**The Emerging Role of Metagenomics in the Diagnosis of Infectious Diseases**

Jun Li and Qiang Feng*

BGI-Research, Building 11, Beishan Industrial Zone, Yantian District, Shenzhen 518083, Guangdong, China

Received: June 27, 2014; Accepted: August 22, 2014; Published: September 08, 2014

*Corresponding author: Qiang Feng, BGI-Research, Building 11, Beishan Industrial Zone, Yantian District, Shenzhen 518083, Guangdong, China, Fax: +86-755-22354236; E-mail: fengqiang@genomics.cn

**Abstract**

Traditional pathogen identification techniques are mainly culture-dependent. Based on recent advances of next-generation sequencing, metagenomics is expected to revolutionize the mode of pathogen identification as a culture-independent technique. Compared to traditional techniques, the most evident advantage of metagenomics is the ability of detecting novel pathogens. As the cost of sequencing drops, metagenomics is expected to be widely applied in the clinical diagnosis system and contribute for the treatment of infectious disease.

**Keywords:** Pathogen; Infectious disease; Metagenomics; Culture-independent identification

**Introduction**

Infectious disease is a vital research area in modern medicine system. A certain type of infectious disease may cause the destructive damage to health and life of populations, for example diarrhea [1]. Infectious disease is derived from a pathogen, such as bacteria or virus [2,3]. By identifying the pathogen, clinicians can make the correct decision for diagnosis and treatment. Thus, the suitable approach of pathogen identification is extremely crucial for infectious disease. Here, we will introduce the transformation of pathogen identification from culture-dependent to culture-independent mode.

**Traditional culture-dependent mode of pathogen identification**

The mode of traditional identification is mainly based on pathogen culture. To identify the specific pathogen, we have to culture the microbe first. So, the suitable medium and conditions of culturing, as well as oxygen and temperature, should be determined. After culturing, staff would sequence the cultured microbe. Then the sequence of microbe would be aligned to known reference sequence for identification. Therefore, traditional identification approaches have several limitations. For example, if a certain microbe cannot be artificially cultured or not well, this potential pathogen will not be identified and the corresponding infectious disease would probably not be diagnosed correctly. Limited by the culture process, the traditional approaches cannot produce high-throughput data. It is not suitable for investigation of outbreak or large sample size of infectious diseases. Furthermore, the most evident disadvantage of traditional identification approaches is that they can merely identify those known pathogens, which have already been discovered to be associated with a particular clinical syndrome. For traditional identification approaches, novel pathogen is a big challenge that could not be solved theoretically. When a kind of unknown or unfamiliar infectious disease appears and causes outbreak, traditional pathogen identification approaches cannot help clinicians to diagnose and make correct treatment. Thus, it is essential to develop a completely innovative approach of pathogen identification without the limitations mentioned above.

**The new culture-independent mode of pathogen identification**

Next-Generation Sequencing (NGS) has contributed too many aspects of modern medicine, such as the progress in personalized medicine [4-6]. NGS refers to non-Sanger-based high-throughput DNA sequencing technologies, in which the sequencing process is parallelized. In principle, NGS extends the synthesis of DNA fragment across millions of reactions in a massively parallel fashion, rather than limited to a single or a few DNA fragment in Sanger sequencing. Thus, millions or billions of DNA strands can be sequenced in parallel, yielding substantially more throughput. This advance enables rapid sequencing of lots of DNA base pairs spanning the entire genome, producing thousands or millions of sequences concurrently in a single sequencing run and obviously lowering the sequencing cost. In the workflow of NGS, the genomic DNA (gDNA) is first fragmented into a library of small fragments that can be accurately sequenced in millions of parallel reaction. The newly identified bases are called reads. The reads will be aligned to a known reference genome if the genome of species is known. Otherwise, the reads will be assembled to construct a de novo genome with the absence of known reference genome.

