, the recall, precision, and F1 score for detecting bacteria and fungi under the optimal threshold at each read length were calculated (Fig. 3, D). The 3 metrics were similar at different read lengths for bacteria; however, the F1 score of fungi was lower at shorter read lengths (<36 bp), mainly because the sequences of shorter reads originating from fungi and from other eukaryotes were similar, which led to FP. Based on the previously discussed results, 36 bp was selected as the optimal shortest read length in subsequent analysis. Therefore, rapid mNGS reports were generated at 36 bp read length (cycle 45 of Read 1), with an estimated TAT of only 10.1 h for testing lower respiratory tract specimens (Figs. 1, B and 2, E, left panel).

LIMITS OF DETECTION

We spiked a mixture of 6 common respiratory pathogenic bacteria into different numbers (1.6×10^4–1.0×10^7) of Jurkat T cells for limit of detection evaluation (Supplemental Methods). The mixture was spiked in 3-fold dilutions. The high level of human host background DNA in these samples had a negative impact on the diagnostic sensitivity of microorganisms. According to the optimal threshold, bacteria had a limit of detection of 5000 to 20 000 cells ml^{-1} and fungi had a limit of detection of 50 to 100 cells mL^{-1} (Supplemental Fig. 1, C).

TEST ACCURACY

To determine the positive and negative results relative to culture or composite standard, we used a more stringent scoring algorithm as illustrated in Fig. 4, A, compared to the previous scoring standard reported by the Charles Chiu Lab (3). The accuracy of rapid mNGS report (36 bp read length) was then compared to a normal mNGS report (74 bp read length) and a 28-bp mNGS result. The specificity and sensitivity of 36 bp mNGS for bacterial detection compared to culture were 83.1% (95% CI 76.3%–88.7%) and 69.9% (95% CI 59.5%–79.0%), respectively (Fig. 4, B). The negative percentage agreement (NPA), positive percentage agreement (PPA), and F1 score for the rapid mNGS strategy when using the composite standard were 100% (95% CI 97.6%–100%), 85.1% (95% CI 79.8%–89.6%), and 92.0% (95% CI 88.9%–94.9%), respectively. Overall, the performance of mNGS testing for bacteria detection was comparable across varied read lengths.

However, rapid and normal mNGS results were superior to the 28 bp mNGS results in fungi identification. The specificity and sensitivity of 36 bp mNGS for fungi detection compared to culture were 95.5% (95% CI 90.9%–98.2%) and 73.9% (95% CI 58.9%–85.7%), respectively, compared to 58.1% (95% CI 50.8%–65.2%) and 80.4% (95% CI 66.1%–90.6%) for 28 bp mNGS results (Fig. 4, C). When using the composite standard, the NPA, PPA, and F1 score for rapid mNGS strategy were 97.4% (95% CI 93.4%–99.3%), 80.3% (95% CI 69.5%–88.5%), and 86.5% (95% CI 80.3%–91.5%), respectively, compared to 59.5% (95% CI 52.1%–66.5%), 85.5% (95% CI 75.6%–92.6%), and 59.6% (95% CI 48.3%–69.0%) for 28 bp mNGS results. Importantly, sensitivities and specificities for bacterial and fungal detection between the 36 bp rapid mNGS and the 74 bp normal method were comparable.

Due to the lack of an unbiased virus detection test, a composite standard using confirmatory research-based virus qPCR was adopted for virus pathogen detection. We only used qPCR to verify the presence of viruses in mNGS viral positive samples. Using the previously validated clinical mNGS thresholds for viruses (20, 21), 8 DNA viruses were detected (Supplemental Table 4). In total, 54 cases were positive for DNA virus in either 36 or 74 bp mNGS reports. For virus pathogen detection, the overall NPA, PPA, and F1 score for rapid mNGS strategy were 98.0% (95% CI 94.4%–99.6%), 96.8% (95% CI 89.0%–99.6%), and 96.1% (95% CI 92.1%–99.2%), respectively (Fig. 4, D).

