A streamlined clinical metagenomic sequencing protocol for rapid pathogen identification

Xiaofang Jia
Shanghai Public Health Clinical Center, Fudan University, Shanghai, China https://orcid.org/0000-0002-7127-0011
Lvyin Hu
Shanghai Public Health Clinical Center, Fudan University, Shanghai, China
Min Wu
Shanghai Public Health Clinical Center, Fudan University, Shanghai, China
Yun Ling
Shanghai Public Health Clinical Center, Fudan University, Shanghai, China
Wei Wang
Shanghai Public Health Clinical Center, Fudan University, Shanghai, China
Hongzhou Lu
Shanghai Public Health Clinical Center, Fudan University, Shanghai, China
Zhenghong Yuan
Key Laboratory of Medical Molecular Virology (MOE/NHC/CAMS), School of Basic Medicine, Shanghai Medical College, Fudan University, Shanghai, China
Zhigang Yi (yizhigang@shphc.org.cn)
Key Laboratory of Medical Molecular Virology (MOE/NHC/CAMS), School of Basic Medicine, Shanghai Medical College, Fudan University, Shanghai, China
Xiaonan Zhang (zhangxiaonan@shphc.org.cn)
Shanghai Public Health Clinical Center, Fudan University, Shanghai, China

Research Article

Keywords: mNGS, Illumina, Nanopore, pathogen identification, virus

Posted Date: August 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-53131/v1

Version of Record: A version of this preprint was published at Scientific Reports on February 23rd, 2021. See the published version at https://doi.org/10.1038/s41598-021-83812-x.

Abstract

Metagenomic next generation sequencing (mNGS) holds promise as a diagnostic tool for unbiased pathogen identification and precision medicine. However, its medical utility depends largely on the assay simplicity and reproducibility. In the current study, we aimed to develop a streamlined Illumina and Oxford Nanopore-based DNA/RNA library preparation protocol and rapid data analysis pipeline. The Illumina sequencing based mNGS method was first developed and evaluated using a set of samples with known etiology. Its sensitivity for RNA virus (Influenza A, H1N1) was $<6.4\times10^2$ EID50/mL, and a good correlation between viral loads and mapped reads was observed. Then the rapid turnaround time of Nanopore sequencing was testified by sequencing of an Influenza A virus and Adenoviruses. Further, 11 respiratory swabs or sputum samples pre-tested for a panel of pathogens were analyzed and the pathogens identified by illumina sequencing showed 81.8% concordance with qPCR-based results. Additional sequencing of cerebrospinal fluid (CSF) samples from HIV-1 positive patients with meningitis/encephalitis detected HIV-1 RNA and toxoplasma gondii sequences. In conclusion, we have developed a simplified protocol which realized facile metagenomic sequencing of a variety of clinical samples and pathogen identification in a clinically meaningful time frame.

Introduction

Historically, laboratory diagnosis of infectious diseases largely relied on microscopic examination and culture on appropriate media or cell lines. The advent of molecular biology techniques and sensitive RNA/DNA detection-by-amplification method has dramatically changed clinical practice on infectious diseases. However, these tests require prior knowledge of the infectious agent and not all molecular tests are readily available for all the suspected pathogens in clinical practice.

By contrast, metagenomic next-generation sequencing (mNGS) is a bias-free method that retains the key advantages of molecular tests and in the meantime, requires no information on the etiology of the disease. It reports a panorama of the microbes (virus, bacteria, fungi and parasite) present in the sample in a single assay [1-5]. In addition to clinical diagnosis, mNGS has also shown its potential in discovery of novel pathogens, a case-in-point was the recent outbreak of infectious pneumonia caused by hCoV-19[6, 7]. Thus, mNGS has widespread microbiological applications, including infectious disease diagnosis in clinical laboratories[8], pathogen discovery in acute and chronic illnesses of unknown origin[9], and outbreak surveillance on a global scale [6, 7, 10].

Despite the significant advantages of the mNGS approach, it also faces several technical and regulatory obstacles in order to be widely applied in clinical practice. Most obviously, the whole process usually takes several days and involves a long chain of wet and dry lab activities whose reliability need to be rigorously validated. In particular, the wet lab procedure usually involves extraction of minute amount of nucleic acids which are subsequently transformed into sequencing-ready libraries with high molecular efficiency. In addition, although genome sequencing technologies continue to develop with remarkable pace[11-17], analytical approaches for reconstructing and classifying metagenomes from mixed samples remain limited in their performance and usability[18]. Finally, pre-validated reference databases and sequence analysis pipelines which are aware of the common pitfalls of pathogen identification are needed for reliable reporting.

