Virology 434 (2012) 162–174

Contents lists available at SciVerse ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Review

Computational tools for viral metagenomics and their application in clinical research

L. Fancello, D. Raoult, C. Desnues *

Aix Marseille University, URMITE, UM63, CNRS 7278, IRD 198, Inserm 1095, 13005 Marseille, France

ARTICLE INFO

Available online 11 October 2012

Keywords:
Virus
Metagenomics
Computational tools
Clinical research
Virome
Emerging disease

ABSTRACT

There are 100 times more virions than eukaryotic cells in a healthy human body. The characterization of human-associated viral communities in a non-pathological state and the detection of viral pathogens in cases of infection are essential for medical care and epidemic surveillance. Viral metagenomics, the sequenced-based analysis of the complete collection of viral genomes directly isolated from an organism or an ecosystem, bypasses the "single-organism-level" point of view of clinical diagnostics and thus the need to isolate and culture the targeted organism. The first part of this review is dedicated to a presentation of past research in viral metagenomics with an emphasis on human-associated viral communities (eukaryotic viruses and bacteriophages). In the second part, we review more precisely the computational challenges posed by the analysis of viral metagenomes, and we illustrate the problem of sequences that do not have homologs in public databases and the possible approaches to characterize them.

Contents

* Corresponding author.
E-mail address: christelle.desnues@univ-amu.fr (C. Desnues).

0042-6822/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.virol.2012.09.025

Viral infections and the need for better viral discovery tools

Viral infections may become more prevalent in the future as multiple factors contribute to the emergence of new viral pathogens (Delwart, 2007; Wang, 2011). The expansion of the human population has led to the removal of barriers between animal and human communities, which favors the development of zoonoses. In addition, modern immunosuppressive therapies create favorable environments for the replication of viruses that are not commonly pathogenic. Furthermore, the spread of viruses worldwide is promoted by globalization and climate change, which extend the active ranges for some viral vectors, and there still exist several common pathologies, such as encephalitis and many respiratory syndromes, for which extensive classical diagnostic testing has failed to determine the etiology and which are thought to be of viral origin (Glaser et al., 2003; Quan et al., 2007).

Thus, an improved detection of newly emerging and re-emerging viruses and a systematic characterization of the full range of viruses that infect humans are needed (Anderson et al., 2003).

Classical methods of viral detection have several limitations. First, most of them are based on isolation and culture of the viral pathogen, but frequently the virus or its host cannot be cultivated under laboratory conditions, or the virus does not exhibit its characteristic cytopathic effects in culture (Specter, 1992). Moreover, these methods target known agents, and they are thus unsuitable for the detection of unexpected pathological agents or for the discovery of new ones. Immunological assays, for example, fail to identify unexpected or unknown viruses because such viruses are usually too divergent to cross-react. With respect to molecular tools, viruses lack a universally conserved genetic marker to target, and PCR assays directed towards conserved sequences within viral groups can only identify close variants of those groups (Staheli et al., 2011; Rose et al., 1998). Although the use of a wide set of different and highly degenerate primers has allowed the identification of numerous viruses (Culley et al., 2003), it does not allow a systematic and comprehensive screening to determine the identity of every virus that may be present.

Viral metagenomics and its first applications

Metagenomics, which is commonly defined as the sequenced-based analysis of the whole collection of genomes directly isolated from a sample (Handelsman et al., 1998), overcomes the principal limitations of the classical tools for viral detection. In fact, unlike traditional techniques for microbial and viral identification, metagenomics does not require prior isolation and clonal culturing for species characterization, nor does it rely on previous assumptions about what organisms are expected to be present or the genomic sequences that are to be targeted. Thus, it is particularly suitable to provide a global overview of the community diversity (species richness and distribution) and functional (metabolic) potential and to identify new species. In principle, it allows the identification of any organism, including those commonly not detected because they are difficult to isolate and grow under laboratory conditions. Such organisms are estimated to constitute between 90% and 99% of microbial species (Rappé and Giovannoni, 2003; Pace, 1997). Indeed the method of viral isolation, library preparation and sequencing affects the type of viruses which are retrieved. These issues have to be considered when analyzing the taxonomical profile of a metagenome and will be discussed later (see "General considerations on technical issues and potential biases in metagenome preparation").

