Rapid Detection and Identification of Infectious Pathogens Based on High-throughput Sequencing

Pei-Xiang Ni[1,2,3], Xin Ding[4], Yin-Xin Zhang[2,3], Xue Yao[2,3], Rui-Xue Sun[2,3], Peng Wang[5], Yan-Ping Gong[2,3], Jia-Li Zhou[2,3], Dong-Fang Li[2,3], Hong-Long Wu[2,3], Xin Yi[2,3], Ling Yang[2,3], Yun Long[4]

[1]T-Life Research Center, Department of Physics, Fudan University, Shanghai 200433, China
[2]Binhai Genomics Institute, BGI-Tianjin, BGI-Shenzhen, Tianjin 300308, China
[3]Tianjin Translational Genomics Center, BGI-Tianjin, BGI-Shenzhen, Tianjin 300308, China
[4]Department of Critical Care Medicine, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China
[5]Clinical Laboratory, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China

Abstract

Background: The dilemma of pathogens identification in patients with unidentified clinical symptoms such as fever of unknown origin exists, which not only poses a challenge to both the diagnostic and therapeutic process by itself, but also to expert physicians.
Methods: In this report, we have attempted to increase the awareness of unidentified pathogens by developing a method to investigate hitherto unidentified infectious pathogens based on unbiased high-throughput sequencing.
Results: Our observations show that this method supplements current diagnostic technology that predominantly relies on information derived five cases from the intensive care unit. This methodological approach detects viruses and corrects the incidence of false positive detection rates of pathogens in a much shorter period. Through our method is followed by polymerase chain reaction validation, we could identify infection with Epstein–Barr virus, and in another case, we could identify infection with *Streptococcus viridians* based on the culture, which was false positive.
Conclusions: This technology is a promising approach to revolutionize rapid diagnosis of infectious pathogens and to guide therapy that might result in the improvement of personalized medicine.

Key words: Epstein–Barr Virus; Next-generation Sequencing; Whole Genome Sequencing

Introduction

Rapid detection and identification of infectious pathogens in clinical samples is very challenging, but it is essential to guide the therapy and predict the outcome. Traditional clinical microbial diagnostic methods like physiology and biochemical identification, serological tests, and automated detection technology, are mainly based on the conventional culturing of clinical samples.[1,2] The culture turnaround time is usually 1–2 days for most samples, or longer for samples that require prolonged incubation, like blood cultures.[3] More recently, once an isolate has been detected, species identification and genotyping could be performed following identification of a susceptibility test.[3] In fact, the whole procedure could take several days or even weeks, for example, in terms of the culture time required for tuberculosis, which is 42 days in the clinical laboratory. Therefore, it is not only a time-consuming process, but also is a barrier that impedes the rapid detection and identification of infectious pathogens. Interestingly, these culture-based methods usually fail to identify those pathogens that cannot be cultivated under standard techniques, such as viruses that can be identified through antibody-specific ELISA or sequence-specific molecular methods such as polymerase chain reaction (PCR) assay and array hybridization technologies.

With traditional culture-based diagnostic techniques being partially superseded by several methods, like matrix-assisted laser desorption ionization time-of-flight mass spectrometry, PCR, real-time PCR and molecular hybridization-based technologies, the speed and accuracy of both the diagnosis and prevention of infectious diseases have been greatly improved.[4-8] For instance, an internal transcribed spacer-targeted oligonucleotide chip was successfully used to detect and identify important bacterial pathogens associated with sepsis directly from the blood sample.[9] Another study also reported the use of a multiplex real-time PCR assay to identify members

Access this article online
Quick Response Code:
Website: www.cmj.org
DOI: 10.4103/0366-6999.154281

Address for correspondence: Prof. Yun Long,
Department of Critical Care Medicine, Peking Union Medical College
Hospital, Chinese Academy of Medical Sciences and Peking Union
Medical College, Beijing, China
E-Mail: ly_icu@aliyun.com

of the *Mycobacterium tuberculosis* complex.[10] Although the accuracy of pathogen identification has been improved, some newly emerging and unknown pathogens still cannot be identified using these current technologies.