Overall, of 153 rapid mNGS reports, 143 (93.5%) were adjudicated clinically relevant by an infectious disease and clinical laboratory specialist (Q.Y.) (Supplemental Table 2). Some FPs and FNs were found in the 10 clinically irrelevant reports. Subsequently, we examined the FN and FP cases in mNGS testing. FN results were obtained in 40 cases (50 species in total) for 36 bp mNGS results, and 50 cases (69 species in total) for 74 bp mNGS results (Supplemental Fig. 3, A;

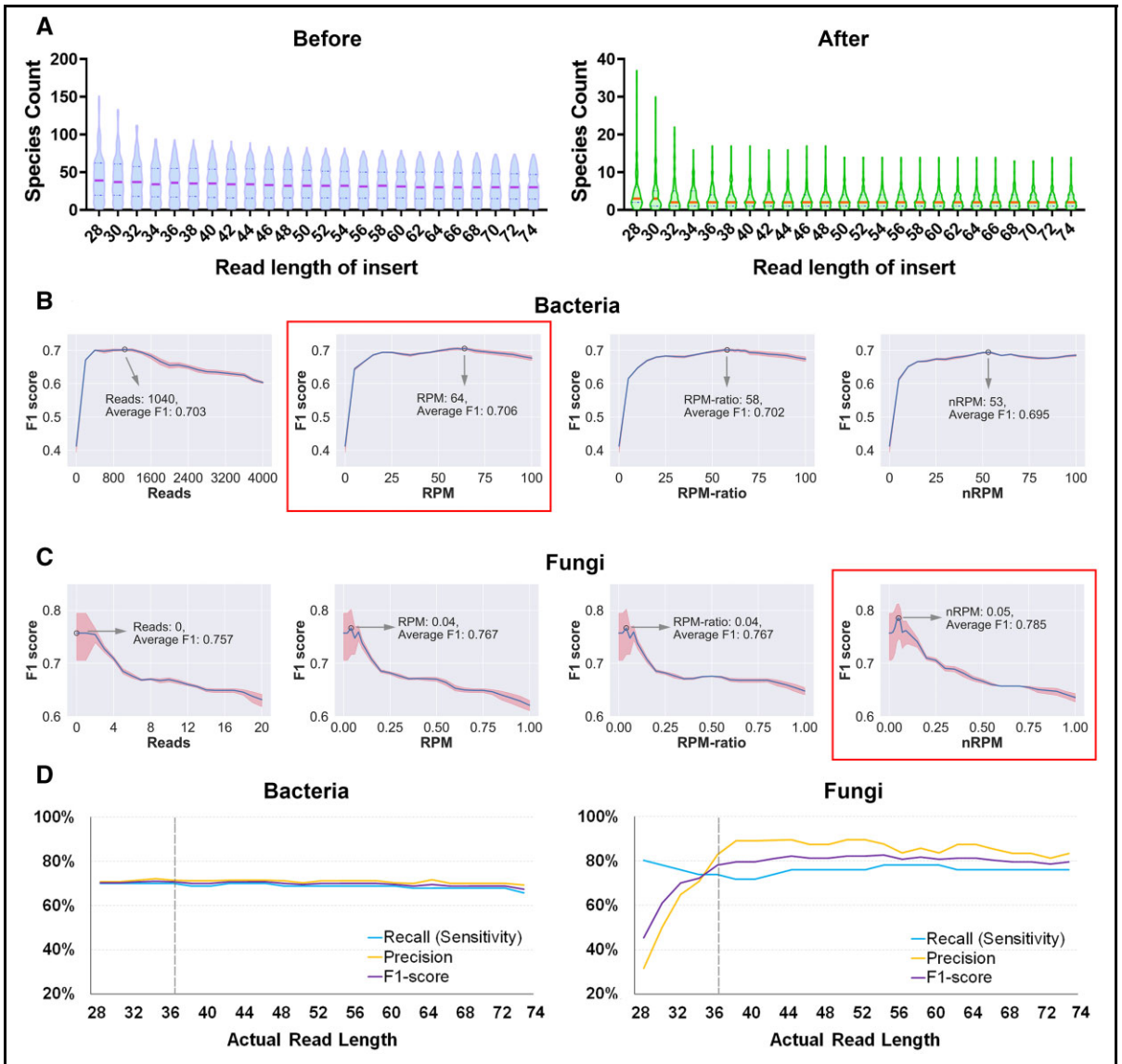

Fig. 3. Determination of optimal threshold and minimum optimal read length. (A), Violin plots of species count in each cycle of read length analysis. The left panel indicates species count after filtering low-level reads and closely related microorganisms. The right panel shows an additional filter of normal respiratory flora and known microbial contaminations. Solid lines in the violin plot represent the median, and dotted lines represent the interquartile; (B and C), Comparison of bacteria (B) and fungi (C) average F1 score (solid blue lines) with 95% CI (light red zone) of all read lengths under varying threshold across different metrics, based on culture. Threshold with maximum average F1 score under given metric is circled and labeled. The red box indicates the final threshold metric used for bacteria or fungi identification; (D), Recall, precision, and F1 score of mNGS analysis for bacteria (left panel) and fungi (right panel), based on culture at each read length under the optimal threshold. Color figure available online at clinchem.org.