Presently, the most popular NGS techniques are offered by Illumina, Roche and Life technologies. Distinct technical characteristics are exhibited by three types of sequencing platforms, Illumina Hiseq, Roche 454 and Ion proton. Hiseq machines could provide the largest amount of high-throughput...
The Emerging Role of Metagenomics in the Diagnosis of Infectious Diseases

data. The latest HiSeq 2500 will generate 600 G in a single run, leading to the lowest cost per G. Large numbers of samples are suitable for sequencing in HiSeq platform because of its cost advantage. Roche 454 can provide the best quality of de novo assembly with the advantages of the longest reads of assembly. Thus, microbes or complex genomes of species are suitable for sequencing by Roche 454. The limitation of Roche 454 is the cost, relatively the highest among the three platforms. The sequencing result of samples by ion proton can be obtained in the shortest time, which is its obvious advantage of Proton. But the amount of data output is much fewer than HiSeq. Proton is often used for sequencing small numbers of clinical samples with short turnaround time. However, the read is the shortest, and cannot provide the assembly quality as good as Roche 454.

No matter what platform is employed, the sequencing cost has dropped at an incredible speed in the past decade. In 2000, it costs 3 billion dollars to sequence a human genome. Along with the rapid technology development, the cost has been decreasing beyond Moore's law. It is widely expected that 1,000 dollars of sequencing a human genome is the threshold of clinical application. According to the decreasing trend of sequencing cost, the goal will be realized ahead of time by Illumina. In the first quarter of 2014, Illumina has announced that $1,000 is sufficient to afford the sequencing of the whole human genome based on the latest platform, HiSeq X Ten. Possibly, the cost will continue to drop to $100 per genome. It is easy to infer that sequencing technologies will be applied in many aspects of clinical medicine, such as prevention, diagnosis and prognosis.

Based on the power of NGS, researchers have developed metagenomics as a culture-independent identification approach. Metagenomics has the potential to revolutionize the identification of both known and novel microbes. As a recent report describes, the life of a 14-year-old boy was saved by NGS [7]. He had fever and headache for 4 months, but cannot be diagnosed correctly by diagnostic workup including brain biopsy. At last, clinicians conducted unbiased NGS test of the cerebrospinal fluid, and immediately detected the pathogen after 48 hours. Some sequence reads were found to correspond to leptospirosis infection. After confirmation by other identification approaches, the boy received personalized treatment and recovered his health quickly. Since NGS technology has developed rapidly in recent years, the sequencing cost continues to decrease in an unexpected speed. The culture-independent identification approach based on NGS is expected to play a significant role in modern medicine of infectious disease.

Shotgun metagenomics based on NGS, is an unbiased, robust and comprehensive technique for pathogen identification. In the metagenomics workflow, the sample would be prepared by enrichment, instead of culturing step in the traditional method. Then, nucleic acids would be extracted from clinical samples. Reads would be generated after sequencing of these nucleic acids. By bioinformatics analysis, microbe genomes would be assembled with long reads generated by NGS. Thus, without previous knowledge of microbes, shotgun metagenomics can technically generate whole-genome sequence data of microbes, not limited to certain regions of the microbe genome in traditional approach. The metagenomics-based identification result would reflect the comprehensive characteristics of microbes in the sample without bias, as shown by the Human Microbiome Project [8]. Recently, it is reported that metagenomics has shown its great power in identifying novel species and strains [9,10], as well as outbreaks of infectious diseases [1,2]. It is vital to detect pathogens as quickly as possible for appropriate treatment and prevention before outbreaks. The rapid and effective way to prevent potential outbreaks is to identify pathogens in a short time. However, the biased culturing step of traditional approach may lead to identification of wrong pathogens. Some slow-growing microbes will also cause time wasting. These problems can be avoided by application of metagenomics. The speed of pathogen discovery by metagenomics is increasing. A single genome of microbe can now be sequenced in a few hours [11,12]. Furthermore, known pathogens related to infectious diseases are limited, while the majority is unknown microbes. The advantages of metagenomics described above, without culturing step and previous knowledge of microbes, allows identification of novel pathogens as quickly as possible for clinical diagnosis [9,10].