In recent years, unbiased high-throughput sequencing (HTS) has been applied in medical microbiology as an emerging and powerful technique due to its low cost and rapid turnaround time. Some notable observations are the detection of a rapid hospital spread of antibiotic-resistant *Klebsiella pneumonia*, distinguishing *Legionella pneumophila* outbreak isolates from nonoutbreak isolates and identifying methicillin-resistant *Staphylococcus aureus* outbreak isolates.[11-14] Recently, the use of metagenomic analysis methods, which applies HTS has been proposed for infectious disease detection.[15,16] Although HTS has an important impact on the detection of pathogens, some challenges remain due to a lack of standard operating procedures.[17]

Here, we have developed an efficient, accurate and comprehensive method based on HTS for the rapid detection and identification of infectious pathogens directly from clinical specimens, which can be used in the context of more challenging and troublesome diseases such as a fever of unknown origin (FUO).

Methods

Clinical sample collection

The blood samples were collected from five patients, when they were transferred immediately to the intensive care unit (Peking Union Medical College Hospital) with the same symptoms as FUO. And then blood samples were processed as the following steps: Centrifuged (4°C, 10 min, 1600 ×*g*) within 8 h, transfer plasma to another tube and delivered in dry ice type environments. Two blood samples, which were referred to as UPDID2017 and UPDID2011-1, were blood culture positive for *Pseudomonas aeruginosa* and *Streptococcus viridians*, respectively. The remaining three samples, which were referred to as UPDID2020, UPDID2026 and UPDID0559-1 were blood culture negative. All samples were collected in sterile tubes and were centrifuged at 1600 ×*g*, and then the plasma was transferred to another tube for performing the follow-up analyses. Ethical approval for the study was obtained from the Institutional Review Board of BGI-Shenzhen. All patients had signed informed consent, and samples could then be used for research only.

Nucleic acid extraction, library preparation and sequencing

Nucleic acid (including DNA and RNA) was extracted directly from the clinical samples with QIAamp Viral RNA Mini Kit (QIAGEN). After extraction, the reverse transcription reaction was performed with PrimeScript RT-PCR Kit to generate single strand cDNA. Next, the double strand cDNA was obtained according to the Second Strand cDNA Synthesis Kit manufacture's instruction. Finally, the reaction product was purified using the QIAquick PCR Purification Kit (QIAGEN). The double stranded cDNA was disrupted into 200–300 bp fragments. The cDNA library was constructed after end repairing, A-Tailing, Adapter ligation and PCR. The library was sequenced using the Proton platform after quality control.

Data processed and analysis

The high-quality sequence data were generated through removing low quality reads, adapter contamination, duplication reads and discarding reads shorter than 35 bp. To eliminate the effect of the human sequence, the data were removed that was mapped to a human reference sequence hg19 using Burrows–Wheeler Alignment.[18] Subsequently, we mapped the remaining sequence to the bacterial, virus, and fungi database.

For a nonredundant bacterial database, we downloaded the complete genome set of version 20140829 from NCBI (*http://www.ncbi.nlm.nih.gov/*). For the nonredundant viral database, we chose 100 different viruses that are associated with human health based on version R194 from GenBank (*http://www.ncbi.nlm.nih.gov/genbank*). For the nonredundant fungal database, we downloaded the fungal genomic sequences associated with human disease from NCBI. At present, our databases contain 580 genus of bacteria, and 110 species of virus that are related to human diseases, and 14 species of fungi that can cause infectious in humans.

The time cost of our projected pipeline is almost 3 days (2 days and 10 h) based on the current technology of library construction and sequencing [Figure 1a]. The complete process of data analysis is shown in detail [Figure 1b]. This could still be extended according to one's requirements.