Supplemental Table 5). In 36 bp mNGS results, 34 (68%) FN species in 29 cases were obtained, which were generally attributed to positive but subthreshold detection, whereas 13 (26%) species in 13 cases were completely undetected. The remaining 3 (6%) FN species in 2 cases were attributed to low intragenus abundance (Supplemental Fig. 3, B; Supplemental Table 5). On the other hand, FPs were recorded in 6 cases (7 species in total) for 36 bp mNGS results and in 5 cases (5 species in total) for 74 bp mNGS results

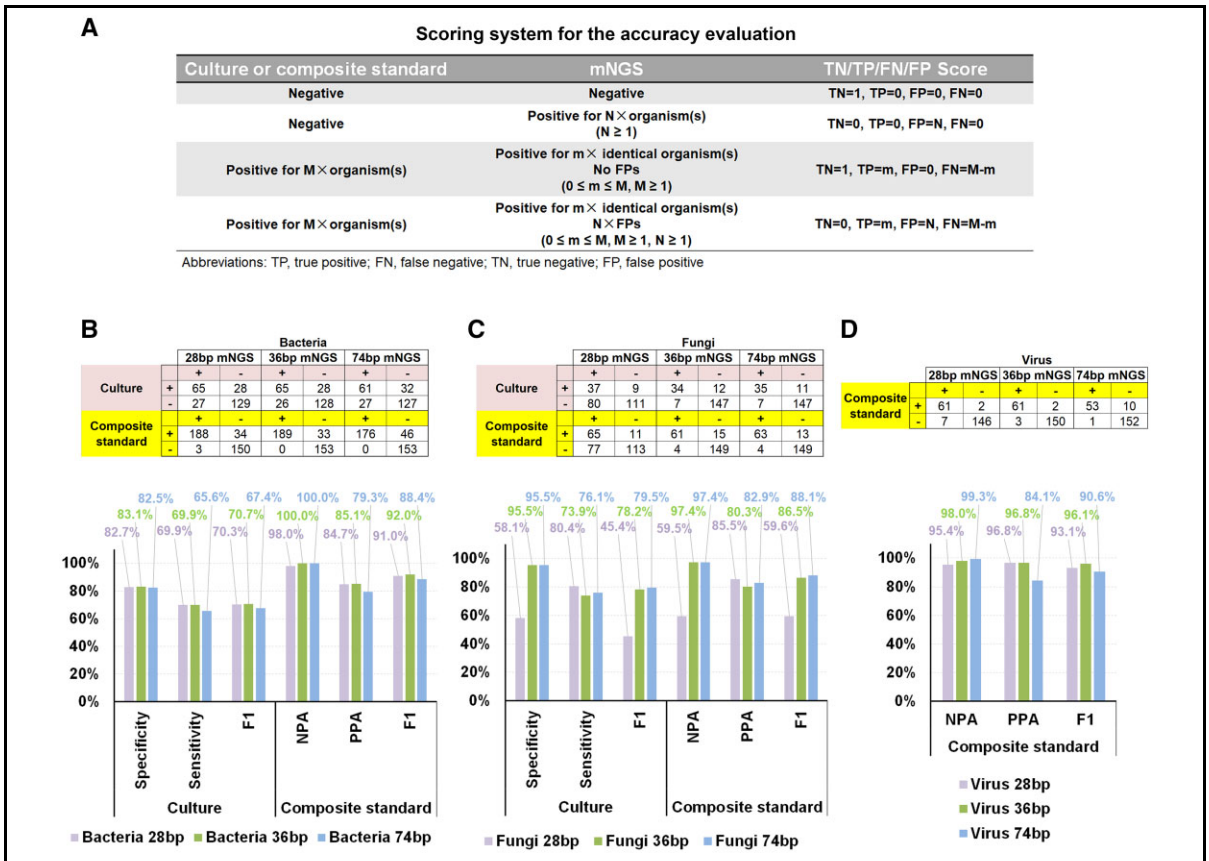

Fig. 4. Accuracy of rapid mNGS testing. (A), Scoring system for the mNGS accuracy evaluation. The number (n×) of false-positive or false-negative microorganisms in a sample was counted as n× FP or FN score value; (B–D), Upper panel, contingency tables of mNGS results for bacteria (B), fungi (C), and virus (D). Lower panel, histogram diagrams of specificity/NPA, sensitivity/PPA, and F1 score for each read length. NPA and PPA indicate specificity and sensitivity, respectively, when a composite standard was used. Color figure available online at clinchem.org.