As a revolutionary microbe identification technique, it is not difficult to conclude evident advantages of metagenomics in comparison with traditional techniques. The core advantage of metagenomics is the culture-independent characteristic. The microbes in the samples are not required to be cultured first for sequencing in the metagenomics workflow. The abundance of numerous microbes is very low inside or outside human body, and they are very difficult or ineffective to culture in the laboratory. Many of these low-abundance microbes are potential pathogens in many infectious diseases. For example, metagenomics have detected HPV type 6 and putative novel HPV types, as well as the molluscum contagiosum virus (MCV), in a human papillomavirus (HPV) negative sample previously detected by PCR. These viruses were all undetected using traditional techniques [13]. In addition to the culture-independent characteristic, another evident advantage of metagenomics is large-scale data output. Because metagenomics is a kind of NGS-based technique, the distinct high-throughput trait of NGS allows metagenomics to detect much more samples than traditional techniques. Sometimes, the pathogenic agent is not a single novel bacterium or virus, but a personalized combination of microbial species. According to the Human Microbiome Project (HMP), the microbiomes of different individuals have a great extent of diversity in composition. HMP has characterized the healthy microbiomes of 242 individuals, collecting microbe communities from 18 body habitats from five sites (oral, nasal, skin, gut, and urogenital) to study the role of these microbes in human health and disease [8]. By comparing the microbial diversity among body sites, the microbiomes differ significantly in taxonomic composition between individuals and between body sites. However, the functions of microbial metabolic pathways at each site remain stable for healthy human [8,14]. The high-throughput characteristics of metagenomics make it feasible to identify various microbes in a single test.
In this way, it is possible to compare taxonomic and functional differences of microbiomes between individuals in various health conditions. This is especially valuable for infectious diseases with complex etiology [15,16]. Thus, metagenomics provides a precious chance to demonstrate the dynamics and functions of potential pathogens, missed by culture-dependent techniques. Moreover, metagenomics will possibly establish the basis of personalized medicine of infectious diseases by revealing the difference of microbiomes among individuals.

The Application of Metagenomics in Infectious Diseases

In general, it is reasonable to anticipate that metagenomics would become commonly applied to the diagnosis and control of infectious diseases in the future. A recently published paper demonstrates the prospect of metagenomics in clinic [17]. Pneumonia can cause high rate of morbidity and mortality, which cannot be treated by new approaches in clinical practice. Lung was previously thought to be sterile without infection based on traditional culture-dependent microbial identification approach. But metagenomics, as the innovative technique of culture-independent microbial identification, has revealed the pathogenesis of pneumonia. Diverse and dynamic communities of microbes were identified by metagenomics in the lung microbiota. The ecosystem of lung microbiota was considered to have all features of a complex adaptive system, fundamentally different from the simple, linear systems demonstrated by the traditional model of pneumonia pathogenesis. Thus, metagenomics is expected to reveal the microbial communities responsible for many infectious diseases. Many potential pathogens, previously ignored by culture-dependent microbe identification techniques, would be supposed to be detected by metagenomics.

In addition, shotgun metagenomics exhibits high resolution in pathogen identification by de novo assembly of pathogen genome. Shotgun metagenomics can construct complete or nearly complete pathogen genome [18,19]. The high resolution of genome sequence provides the comprehensive characterization of microbes in the microbiome at deeper level. For example, the infectivity of diseases has long been underestimated clearly at the genotype level, caused by the resolution limitation of traditional techniques. The resolution of metagenomics technique is sufficient to detect person-to-person transmission events of epidemic diseases together with epidemiological information [15]. The problem of antibiotic-resistance is ordinary in infectious disease after long-term medication. Due to its advantages, metagenomics has been applied for the discovery of novel resistance genes [20]. With the widespread of metagenomics, a wider range of application in infectious diseases will appear to improve the understanding of antibiotic-resistance mechanisms.