Statistical analysis

We used the software named SoapCoverage from the SOAP website (*http://soap.genomics.org.cn/*) to calculate the depth and coverage of each position. For bacterium, the result with coverage more than 1% and the number of stringent reads more than 10 was adopted. For fungi, the result with coverage of more than 0.1% and the number of stringent reads of more than 50 was reserved. Because of the high virus mutation rate, we chose the result with coverage of more than 10% and sequence length more than 1000 bp. The relative abundance of each microbe was made according to the following definition that modified the RPKM definition, because the pathogen content is ultimately low:[19]

$$\text{RPMM} = (10^{12} \times \text{C})/(\text{N} \times \text{L})$$

The RPMM represents the relative abundance of 1 million base-pairs in length, C is the number of reads that aligned to this microbe reference, N is the total number of reads, and L is the length of this microbe.

Polymerase chain reaction and Sanger sequencing validation

We carried out 16s rRNA identification for *P. aeruginosa* and *S. viridians*, and sequence-specific PCR identification for Epstein–Barr virus (EBV) with a target fragment of 250 bp to validate the result, respectively. The primers for 16S are (forward and reverse):

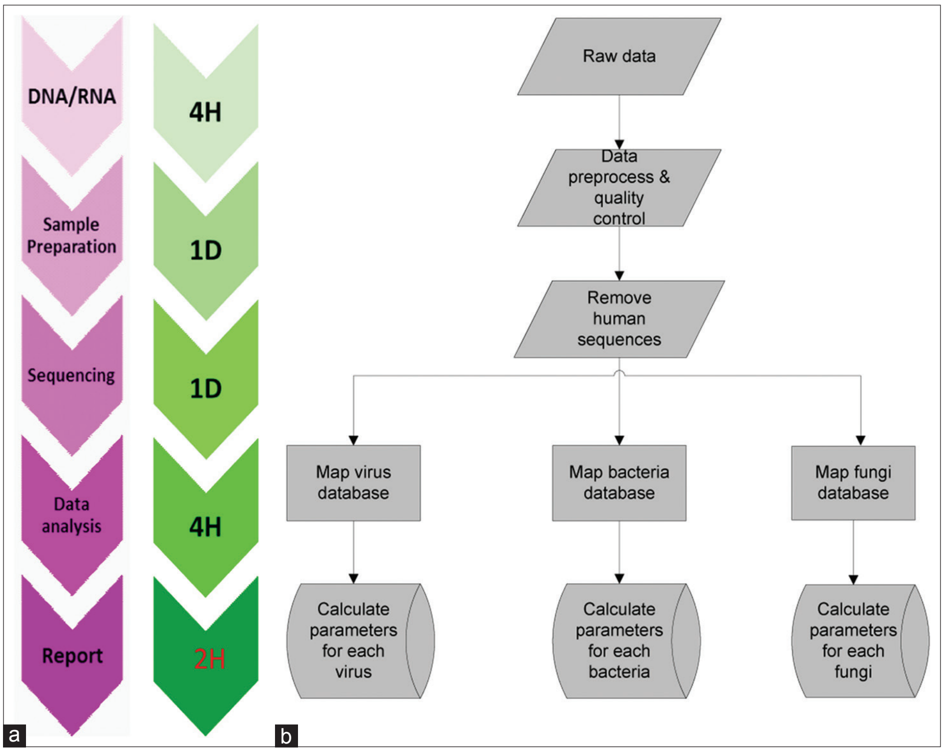

Figure 1: The pipeline of pathogen detection that contains detailed information. (a) The framework of rapid pathogen detection, which includes DNA/RNA extraction, sample preparation, DNA library construction, sequencing, and data analysis. The analysis took 2 days and 10 h. H stands for hour and D stands for day; (b) The complete data analysis process, followed by QC, trim host sequences, and database alignment.

5'-CAACGCGAAGAACCTTACC-3', and 5'-GACGGGCRGTGWGTRCA-3'

The EBV specific primers we used were:

5'-TCTGGCAGCTTTTGGCCTT-3', and 5'-GTTGGAGTTAGAGTCAGAT-3'

Subsequently, we sequenced these target products by Sanger sequencing.