In this study, we attempted to address some of the issues by developing a broadly applicable, time and cost-effective mNGS method. Total nucleic acid from virus stocks and clinical samples including throat swab, sputum and cerebrospinal fluid (CSF) were extracted and used to construct separate DNA and RNA libraries, which were further analyzed on Illumina or Nanopore sequencing platforms. Our mNGS techniques shows good sensitivity and specificity with reference to conventional clinical tests and helped identifying additional respiratory viruses, HIV and toxoplasma gondii from clinical samples.

Results

Establishment of Illumina based mNGS method

First, a sensitive and streamlined metagenomic next-generation sequencing (mNGS) protocol was developed and evaluated using a series of virus positive samples. The general assay workflow was depicted in Figure 1. Efficient sample lysis was done using chaotropic salt-based buffer in combination with bead beating, this was followed by magnetic beads-based semi-automatic nucleic acid extraction. This process required about one hour. Another 4 or 7 hours were needed for generation of Illumina sequencing libraries starting from DNA or RNA respectively. Less than one working day (8 hr) is required for an experienced technician to process around 20 samples into sequencing-ready libraries. We tested this assay using representative DNA and RNA viruses (HBV positive serum, human adenovirus type 7, AdV7, human adenovirus type 3 AdV3 and Influenza A/Puerto Rico/8/1934 H1N1 (PR8). The resulting sequencing reads (9.86-70.68 million filtered reads) enabled recovery of full-length viral genomes with average coverage depth ranging from 3262.23 to 12745.27 (Table 1).

To further assess the sensitivity of virus identification using our method, especially for RNA virus, a dilution series of PR8 supernatant were tested. While 1530.02× (100% coverage) average depth was obtained for the original virus stock, a depth of 174.66× (90.30% coverage) and 11.98× (25.60% coverage) was achieved when the virus stock was 1/100 and 1/10,000 diluted (Table 2 and Figure 2). A good correlation between sample viral loads and number of total mapped reads was observed ($p=0.02$, $r=0.99$, linear regression) while reads generated from negative control showed no mapping. Although the genome coverage of virus with highest dilution factor (1/10,000, 6.4×10^2 EID50) was deceased to 25.60% with a total of 1605 reads mapped to PR8 genome, it was still more than enough for a reliable identification. These results suggested that the limit of detection for PR8 is well below 6.4×10^2 EID50.

Nanopore sequencing of RNA and DNA viruses

The single-molecule sequencing technology from Oxford Nanopore has the advantage of real-time data acquisition, which could significantly save the overall turn-around time. We first evaluated its performance on Influenza A virus using PR8 stock as a positive control. As shown in the cumulative read plot (Figure 3A, Table 1), within the first minute, viral reads were sequenced and continued to accumulate. In the first 2 hours, 2123 of the total 61432 (3.46% mapping rate) reads were mapped to one of its eight segments. At the end of the run, 13462 reads were mapped within 0.37 million reads. A near full coverage (99.46%)

of was obtained with average depth of 407 (Table 1). The genomic coverage plot of PR8 was shown in Figure 3B. After finishing sequencing the PR8 virus, we washed the sequencing chip and reloaded it with barcoded libraries generated with Adenovirus 3 and 7 DNA. Although the data generated was low due to the inactivation of most pores, we still found 11 from 22 reads in AdV7 and 3 out of 22 reads in AdV3. With such scarce reads data, nanopore sequencing allowed successfully assembly of 67.50% and 23.10% genomes sequences of AdV7 and AdV3 stock respectively (Figure 3C-D, Table 1). These results reflected the real-time capability of nanopore sequencing.

Validation with clinical samples

The established mNGS protocols were further tested with clinical samples. Eleven throat swab or sputum samples which had been tested for 41 known respiratory pathogens using Taqman array card read-time PCR were sequenced (Table 3). DNA and RNA libraries were constructed independently for each clinical sample. RNA or DNA sequencing results with more matching reads for each sample were shown in Table 3. For 10 samples tested positive by Taqman array, 8 were positively detected by our mNGS workflow, which include two FluA H1N1, one FluA H3N2, two Rhinoviruses, one Coronavirus OC43 and two Adenoviruses. Sequencing reads from two samples, positive for FluA (Sample # 10) and Haemophilus influenzae (Sample # 11) respectively, did not reach statistical criteria for pathogen calling. Sequencing results from one tested negative sample using array card did not show significant reads from these pathogens. Thus, our current mNGS sequencing method operated with 81.8% (9 in 11) concordance with qPCR-based results. Due to the usually low level of pathogen nucleic acids, these samples yielded mapped reads from 1.00 to 85.29 mapped reads per million (RPM) with a genome coverage of 2.03%–98.75%. For sample #1 with a Ct value 19.02, near full length H1N1 viral genomes (98.75%) and an average depth of 13.84 was obtained (Table 3).