Metagenomics has a wide variety of applications from ecology and environmental sciences (Breitbart et al., 2002; Dinsdale et al., 2008) to the chemical industry (Lorenz and Eck, 2005) and human health (Turnbaugh et al., 2007; Ravel et al., 2010; Sullivan et al., 2011; Nakamura et al., 2009; Minot et al., 2011). Historically, it was first associated with the study of uncultured microbial organisms (bacteria and archaea) in environmental samples (Handelsman et al., 1998; Hugenholtz and Tyson, 2008). More recently, it has also been applied to the characterization of viral communities, a task that it is particularly suited for because the small size of viral genomes makes their coverage more comprehensive using the same number of metagenomic sequences. The first example of viral metagenomics was performed by Breitbart et al. in 2002. This study revealed that viral diversity had been widely underestimated because, in approximately 200 l of marine water, more than 7000 different viral genotypes were found. This high degree of viral genetic diversity has been confirmed by further metagenomic studies of marine water (Angly et al., 2006), marine sediments (Breitbart et al., 2004) and freshwater (Lopez-Bueno et al., 2009). Today, viruses are considered the most abundant and diverse living forms on earth (Culley et al., 2006; Suttle, 2005). Their diversity has been explored by metagenomics in a wide variety of environments: oceans (Williamson et al., 2008), stromatolites (Desnues et al., 2008), acidic hot springs (Rice et al., 2001), and subterranean and hypersaline environments (Dinsdale et al., 2008).

Identifying human-associated viral communities (the human virome)

A preliminary step in identifying viral agents that cause disease is the characterization of the viral microflora associated with humans in a non-pathological state. To date, only a few viral metagenomic studies have been performed on human samples. Moreover, due to the limited availability and size of human samples, most of these studies used fecal samples (Reyes et al., 2010; Breitbart et al., 2008,2003; Minot et al., 2011,2012; Zhang et al., 2006; Kim et al., 2011).

The first contribution to the assessment of the human virome by metagenomics was made in 2003 by Breitbart et al. who studied the DNA virus community that was associated with the human gut through partial shotgun sequencing of the feces of a healthy adult. Most of the sequences generated were unknown (59% according to a tblastx search against the Genbank non-redundant database with an *E*-value $< 1e-03$). Among the identifiable viral sequences, the majority were phages (Breitbart et al., 2003). The community was estimated to have a high richness (approximately 1200 different genotypes) and diversity as estimated by the Shannon–Wiener index ($H' = 6.4$ nats) which determines species diversity on the basis of both the number of species and the relative contribution of each of these species to the total number of individuals in a community. Breitbart et al. performed an analogous study in 2008 using the feces of a 1-week-old infant. Similarly to the 2003 study, an elevated percentage of unknown sequences (66%) and a significant abundance of phages were found. Similar observations were also reported by two recent studies on the DNA virome of the human gut (Reyes et al., 2010; Minot et al., 2011) in which the percentage of unknown sequences was 81% and 98%, respectively, and phages dominated the viral community. However, the richness and diversity of these viral communities were significantly lower in comparison with the results obtained by Breitbart in 2003 and in particular to the 1-week-old infant, whose virome richness was 8 genotypes and whose Shannon–Wiener index was only 1.63 nats. In addition to the DNA viruses, the RNA viruses of the human gut have also been studied (Zhang et al., 2006; Nakamura et al., 2009). In a study performed using stool samples from two healthy adults, Zhang et al. found that only 8.9% of the sequences were unknown (tblastx search with $E < 1e^{-03}$) and that among the identifiable viral sequences there was an insignificant number of phages. The majority of the identifiable viruses were plant viruses (91.5%). Among these viruses, they found viruses that infect consumable