Results

All these samples were cultured in the laboratory and performed sequencing with proton platform. Culture results and data statistics are listed in Tables 1 and 2, respectively. Species of *P. aeruginosa* and *S. viridians* in sample UPDID2017 and UPDID2011-1 were found to be positively grown. But in the other three samples, no microbes had been detected. For each sample, we adopted pooling sequencing with an 8 bp index for all of the samples, which improved the utilization of proton. The raw-data output was 5.1M, 46.2M, 4.9M, 6.1M and 55.7M reads for UPDID2017, UPDID2011-1, UPDID2020, UPDID2026 and UPDID0559-1 respectively, which fulfilled the analytical standard quantity.

After removing the human sequence, the remaining data were mapped to the bacterial database with the mapping rate from 0.08% to 15.22%. However, for viruses whose genome is short, the proportions were all lower than the bacterial data. Only UPDID2026 and UPDID0559-1 samples had a higher proportion of 2.27% and 2%.

Using our method based on proton sequencing, we detected the presence of *P. aeruginosa* in sample UPDID2017 with the coverage of 6.7%. This was consistent with the blood culture result. *P. aeruginosa* is a common opportunistic pathogen, and is capable of causing serious infections that can cause wound infection, otitis media, meningitis, respiratory tract infection, urinary tract infection and septicemia, with varying degrees of fever.[20] In addition, we also found traces of other bacteria with low coverage and depth, such as *Azotobacter vinelandii* (a nitrogen fixing bacteria), *Escherichia coli* and *Shigella sonnei*. *E. coli* is an opportunistic pathogen that mainly colonizes the human gut and sometimes causes intestinal diseases and genitourinary disorders. By contrast, *Shigella* is a pathogen that can cause bacillary dysentery. Given their low coverage and depth, we thought they may be the homologous sequences that cannot cause such infectious symptoms. According to our parameters and clinical symptoms, we confirmed that *P. aeruginosa* was the cause of fever in sample UPDID2017.

In sample UPDID2011-1 that was positive for *S. viridians* after blood culture, we did not find any trace of *S. viridians* or other bacteria or viral species, which suggested the possibility of a false positive in the blood culture. In fact, *S. viridians* is an important bacteria among normal human flora that mainly colonizes the oral microbiome, followed by the respiratory tract, gastrointestinal tract, and female genital tract. So, we suspected contaminants during blood culture.

Interestingly, we found EBV, namely human herpesvirus 4, with the coverage of 86% in the sample UPDID0559-1, which was negative when we performed blood culture. EBV is the cause of nasopharyngeal carcinoma, which is also associated with chronic fatigue syndrome, Infectious mononucleosis, and lymphoproliferative diseases.[21] In addition to EBV, we did not detect any other pathogens. Due to the uncultured nature of EBV using the conventional blood culture approach, we concluded that EBV was closely related to the fever in this patient. For the two remaining blood culture negative samples, in addition to *E. coli* and *Shigella* that were detected in other samples as fragmented and homologue sequences, we did not find any closely or relevant pathogens.

To profile for positive results whatever based on culture and sequencing, detailed information on data statistics is listed in Table 3, and the coverage and depth of *P. aeruginosa*, *S. viridians* and EBV along the genomic sequence in sample UPDID2017, UPDID2011-1 and UPDID0559-1 were figured out. We found that the coverage distribution of *P. aeruginosa* and EBV genomes were dispersed and uniform, although *S. viridians* contained reads in only some special regions [Figure 2].

Sanger sequencing was carried out on the original sample with 16s rRNA identification for UPDID2017 and UPDID2011-1, as well as sequence-specific PCR identification for UPDID0559-1. The results showed that *P. aeruginosa* and EBV were present in the original samples UPDID2017 and UPDID0559-1, and in RTI2011-1 there was no trace of *S. viridians*, respectively [Figure 3].