We then performed sequencing on 20 CSF samples from patients with meningitis/encephalitis of various causes (AIDS related diseases, suspected CNS infection). HIV-1 sequences were identified in two of these patients with a mapping rate of 4.14 and 1425.45 RPM (Table 3), their serum HIV-1 RNA levels were 3.14E+04 and 7.57E+05 copies/mL, respectively. This confirmed previous reports of cerebral infection of HIV-1 in some AIDS patients [19]. Furthermore, toxoplasma gondii sequences (RPMsample=18024) were identified in one of these two samples (sample # 12, Table 3) with a mapping ratio of 61.7 (RPM-r) compared to the blank control (RPM_{NTC} =292, data not shown). Indeed, antibody against toxoplasma gondii was positive for this patient. This indicated the feasibility to identify potential parasite infection in CSF samples using our protocol.

Discussion

mNGS provides an unsurpassed level of genetic information for pathogen discovery, diagnosis, characterization, and genotypic classification that would not be available by using current clinical laboratory methodologies. It provides opportunities to screen for pathogens in outbreak situations from known and novel unexpected pathogens in a single run. Although highly promising in concept, it still faces a series of technical challenges for routine use in clinical laboratories.

The goal of our current study is to develop an easy-to-perform mNGS assay with minimal hands-on time while retaining adequate sensitivity to identify pathogens in clinical samples that could aid in the diagnosis of infectious etiologies in patients. The current protocol incorporated construction of RNA and DNA libraries from total nucleic acid extracted semi-automatically. Effective RNA or DNA extraction from clinical specimens is critical for molecular pathogen detection. In this study, we utilized a glass bead method to break the cell wall to release nucleic acids that can be extracted by the lysis buffer. The sensitivity of our current protocol, especially for RNA virus, was validated by identification of Influenza A virus with less than 6.4×10^2 EID50 input. This method was found to be sensitive enough for detecting a series of RNA virus including influenza A virus, human rhinovirus, human coronavirus, and HIV in respiratory or central nervous system samples. In addition, library preparation protocols for two major sequencing platforms (Illumina and Oxford Nanopore) were developed. Pathogens could be identified in a clinically meaningful time frame. Using Illumina, one working day (8hrs) was enough for nucleic acids extraction and preparation of sequencing-ready library. With Nanopore sequencing, similar library preparation time was required but sequencing and pathogen calling can be performed in real time which made report within 24 hrs feasible. This approach has been increasingly used for molecular epidemiology of emerging infectious diseases [20-22]. Indeed, we quickly utilized this methodology in response to the COVID-19 outbreak. The sequencing results showed a 96.4% sensitivity on qRT-PCR confirmed COVID-19 clinical samples and 35.7% of them yielded >90% genome coverage (unpublished data).

In summary, a simplistic NGS workflow that realized time and human-cost efficient conversion from clinical sample to sequencing-ready library was developed without obvious sacrifice in sensitivity. Our assay performed well in identifying DNA/RNA viruses in our validation test set. The efficiency of current protocol in identifying bacteria, fungi and parasites should be further evaluated. Additional workflow improvements are still needed and are under way in several aspects. First, better sensitivity and genome coverage could be achieved by incorporating a targeted sequence capture panel [23] although retaining assay simplicity would be a challenge. Second, developing a fully automated sample-to-library procedure would be instrumental for wider use of mNGS and minimization of contaminations in average clinical laboratories. Finally, improved sequence analytics that are efficient, bias-free and rigorously validated would ensure reproducibility of reports.

Methods

Ethics statement

This study was approved by the Shanghai Public Health Clinical Center Ethics Committee. All the experiment protocol for involving humans was in accordance to guidelines of Declaration of Helsinki. Informed consents had been obtained from all the enrolled patients.

Sample collections and study subjects

HBV positive serum, human adenovirus type 7 (AdV7) and human adenovirus type 3 (AdV3) were isolated and collected in our previous studies [24, 25]. Influenza A virus (A/Puerto Rico/8/1934 H1N1) (PR8 for short) was provided by Prof. Zejun Li (Shanghai Veterinary Research Institute). Eleven clinical throat swab or sputum samples which had been tested for 41 respiratory pathogens (Human Adenoviruses, Human Bocavirus, Human Herpesviruses, Influenza A, Influenza B, Human Parainfluenza viruses, Coronaviruses, Rhinovirus, Enteroviruses, Haemophilus influenzae etc) using 384-well pre-configured TaqMan real-time PCR Array cards (Thermo Fisher, #4398986) were used to validate the clinical performance of our mNGS method. Another 20 CSF samples taken from cases from AIDS related meningitis or encephalitis cases with suspected infections were used to test our method.