The Current Challenges of Metagenomics

At present, metagenomics has proven its evident advantages for pathogen identification. It is widely recognized as the most prospective technique in the detection of novel microbes. Metagenomics has been used in a clinical diagnostic setting to identify the cause of outbreaks of viral infection [21]. However, metagenomics still can not completely replace traditional culture-based identification approach. For known pathogens, current metagenomics technique exhibits limited sensitivity for pathogen identification compared to traditional techniques. According to metagenomics, the study of diarrhea, a draft genome of the outbreak strain, \textit{Shiga-toxigenic Escherichia coli} (STEC) O104:H4 was quickly obtained by de novo assembly. The pathogen was immediately diagnosed by this culture-independent technique. Then, sequences from the \textit{Shiga}-toxin genes were detected in 27 of 40 STEC-positive samples. The result implied that the sensitivity of metagenomics was only 67% (27/40) compared to traditional culture-dependent technique [1]. The research results strongly suggests the potential of metagenomics as a culture-independent approach for rapid identification of bacterial pathogens during an outbreak of diarrheal disease.

Another Challenge of Metagenomics is the Large-Scale Data

Due to numerous samples in a certain biological context, metagenomics would generate enormous data. For example, more than 500 gigabases of sequence data were generated in a human gut microbiome project, from which the human gut microbiome gene catalog was assembled. Because of the complexity community in the metagenomic samples and the contamination in the environment, the massive amount of generated sequence data would inevitably contain noisy data. Certain fraction of sequence data may be derived from undesirable microbes in the environment. Thus, the significant challenge of metagenomics bioinformatics is to distinguish useful sequence data from the noisy background to assemble the genome of desirable pathogens in the environment. Furthermore, the massive amount of generated data will challenge the storage and management of computing system. Improvement in data-management and bioinformatics tools would significantly speed up the pathogen identification by metagenomics. These computational challenges will be well solved only with development of appropriate software and algorithms.

Although metagenomics has great potential for prevention, diagnosis and prognosis, challenges still lie ahead. Further technical advances are essential, including improving diagnostic sensitivity, speeding up and simplifying workflows of experiments and bioinformatics and reducing costs. All these progresses rely on the development of sequencing technologies. Thus, metagenomics currently serves as the complementary technique of traditional techniques. Although metagenomics is expected to become the standard technique for pathogen screening in the future, the application of metagenomics for clinical diagnosis is still in its infancy before the achievement of technical advances above.

Conclusion

Advances in sequencing technologies and bioinformatics revolutionize the techniques of pathogen identification from culture-dependent to culture-independent mode. Recent studies have shown that shotgun metagenomics has the ability to identify
novel and rare pathogens in infectious diseases in a short time. The high resolution based on de novo assembly of whole-genome sequences of pathogens allows us to deepen our research level from pathogen identification to antibiotic resistance and transmission, etc. Due to the limitations of sensitivity, cost and other problems, metagenomics is not the current standard pathogen identification approach. Along with the technical advances of sequencing, metagenomics will be increasingly applied to clinical diagnosis of more novel infectious diseases in the future. However, for clinical diagnosis of infectious diseases, the identification of potential pathogen genome in a clinical sample is just the first step. Correlation is not equal to causation. The difficulty of disease research is the process of establishing the causation based on the result of correlation. More efforts to validate the potential pathogens, such as investigation in larger cohorts of samples or experiments on animal models, are required to establish the causal link between identification of pathogen and causation of a disease. Although shotgun metagenomics is a relatively new field, we are moving towards an era that metagenomics will be integrated into the clinical diagnosis system and significantly contribute to the treatment of infectious diseases.

References