Table 1: The result of blood culture in clinical laboratory

Sample	*Pseudomonas aeruginosa*	*Streptococcus viridians*	Other
UPDID2017	+	−	−
UPDID2011-1	−	+	−
UPDID2020	−	−	−
UPDID2026	−	−	−
UPDID0559-1	−	−	−

"+" represents positive, and "−" represents negative outcomes.

Table 2: Data statistics

Sample	Raw data	Clean data	Nonhuman data	Bacterial	Virus	Fungi
UPDID2017	5.1M	4.9M	277.5K	4334	17	538
UPDID2011-1	46.2M	40.7M	2.9M	115	2	1217
UPDID2020	4.9M	4.7M	257.6K	4467	55	545
UPDID2026	6.1M	5.8M	295K	2873	51	478
UPDID0559-1	55.7M	52.9M	2.2M	320	5.3K	1138

Discussion

The current diagnostic paradigm consisted of four main stages, starting with the detection of a pathogen in the sample.[3] If a clinically relevant pathogen was detected, then this might be further tested for identification of drug susceptibility and epidemiological typing.[3] However, for most viruses, drug susceptibility, and epidemiological typing are not performed. The traditional method of detection and identification always takes several days or even weeks, which might delay the most optimal treatment times. Besides this, due to the limited resolution of the method itself, the accuracy still needs to be improved.

Therefore, in order to detect the pathogen directly in the clinical sample, we developed a rapid, efficient and accurate method containing experimental and bioinformatics analysis processes based on proton sequencing. The whole process including sample preparation, library construction, sequencing, data analysis and issuing reports, could take almost 3 days (2 days and 10 h). Compared with the traditional clinical microbial diagnostic methods based on conventional culturing, it shortens the whole period greatly. In addition, its ability to detect pathogens that are not cultivable or unknown has improved dramatically.

Despite the advantage of our method shows, there are still several problems to be improved upon. In regard to the experiment process, more technologies are essential to be developed such as with low DNA input based on proton sequencing. For most of the clinical samples, like blood samples, the nucleic acid content that is used for library construction after removing the majority of human cells is very low, and even at the nanogram (ng) scale, which will cause false negative result. This process is not only applied to blood samples, but it is also necessary to develop various methods of sample preparation for different types of samples. To optimize this process further, sample preparation must be simplified, which would reduce the turnaround time to a few hours and the need for highly skilled and trained technical staff.[3]

In the context of bioinformatics analysis, considering the present opportunistic pathogens belong to the human normal microbial flora, the threshold to determine whether a pathogen is clinically relevant should be more stringent. Thus, the research aiming at the distribution, coverage and abundance of the normal microbial flora should be carried out. Besides this, the database we used will be updated continuously. It would not only encompass bacterial and viral genomes, but will also incorporate fungal and protist genomes. It will also represent a catalogue of point mutations

Table 3: The annotation result for each sample

Sample no.	Species	Coverage	Abundance (RPMM)	Unique reads	GI accession
UPDID2017	*Pseudomonas aeruginosa*	430352/6764661	117.42	3332	NC_017549.1
UPDID0559	Human herpesvirus 4	144909/171823	583.61	4884	NC_007605

RPMM: Represents the relative abundance of 1 million base pairs in length, which was similar with the definition of RPKM value.