Nucleic acid extraction

Samples were lysed by a guanidinium isothiocyanate based buffer and were subjected to bead-beating (Figure 1A). Subsequent carboxyl-coated magnetic beads (SpeadBead Magnetic Carboxylate, GE healthcare) were used to bind total nucleic acid (TNA) which was washed with isopropanol, 80% ethanol and dissolved in nuclease free water. TNA was then collected and split into aliquots for subsequent DNA and RNA library preparation for illumina or Nanpore sequencing.

Illumina library preparation and sequencing

DNA and RNA libraries were constructed independently for each clinical sample. For DNA libraries, we used a Tn5 transposase based tagmentation method (TruePrep TD503, Vazyme) followed by PCR (13-16 cycles) with indexed primers (TruePrep Index Kit V2, Vazyme). For RNA libraries, viral RNA was reverse transcribed and amplified (20 cycles) using SMARTer universal low input RNA kit (Takara) in which ribosomal RNA depletion was omitted. About 5 ng amplified product was used for library construction using TruePrep DNA kit (TD503, Vazyme). The amplified product (13-15 cycles) was purified using AMPure XP beads. Sequencing was performed on NovaSeq 6000 with 2×150 bp paired-end sequencing protocol and 10 to 150 million reads were generated for each sample. For each batch of samples, a pure water control or optionally a negative sample control (specific-pathogen-free) was included and analyzed in parallel.

Nanopore library preparation and sequencing

An influenza A strain, PR8 and two adenoviruses B (AdV3 and AdV7) stocks were analyzed by Nanopore sequencing as representative of RNA and DNA virus. TNA were extracted from these samples. For influenza A H1N1, viral RNA was reverse transcribed and amplified (35 cycles) using SMARTer universal low input RNA kit. About 1 μg amplified product was used for library construction using the SQK-LSK108 kit (Oxford Nanopore Technologies). For adenovirus B, TNA extracted from AdV3 and AdV7 stocks were used for library construction using the SQK-RPB004 kit (Oxford Nanopore Technologies) with 25 cycles of amplification. Each sample was amplified with a unique barcode primer provided in the kit. Libraries were sequenced on the MinION platform using R9 flow cells. The H1N1 sample was first loaded onto the R9 flow cell. After sequencing the H1N1 virus for 24 h, we washed the sequencing flow cell and reloaded it with barcoded libraries generated with Adenovirus 3 and 7 DNA. The MinION was run for up to 24 h for each group samples and the first 2 h of data were used for data process and alignment to evaluate the possibility of quick pathogen identification.

Data analysis

Paired-end 150 base pair sequences generated by Illumina sequencing were processed for classification and mapping using our rapid computational pathogen detection pipeline (Figure 1B). First, reads are preprocessed by Fastp v 0.20.0 [26] for trimming of adapters and removal of low-quality ($q<20$), too short (less than 30) and low-complexity sequences. Second, the qualified reads were mapped to the human reference genome using bowtie2 v 2.3.5 [27] and samtools v 1.9 [28] to remove human sequences. Third, the remaining unique, nonhuman sequences were then taxonomically classified against the viral genomes or NCBI nucleotide sequences (NT database, 98 GB) using Centrifuge v 1.0.4 [29]. Fourth, the unique, nonhuman reads were then mapped against the curated RVDB viral sequence database [30] or the reference sequence of the specific pathogen selected from the Centrifuge output summary using bowtie2 (v2.3.5). Genome alignments and genome coverage (%) were visualized using Tablet (v19.09.03) [31]. The sequencing data was analyzed in terms of the numbers of filtered reads, the number of reads aligned to the species-specific sequence, the number of mapped reads per million filtered reads, genome coverage (%) and coverage depth (average and maximum).

For nanopore sequencing data, raw FAST5 files from the MinION instrument are base-called by the Guppy (v 3.2.4). Base-called FASTQ files were processed by filtlong software (v0.2.0) for removal of low-quality ($q>7$) and too short (less than 100) sequences. The qualified reads were then aligned to the specific viral sequence using minimap2 (v 2.17-r941). Mapped reads were exported to a bam file using samtools and visualized using Tablet.

