Unbiased Next-Generation Sequencing and New Pathogen Discovery: Undeniable Advantages and Still-Existing Drawbacks

Arianna Calistri and Giorgio Palù
Department of Molecular Medicine, University of Padova, Italy

(See the Major Article by Brown et al on pages 881–8, and Brief Report by Naccache et al on pages 919–23.)

Keywords. next-generation sequencing (NGS); whole-genome sequencing; computational pipelines; encephalitis; astroviruses.

Traditional tools for microbial diagnosis might be unsuccessful in identifying the causative agent of an infection. Since the advent of polymerase chain reaction (PCR) technology, the nucleic acid–based detection of pathogens in clinical samples has significantly improved the percentage of successful identifications. The most commonly used PCR-based methodologies require the knowledge of the microorganism’s genome sequences; however, this knowledge is not always available. A typical case is represented by the outbreaks of emerging pathogens, where the causative agent was never before identified, at least in the human population. Even more important for the daily practice is the case of diseases for which the diagnosis of the etiological determinant still poses a challenge in a high percentage of cases. A paradigmatic example is represented by encephalitis, where, in up to 60%–75% of cases, the identification of the causing pathogen remains unmet, even after broad-range PCR-based screenings [1–3]. Different reasons might account for this failure. For instance, the causative agent might be represented by a microorganism that was never before associated with encephalitis (unexpected) or that is completely unknown. The discovery of novel pathogens through the analysis of their genome requires in these cases a random, sequence-independent, and thus “universal” amplification strategy [4].

Unbiased next-generation sequencing (NGS) is emerging as an attractive approach for the detection of pathogens in different types of samples, including clinical specimens. NGS applied to microbial diagnosis ideally should allow the amplification and analysis of any type of genetic material present in the clinical sample with a dose-independent efficiency, without the need to design specific primers to preamplify target sequences. Because random/unbiased amplification amplifies the host nucleic acids along with the microbial ones, searching for the microbial nucleic acids is like looking for a needle in a haystack. Indeed, even though different strategies have been developed to either enrich the samples with microorganism nucleic acids or to decrease the amount of the host nucleic acids [5], the level of host background still limits the sensitivity of NGS in pathogen discovery. Moreover, the choice of the NGS platform is crucial, with read length and number of sequences generated per runs as key parameters guiding the selection [6].

Taking into account these aspects, the potentiality of unbiased NGS in recovering genetic material of unknown pathogens, especially viruses, has been already proven [5].

In the present issue of Clinical Infectious Diseases, 2 different articles, one by Brown et al [7] and one by Naccache et al [8], address the problem of the diagnosis of unexplained encephalitis in 2 immunocompromised patients, an 18-month-old boy and a 42-year-old man, respectively, by employing an unbiased NGS approach.

In both cases, high-throughput RNA sequencing (RNA-Seq, Illumina Miseq) allowed the identification of a novel astrovirus, HAstV-VA1/HMO-C-UK1, as the potential causative agent of the fatal encephalitis. Astroviruses are nonenveloped, single-stranded RNA viruses, commonly associated with gastroenteritis in humans and only recently reported as possible causative agents of encephalitis [9, 10]. In one case, the patient was found to be infected by HAstV-SG, an astrovirus of the VA/HMO clade [9]. In the second case, the
human astrovirus 4 was detected in the brain tissues by reverse transcription (RT) PCR [10]. In both cases, the patients were severely immunocompromised.

These studies, along with the 2 reported in this issue of *Clinical Infectious Diseases*, suggest that the spectrum of causative agents associated with encephalitis should be expanded to astroviruses, at least in immunocompromised patients.

The work by Brown et al [7] and by Naccache et al [8] further supports the high potential of unbiased NGS in recovering novel viruses from clinical samples of patients suffering from diseases of unknown etiology. However, as recognized by both teams of authors, challenges still need to be faced before NGS can become fully exploitable at the clinical laboratory level.

First, in addition to the high cost of the equipment and the time necessary to successfully complete the entire procedure, one of the main challenges to be addressed is represented by the significant computational hurdles posed by the NGS approach. This is due not only to the large amount of sequencing data generated, but also to the fact that the identification of pathogens’ nucleic acids in clinical samples is complicated by the presence of the usual preponderant host background. To overcome this problem, the most widely used strategy is represented by computational subtraction, in which reads are first sequentially aligned to reference databases to filter out sequences corresponding to host background [11]. In the article by Naccache et al, the authors employ a rapid cloud-compatible bioinformatic pipeline (sequence-based ultra-rapid pathogen identification [SURPI]) that they recently developed to fish the pathogen sequence in the National Institutes of Health GenBank reference database [8]. SURPI was previously used by the same authors to diagnose neuroleptospirosis in a child affected by fulminant encephalitis [12].