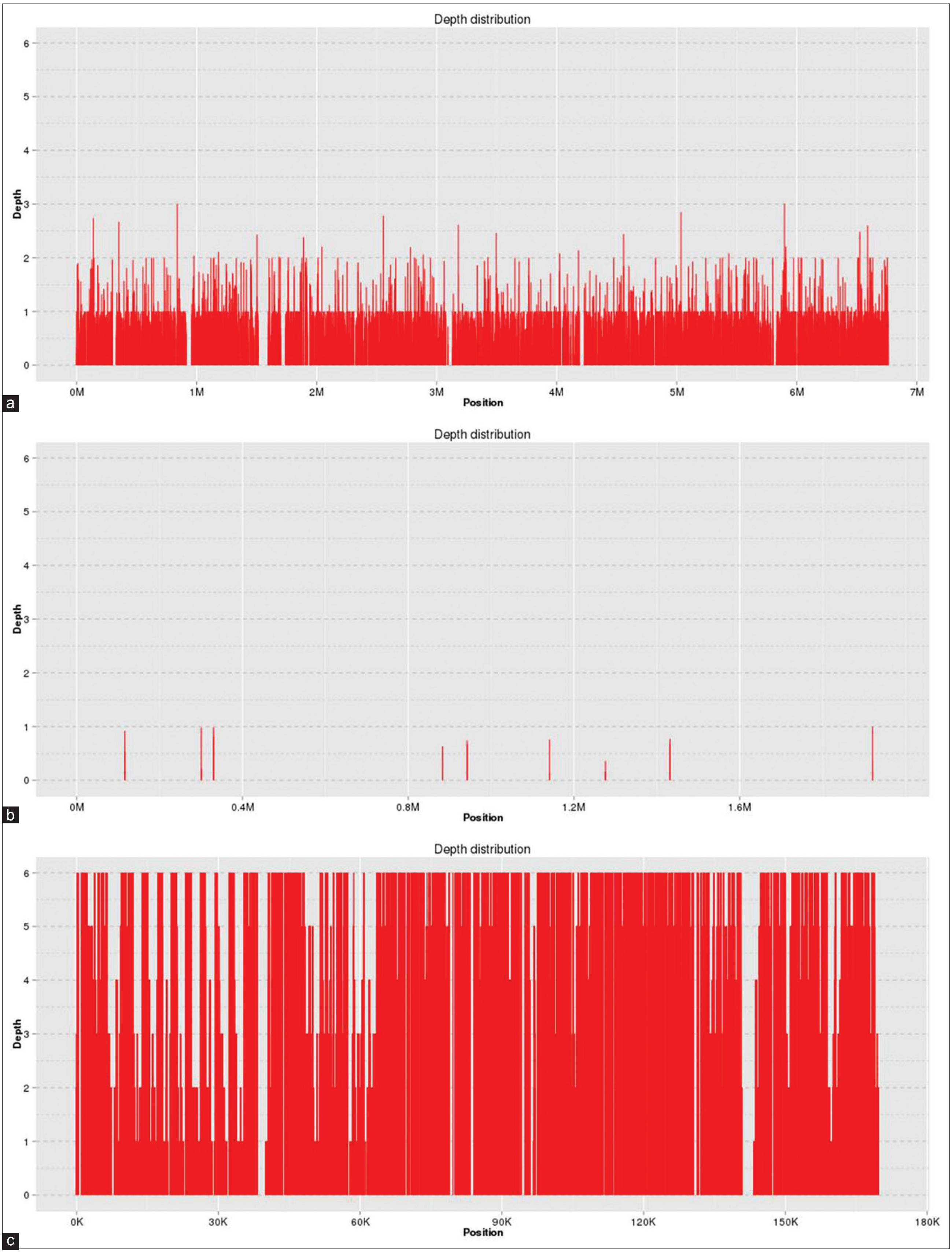

Figure 2: (a) The coverage profile of *Streptococcus viridians* we detected. The horizontal axis represents the position of each species while the vertical axis is the corresponding bin depth; (b) The coverage profile of *Pseudomonas aeruginosa* we detected. The horizontal axis represents the position of each species while the vertical axis is the corresponding bin depth; (c) The coverage profile of Epstein–Barr virus we detected. The horizontal axis represents the position of each species while the vertical axis is the corresponding bin depth.