Positive reporting threshold and assay controls

For each batch of illumina sequencing library, “no template” control (NTC), i.e., nuclease-free water, was processed in parallel with samples and the resulting reads were used as background reference. Pathogen reporting threshold criteria were established to minimize false-positive results from contaminating microbial sequences. Identification of RNA viruses were reported based on analysis of RNA mNGS libraries, whereas DNA viruses, bacteria, fungi, and parasites were reported based on analysis of DNA or RNA library, depending on the abundance of pathogen mapped reads. For viruses, the threshold criteria were based on the detection of non-overlapping reads from ≥3 distinct genomic regions. For identification of bacteria, fungi, and parasites, a reads per million (RPM) ratio metric (RPM-r) was used, defined as RPM-r $=RPM_{sample}/RPM_{NTC}$, with the minimum RPM_{NTC} set to 1 [32]. A minimum threshold of RPM-r ≥ 10 was designated for reporting the detection of a bacterium, fungus, or parasite.

Declarations

Acknowledgements

The authors acknowledge funding received from the following sources: The National Science and Technology Major Project of China (2017ZX10103009-001, 2018ZX10305409-001-005); The National Natural Science Foundation of China (grant no. 81801991, 81873962, 81671998, 91542207, 91842309), Chinese foundation for hepatitis prevention and control-TianQing liver disease research fund subject (TQGB20200164). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

XN Z, ZG Y, HZ L and ZH Y conceived the study. XF J, M W and XN Z performed the experiments and analyzed the data in the study. W W and Y L provided samples. XF J, LY H, and XN Z drafted the paper and all authors reviewed and approved the manuscript.

Competing interests

The author(s) declare no competing interests.

References

1. Miller, S., et al., Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Research*, 29, 831-842 (2019).
2. Matranga, C. B., et al., Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol*, 15, 519 (2014).
3. Tang, P.; Croxen, M. A.; Hasan, M. R.; Hsiao, W. W.; Hoang, L. M., Infection control in the new age of genomic epidemiology. *Am J Infect Control*, 45, 170-179 (2017).
4. Lu, R., et al., Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*, 395, 565-574 (2020).
5. Simner, P. J.; Miller, S.; Carroll, K. C., Understanding the Promises and Hurdles of Metagenomic Next-Generation Sequencing as a Diagnostic Tool for Infectious Diseases. *Clin Infect Dis*, 66, 778-788 (2018).
6. Zhou, P., et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579, 270-273 (2020).
7. Wu, F., et al., A new coronavirus associated with human respiratory disease in China. *Nature*, 579, 265-269 (2020).
8. Dunne, W. M., Jr.; Westblade, L. F.; Ford, B., Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. *Eur J Clin Microbiol Infect Dis*, 31, 1719-26 (2012).
9. Chiu, C. Y., Viral pathogen discovery. *Curr Opin Microbiol*, 16, 468-78 (2013).
10. Zaki, A. M.; van Boheemen, S.; Bestebroer, T. M.; Osterhaus, A. D.; Fouchier, R. A., Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med*, 367, 1814-20 (2012).
11. Cazanave, C., et al., Rapid molecular microbiologic diagnosis of prosthetic joint infection. *J Clin Microbiol*, 51, 2280-7 (2013).
12. Naccache, S. N., et al., Diagnosis of neuroinvasive astrovirus infection in an immunocompromised adult with encephalitis by unbiased next-generation sequencing. *Clin Infect Dis*, 60, 919-23 (2015).
13. Wilson, M. R., et al., Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med*, 370, 2408-17 (2014).
14. Salzberg, S. L., et al., Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system. *Neurol Neuroimmunol Neuroinflamm*, 3, e251 (2016).
15. Picelli, S., et al., Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc*, 9, 171-81 (2014).
16. Picelli, S., et al., Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods*, 10, 1096-8 (2013).
17. Di, L., et al., RNA sequencing by direct tagmentation of RNA/DNA hybrids. *Proc Natl Acad Sci U S A*, 117, 2886-2893 (2020).
18. Rose, R.; Constantinides, B.; Tapinos, A.; Robertson, D. L.; Prosperi, M., Challenges in the analysis of viral metagenomes. *Virus Evol*, 2, vew022 (2016).
19. Spudich, S., et al., Persistent HIV-infected cells in cerebrospinal fluid are associated with poorer neurocognitive performance. *J Clin Invest*, 129, 3339-3346 (2019).
20. Lu, J., et al., Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell*, 181, 997-1003 (2020).
21. Fauver, J. R., et al., Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell*, 181, 990-996 (2020).
22. Faria, N. R., et al., Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature*, 546, 406-410 (2017).
23. Muller, C. A., et al., Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. *Nat Methods*, 16, 429-436 (2019).
24. Bai, L., et al., Extracellular Hepatitis B Virus RNAs Are Heterogeneous in Length and Circulate as Capsid-Antibody Complexes in Addition to Virions in Chronic Hepatitis B Patients. *J Virol*, 92, (2018).
25. Zhang, W.; Huang, L., Genome Analysis of A Novel Recombinant Human Adenovirus Type 1 in China. *Sci Rep*, 9, 4298 (2019).
26. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J., fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34, i884-i890 (2018).
27. Langdon, W. B., Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min*, 8, 1 (2015).
28. Li, H., et al., The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078-9 (2009).