(Supplemental Fig. 3, C and Supplemental Table 5). Of note, no bacteria FP in mNGS results was found compared to the composite standard. We speculated that some microorganisms grow slowly (e.g., *Streptococcus pneumonia* and fungi) in culture or are unculturable, but mNGS and other PCR-based testing methods with higher sensitivity could detect them leading to FP results compared to culture. Except for human alphaherpesvirus 2 in case 047, human betaherpesvirus 6B in case 056 and human gammaherpesvirus 4 in case 101, the presence of DNA virus in each of other positive cases was all confirmed by qPCR (Supplemental Table 4).

Discussion

A rapid and real-time mNGS solution was developed for unbiased metagenomic detection of pathogens. This approach had the following key advances: (i) internal index sequencing adaptors with different lengths of barcode sequences to ensure nucleotide diversity and optimal performance in multiplexed sequencing and to achieve the early assignment of reads by first sequencing the barcodes; (ii) an automated and clinically validated bioinformatics pipeline for real-time analysis of mNGS data; (iii) the automated library preparation device enhanced rapid mNGS detection of pathogens via the Illumina sequencing platform with TAT of 10.1 h for genomic DNA libraries or 9.1 h for cell-free DNA libraries; (iv) besides the cell-free DNA libraries prepared from body fluids, the solution could also be used for rapid sequencing and analysis of genomic DNA libraries in various samples; (v) a comprehensive evaluation of 4 mNGS threshold metrics for pathogen identification reported in previous studies (3, 20); and (vi) an accuracy evaluation of different read lengths in metagenomic sequencing. The sensitivities and specificities for bacterial and fungal detection of 36 bp and 74 bp read lengths were comparable. Therefore, mNGS reports of 36 bp read lengths can

provide timely and valuable information on etiology during severe infections or other urgent cases. Notably, although 28 bp is not the minimum optimal read length due to insufficient specificity, reliable mNGS results can be obtained earlier for some samples with relatively high pathogen burden, with an estimated TAT of 9.7 h. For instance, the 28 bp mNGS results identified pathogens similar to those in the culture results in Case 001 (*Candida albicans*, 7377 reads; *Acinetobacter baumannii*, 2629 reads) (Supplemental Movie). These results indicate the value of monitoring and analyzing mNGS results in real-time and the potential clinical benefits of shortening TAT and timely diagnosis. Comparison to other existing rapid metagenomic detection methods was outlined in the Supplemental Discussion.

Our study has some limitations. First, only detection accuracy in lower respiratory tract specimens was evaluated. Therefore, further studies should be performed to evaluate the test accuracy across other samples. Second, the internal index adaptors are not recommended when having only 1 to 3 libraries to sequence on the Illumina system since the nucleotide diversity of the barcode sequence will be low. However, it is not necessary to sequence barcodes when having only 1 library. Besides, the internal index adaptors can be combined to prepare 3 to 6 libraries for 1 single sample, ensuring well-balanced base composition of the barcode. Third, short read length (36 bp or even shorter) may result in loss or inaccurate mapping of the data, negatively impacting test specificity and sensitivity for some taxa. For instance, the genome sequences of different *Streptococcus* species are highly similar. Therefore, intragenus cross-contamination may distort taxonomic distributions and abundance rank, thus leading to 2 FN results of *Streptococcus pneumonia* in 36 bp mNGS analysis (Supplemental Fig. 3, A; Supplemental Table 5). Fourth, our custom analysis pipeline combines several well-established bioinformatics tools (Kraken2, Bowtie2, and BLAST) to align sequences. Using different custom analysis pipelines and other reference databases may lead to different thresholds for pathogen identification and different optimal shortest read length, but internal index adaptors and our strategy are still broadly applicable. Fifth, detection and analysis of RNA viruses were not conducted since patients with suspected RNA virus infections were excluded. Finally, this study was a retrospective and proof-of-concept study by using the residual samples after routine clinical testing in the microbiology laboratory. We could not confirm that a microbe found by various methods was pathogenic because our study did not affect the normal diagnosis and treatment of patients.