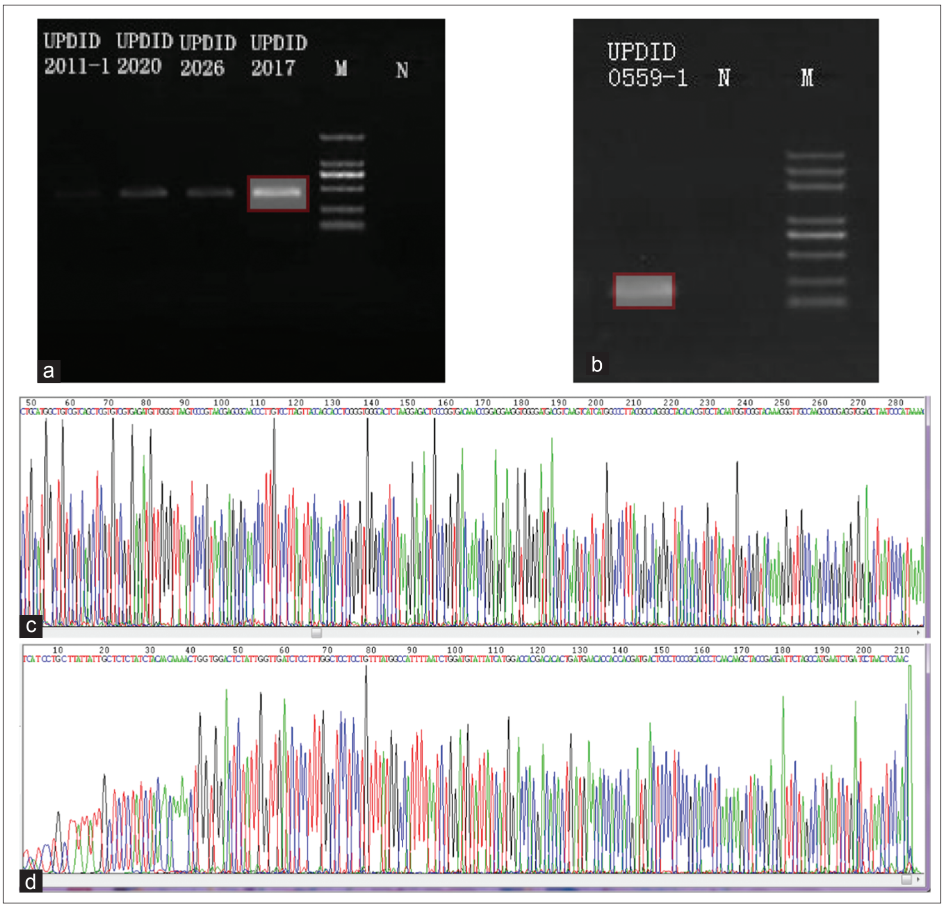

Figure 3: Polymerase chain reaction (PCR) result and Sanger sequencing result of pathogen validation. (a) The result of 16s PCR validation (UPDID2011-1, UPDID2020, UPDID2026 and UPDID2017). The marker we used was D2000; (b) The Epstein–Barr virus sequence-specific PCR identification of UPDID0559-1. The marker we used is Trans 2K Plus DNA Maker; (c and d) Is the Sanger sequence of target fragments. "M" and "N" represent marker and negative control, respectively.

or genes that account for drug resistance, which would also be added.[3] In the present study, we only employed our method to perform detection and identification of clinically relevant pathogens. However, both drug susceptibility and genotyping will be carried out in the future.

For those sequences that are both unmapped to human sequences and microbial databases, we should determine whether they are contaminated or whether this is due to homologous sequence of other eukaryotic genomes, or novel sequences of the human genome, or even perhaps novel microbial genomic sequences that we had not discovered thus far. However, unknown pathogen identification should be considered such as mapping to the NR database according to protein sequence homology. The sensitivity of HTS will cause false positive result.

Finally, the advancement of HTS sequencing technology will reduce the time and bring down the cost of HTS. With the increasing demand for rapid detection and identification of infectious pathogens in clinical samples, the method we developed would play an important role in the clinical laboratory and potentially revolutionize the capability of identifying infectious and novel pathogens.