29. Kim, D.; Song, L.; Breitwieser, F. P.; Salzberg, S. L., Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res*, 26, 1721-1729 (2016).
30. Goodacre, N.; Aljanahi, A.; Nandakumar, S.; Mikailov, M.; Khan, A. S., A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. *mSphere*, 3, e00069-18 (2018).
31. Milne, I., et al., Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform*, 14, 193-202 (2013).
32. Wilson, M. R., et al., Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *N Engl J Med*, 380, 2327-2340 (2019).

Tables

Table 1. Illumina and Nanopore sequencing results of positive control samples.

No.	Sample	Virus type	Virus titer	Sequencing method	Organism identified from metagenomic pipeline	NO. of filtered reads ($\times10^6$)	Total mapped reads	mapped reads/million (RPM)	Coverage (%)	Ave Coverage depth	Max Coverage depth
1	HBV positive serum	DNA virus	1.0×10^6 (copies/mL)	Illumina	Hepatitis B virus	9.86	318866	32339.35	100.00	12745.27	73727
2	Influenza A virus Puerto Rico/8/1934 (H1N1)	RNA virus	3.2×10^7 (EID50/mL)	Illumina	Influenza A virus (H1N1)	70.68	422426	5976.60	100.00	3262.33	11954
				Nanopore	Influenza A virus (H1N1)	0.37	13462	36383.78	99.46	407.00	1598
3	AdV7 virus stock	DNA virus	2×10^7 (TCID50/ml)	Illumina	Human adenovirus type 7	25.41	1694032	66667.92	100.00	6664.40	46698
				Nanopore	Human adenovirus type 7	0.000022	11	500000.00	67.50	1.29	4
4	AdV3 virus stock	DNA virus	6.4×10^5 (TCID50/ml)	Illumina	Human Adenovirus type 3	23.87	1266937	53076.53	100.00	5079.10	26655
				Nanopore	Human Adenovirus type 3	0.000022	3	136363.00	23.10	0.23	1

Table 2. Illumina sequencing results of serially diluted PR8 Influenza virus.

Sample No.	Sample Type	Dilution factor	Virus input (EID50/mL)	NO. of filtered reads ($\times10^6$)	Total mapped reads	Mapped reads/million	Coverage (%)	Ave Coverage depth	Max Coverage depth
1	PR8	/	6.4×10^6	20.72	205036	9895.56	100.00	1530.02	7684
2	PR8	1/100	6.4×10^4	20.45	20570	1005.87	90.30	174.66	742
3	PR8	1/10000	6.4×10^2	21.41	1605	74.96	25.60	11.98	325
4	Blank control	/	-	25.59	0	0	0	0	0

Table 3. Illumina sequencing results of respiratory and central nervous system samples.

Sample ID	Sample type	Organism identified from clinical testing	qPCR‡	Microbes identified	NO. of filtered reads ($\times 10^6$)	Lib type	Total mapped reads	mapped reads/million (RPM)	Coverage (%)	Ave Coverage depth	Max Coverage depth
1	Sputum	FluA H1N1*	19.02	FluA (H1N1)	16.72	RNA	1426	85.29	98.75	13.84	75
2	Throat swab	FluA H3N2*	26.74	FluA (H3N2)	9.60	RNA	14	1.46	7.15	0.40	5
3	Throat swab	FluA H1N1*	27.16	FluA (H1N1)	18.15	RNA	76	4.19	8.67	1.19	20
4	Throat swab	Rhinovirus*	25.99	Human rhinovirus	14.03	RNA	14	1.00	2.03	0.27	4
5	Throat swab	Rhinovirus*	32.32	Human rhinovirus	14.20	RNA	539	37.96	62.59	10.57	123
6	Throat swab	Coronavirus OC43*	29.31	Human coronavirus	15.90	RNA	582	36.60	8.101	2.53	153
7	Throat swab	Adenovirus*	33.13	Human adenovirus B1	21.50	DNA	150	6.98	25.15	0.50	8
8	Throat swab	Adenovirus*	24.22	Human adenovirus B1	20.24	DNA	346	17.09	53.66	1.35	21
9	Throat swab	Haemophilus influenzae *	29.60	N.D.							
10	Throat swab	FluA H1N1*	26.41	N.D.							
11	Throat swab	none*	N.A.	N.D.					/	/	/
12	CSF	HIV-1	7.57×10^5	HIV-1	11.53	RNA	22964	1425.45	96.73	350.00	1237
13	CSF	Toxoplasma gondii†	/	Toxoplasma gondii	16.11	DNA	290365	18024	0.90	0.22	12
		HIV-1	3.14×10^4	HIV-1	13.76	RNA	57	4.14	22.22	0.77	9