Additional bioinformatics analysis pipelines have been specifically developed for the recovery of pathogen sequences from NGS data [13–16]. Validation of these tools on a large number of clinical samples is required. Another aspect to be taken into consideration is that well-annotated reference databases for pathogens are still under construction and might be biased by prevalent sequences or may contain sequencing mistakes. In the study by Brown and coworkers, only 0.4% of the total reads could not be assigned to the human genome, and among 20 588 062 paired-end reads, 46 were assigned to the HAstV-VA1 sequence [7]. Interestingly, whereas 20% of the “nonhost” reads did not match to any database sequence, 79% were represented by sequences of the phiX bacteriophage and 500 reads were assigned to environmental bacterial species. The recovery of microbial sequences other than the pathogenic sequences might further complicate the NGS data analysis and the identification of a single causal infectious agent. It must be kept in mind that brain biopsies and cerebrospinal fluid (CSF), the main samples employed in these 2 studies, are quite special specimens in so much as few nucleic acids other than the ones belonging to the host are expected to be detected. Different samples, such as stools, that are colonized by many different commensal organisms, surely pose additional challenges. For these specimens, the sensitivity and specificity of the NGS methodologies, as well as the strength and accessibility of the computational analysis, need further improvement, and abundant sets of metagenomic data for disease-causing and normal flora are required.

Strictly related to this aspect is the validation of the data obtained by NGS when applied to pathogen discovery. Indeed, to be identified as causative agent of a disease, a microorganism should satisfy different requirements and ideally meet the Koch postulates. Thus, finding specific sequences by NGS is only the first step in determining whether or not a certain microorganism is associated with disease.

Yet another critical step is represented by the assembly of the whole genome of the candidate pathogen. Indeed, this information represents the basis for carrying out a detailed phylogenetic analysis that can provide, for instance, hints on the potential pathogenicity of the identified microorganism. This aspect is well described in the study by Naccache et al [8]. These authors were able to show that HAstV-VA1/HMO-C-UK1 shares 95% identity to HAstV-SG, previously detected in the brain tissue of an immunocompromised child with fatal encephalitis [9], and 52%–54% identity with astroviruses linked to neurological diseases in minks and cattle [8]. Furthermore, knowledge of the full sequence of the microorganism genome, or coverage of most of it, is necessary for the development of specific molecular or serological assays, which are essential to confirm the presence of the candidate causative agent in clinical samples. In this context, Brown et al were able to confirm the presence of HAstV-VA1/HMO-C-UK1 in the brain tissue, CSF, stool, and serum of their patient by real-time PCR assay and in the cell bodies/processes of the pyramidal neurons by immunohistochemistry. Naccache and coworkers were able to identify the astrovirus sequence in brain tissues of their patient by RT-PCR and by in situ hybridization. These molecular analyses are especially important when the new pathogen belongs to microorganisms difficult to grow, as is the case of astroviruses that cannot be easily isolated in vitro. Indeed, neither Brown et al nor Naccache et al report attempts to isolate the novel astrovirus from the clinical samples. In this respect, it has to be mentioned that in the case of Naccache and colleagues, low astrovirus levels in the CSF indicate inefficient production of free virions. Moreover, in their patient, only the CSF sample obtained in the intraoperative setting was found to be weakly positive.
by RT-PCR (cycle threshold = 34), whereas CSF samples collected 1 and 10 days prior to the procedure were negative [8]. These findings would raise the question of the need for biopsy to achieve a correct diagnosis of encephalitis, with clearly related ethical and medical issues. Besides, Brown et al, in addition to the brain biopsy, were able to detect HAsTV-VA1/HMO-C-UK1 in 3 of 3 CSF samples, as well as in stools and in the serum of their patient, before and after the onset of neurological symptoms [7]. These results emphasize the need of extending the analysis to a broad range of specimens in the efforts of achieving a diagnosis of neurological disease.

Finally, an epidemiological analysis aimed, for instance, at the identification of the candidate pathogen in patients and in asymptomatic individuals or in a specific geographic region/location could further support a causative link with disease. Brown and colleagues carried out such an epidemiological screening in a small scale (within the ward where their patient was hospitalized) and in a larger scale (in the local and in other hospitals) [7]. This analysis indicated a low prevalence of HAsTV in the hospitalized population, confirmed by a cross-sectional community study of diarrhea [17]. It further supported the presence of HAsTV in immunocompromised hosts and suggested the possibility of indirect transmission of the infectious agent between 2 patients.

It has to be mentioned that in both studies the patient died. Even though fatality, especially in the case of viral infections in immunocompromised hosts, is not a surprising event, it cannot be excluded that the timely identification of a potential causative agent might improve patient management, while providing a better control of the spread of infection.

In this respect, in the study of Naccache et al, after diagnosis of astrovirus infection, a treatment with ribavirin and of intravenous immunoglobulin high doses was attempted that might have temporarily stabilized the patient’s condition [8].

In conclusion, both studies show that unbiased NGS data analyzed with an efficient computational pipeline and followed by a wide array of confirmatory analyses might represent a powerful tool for the diagnosis of the causative agents of encephalitis and, in more general terms, of unknown/unexpected pathogens.

The more well-conducted studies applying NGS to the clinical setting that are published, the more we can expect that method standardization and development of simplified equipment and bioinformatic tools will lead to practical applications of this technology in diagnostic laboratories in the near future.

**Note**

**Potential conflict of interest.** Both authors: No potential conflicts of interest. Both authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

**References**