References

1. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Møller N, *et al.* Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. J Clin Microbiol 2014;52:139-46.
2. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, *et al.* Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med 2014;370:2408-17.
3. Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, *et al.* Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog 2012;8:e1002824.
4. Neville SA, Lecordier A, Ziochos H, Chater MJ, Gosbell IB, Maley MW, *et al.* Utility of matrix-assisted laser desorption ionization-time of flight mass spectrometry following introduction for routine laboratory bacterial identification. J Clin Microbiol 2011;49:2980-4.
5. La Scola B, Raoult D. Direct identification of bacteria in positive blood culture bottles by matrix-assisted laser desorption ionisation time-of-flight mass spectrometry. PLoS One 2009;4:e8041.
6. Drancourt M. Detection of microorganisms in blood specimens using matrix-assisted laser desorption ionization time-of-flight mass spectrometry: A review. Clin Microbiol Infect 2010;16:1620-5.
7. Wallet F, Nseir S, Baumann L, Herwegh S, Sendid B, Boulo M, *et al.* Preliminary clinical study using a multiplex real-time PCR test for the detection of bacterial and fungal DNA directly in blood. Clin Microbiol Infect 2010;16:774-9.

8. Padmanabhan R, Mishra AK, Raoult D, Fournier PE. Genomics and metagenomics in medical microbiology. J Microbiol Methods 2013;95:415-24.
9. Kim CM, Song ES, Jang HJ, Kim HJ, Lee S, Shin JH, *et al.* Development and evaluation of oligonucleotide chip based on the 16S-23S rRNA gene spacer region for detection of pathogenic microorganisms associated with sepsis. J Clin Microbiol 2010;48:1578-83.
10. Reddington K, O'Grady J, Dorai-Raj S, Maher M, van Soolingen D, Barry T. Novel multiplex real-time PCR diagnostic assay for identification and differentiation of *Mycobacterium tuberculosis*, *Mycobacterium canettii*, and *Mycobacterium tuberculosis* complex strains. J Clin Microbiol 2011;49:651-7.
11. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, NISC Comparative Sequencing Program Group, Henderson DK, *et al.* Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Sci Transl Med 2012;4:148ra116.
12. Reuter S, Harrison TG, Köser CU, Ellington MJ, Smith GP, Parkhill J, *et al.* A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak. BMJ Open 2013;3:e002175.
13. Köser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, *et al.* Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med 2012;366:2267-75.
14. Harrison EM, Paterson GK, Holden MT, Larsen J, Stegger M, Larsen AR, *et al.* Whole genome sequencing identifies zoonotic transmission of MRSA isolates with the novel mecA homologue mecC. EMBO Mol Med 2013;5:509-15.
15. Chan JZ, Sergeant MJ, Lee OY, Minnikin DE, Besra GS, Pap I, *et al.* Metagenomic analysis of tuberculosis in a mummy. N Engl J Med 2013;369:289-90.
16. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ, Quick J, *et al.* A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. JAMA 2013;309:1502-10.
17. Fricke WF, Rasko DA. Bacterial genome sequencing in the clinic: Bioinformatic challenges and solutions. Nat Rev Genet 2014;15:49-55.
18. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.
19. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621-8.
20. Mathee K, Narasimhan G, Valdes C, Qiu X, Matewish JM, Koehrsen M, *et al.* Dynamics of *Pseudomonas aeruginosa* genome evolution. Proc Natl Acad Sci U S A 2008;105:3100-5.
21. Maeda E, Akahane M, Kiryu S, Kato N, Yoshikawa T, Hayashi N, *et al.* Spectrum of Epstein-Barr virus-related diseases: A pictorial review. Jpn J Radiol 2009;27:4-19.

Received: 04-12-2014 **Edited by:** Xiu-Yuan Hao
How to cite this article: Ni PX, Ding X, Zhang YX, Yao X, Sun RX, Wang P, Gong YP, Zhou JL, Li DF, Wu HL, Yi X, Yang L, Long Y. Rapid Detection and Identification of Infectious Pathogens Based on High-throughput Sequencing. Chin Med J 2015;128:877-83.

Source of Support: This study was supported by a grant from National Health and Family Planning Commission of the People's Republic of China (Special Fund for Health Scientific Research in the Public Interest) Program: No.201202011. **Conflict of Interest:** None declared.