Notes: * Samples tested for respiratory pathogens detected by a customized respiratory TaqMan array card read-time PCR method.

Independent DNA and RNA library were prepared for each sample. Data of the indicated library type were listed in this table.

† Antibody against toxoplasma gondii tested positive. ‡ qPCR Ct value or virus titers (copies/

Figures

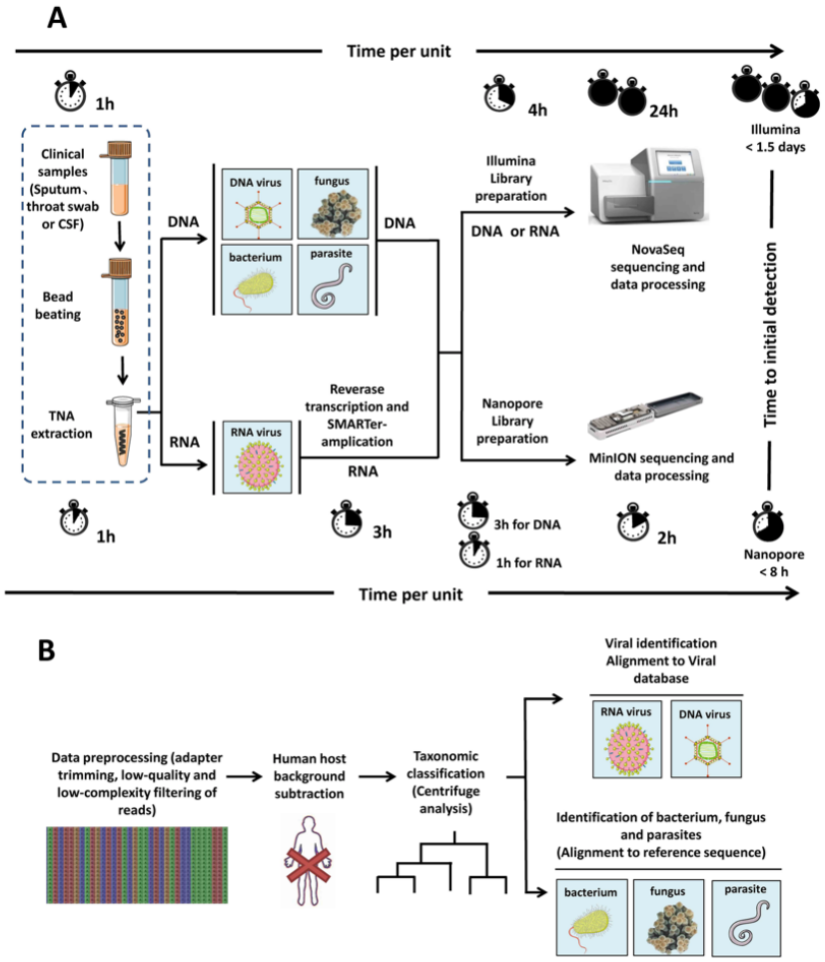

Figure 1

Schematic of the mNGS assay workflow. (A) Time-line of mNGS workflow. Total nucleic acids (TNA) were extracted by bead beating and guanidinium isothiocyanate based lysis. TNA was then collected and split into two aliquots for subsequent DNA and RNA library preparation which were further analyzed by illumina or Nanopore based sequencing. The time consumption of each step and the total time spent were indicated. (B) Sequence analysis workflow. Sequences generated by illumina and Naonopore were processed for alignment and classification. Reads are preprocessed by trimming of adapters and removal of low-quality/low-complexity sequences, followed by centrifuge software analysis to taxonomic classify microbial reads to family, genus, or species and alignment to the specific sequence of candidate pathogen.