Supplemental Material

Supplemental material is available at *Clinical Chemistry* online.

Nonstandard Abbreviations: mNGS, metagenomic next-generation sequencing; TAT, turnaround time; %PF, percent cluster passing filter; RPM, reads count per million total reads; qPCR, quantitative PCR; FP, false positive; FN, false negative; NPA, negative percentage agreement; PPA, positive percentage agreement; ITS, internal transcribed spacer.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 4 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved.*

D. Zhang, J. Zhang, J. Du, Y. Zhou, J. Wang, and C. Ouyang conceived and designed the experiments. J. Zhang, P. Wu, Z. Sun, D. Zhang, J. Du, and J. Chen performed the experiments. J. Wang and Z. Liu developed the analysis pipeline. Y. Zhou, Z. Liu, and C. Ouyang analyzed the data. J. Zhang, Y. Zhou, C. Ouyang, D. Zhang, J. Du, and Q. Yang wrote the manuscript. Y. Xu helped to revise the manuscript. Q. Yang and C. Ouyang supervised the project and helped in result interpretation. All authors read and approved the final manuscript.

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: Y. Zhou, P. Wu, Z. Liu, Z. Sun, J. Wang, W. Ding, J. Chen, J. Wang, and C. Ouyang are employees of Hangzhou Matridx Biotechnology CO., Ltd.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: The National Natural Science Foundation of China (82072318), National Key Research and Development Program of China (2018YFE0101800), Beijing Key Clinical Specialty for Laboratory Medicine-Excellent Project (No. ZK201000) supported this study.
Expert Testimony: None declared.
Patents: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

References

1. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol 2019;37:783–92.
2. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med 2015;7:99.
3. Gu W, Deng X, Lee M, Sucu YD, Arevalo S, Stryke D, et al. Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. Nat Med 2021;27:115–24.

4. Jia X, Hu L, Wu M, Ling Y, Wang W, Lu H, et al. A streamlined clinical metagenomic sequencing protocol for rapid pathogen identification. Sci Rep 2021;11:4405.
5. Wang M, Fu A, Hu B, Tong Y, Liu R, Liu Z, et al. Nanopore targeted sequencing for the accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. Small 2020;16:e2002169.
6. Friedman ND, Temkin E, Carmeli Y. The negative impact of antibiotic resistance. Clin Microbiol Infect 2016;22:416–22.
7. Lindner MS, Strauch B, Schulze JM, Tausch SH, Dabrowski PW, Nitsche A, et al. HiLive: real-time mapping of Illumina reads while sequencing. Bioinformatics 2017;33:917–319.
8. Loka TP, Tausch SH, Dabrowski PW, Radonic A, Nitsche A, Renard BY. PriLive: privacy-preserving real-time filtering for next-generation sequencing. Bioinformatics 2018;34:2376–83.
9. Loka TP, Tausch SH, Renard BY. Reliable variant calling during runtime of Illumina sequencing. Sci Rep 2019;9:16502.
10. Tausch SH, Strauch B, Andrusch A, Loka TP, Lindner MS, Nitsche A, et al. LiveKraken–real-time metagenomic classification of illumina data. Bioinformatics 2018;34:3750–2.
11. Illumina. Calculating percent passing filter for patterned and nonpatterned flow cells. https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/hiseq-x-percent-pf-technical-note-770-2014-043.pdf (Accessed December 2021).
12. Illumina. What is nucleotide diversity and why is it important? https://support.illumina.com/bulletins/2016/07/what-is-nucleotide-diversity-and-why-is-it-important.html (Accessed December 2021).
13. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019;20:257.
14. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9.
15. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
16. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 2018;19:332.
17. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 2012;40:e3.
18. MacConaill LE, Burns RT, Nag A, Coleman HA, Slevin MK, Giorda K, et al. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. BMC Genomics 2018;19:30.
19. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Henaff E, Alexander N, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol 2017;18:182.
20. Miller S, Naccache SN, Samayoa E, Messacar K, Arevalo S, Federman S, et al. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. Genome Res 2019;29:831–42.
21. Chen H, Yin Y, Gao H, Guo Y, Dong Z, Wang X, 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 2020;71:S416–S26.