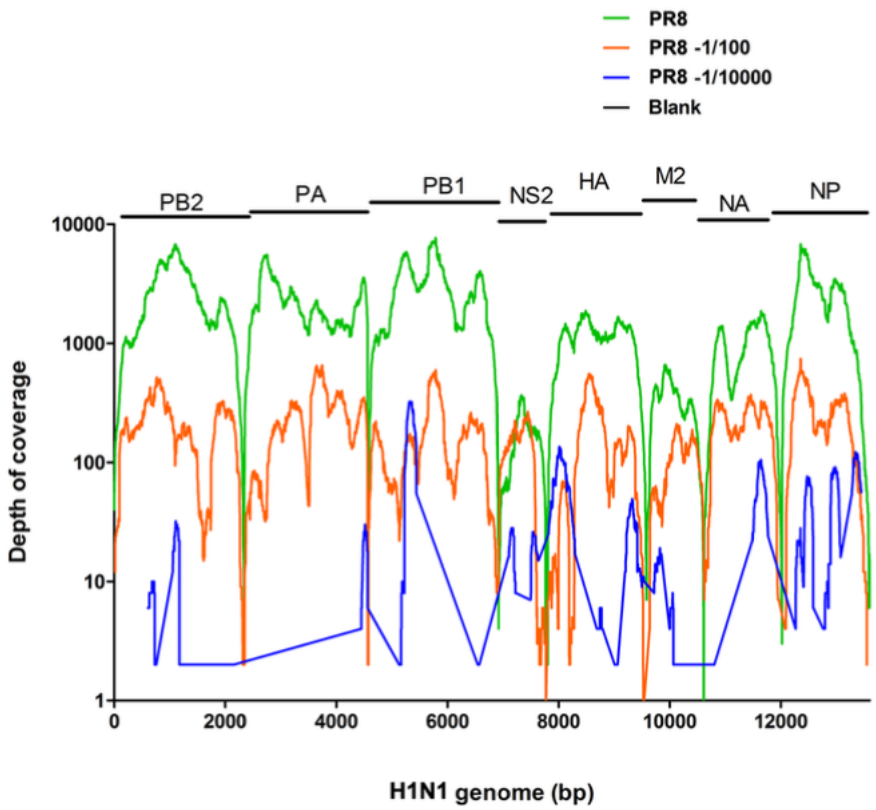

Figure 2

Sensitivity of the mNGS workflow. Genomic coverage from serially (undiluted, 1/100 and 1/10000) diluted PR8 supernatant and blank control.

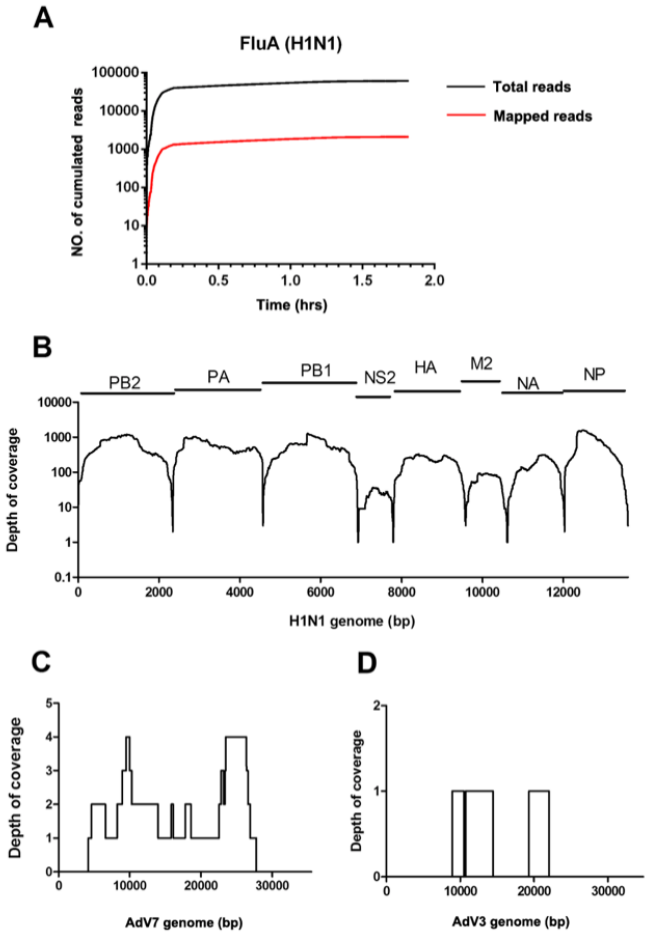

Figure 3

Performance of nanopore sequencing on selected RNA and DNA virus. (A) Cumulative read plot of H1N1. Genomic coverage plot of H1N1(B), AdV7(C) and AdV3(D).