Virus Research 239 (2017) 82–86

Contents lists available at ScienceDirect

Virus Research

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

Review

Deep sequencing for discovery and evolutionary analysis of plant viruses

Marilyn J. Roossinck

Center for Infectious Disease Dynamics, Department of Plant Pathology and Environmental Microbiology, Pennsylvania State University, University Park, PA 16802, United States

ARTICLE INFO

Article history:
Received 15 July 2016
Accepted 18 November 2016
Available online 19 November 2016

Keywords:
Bioinformatics
Double-stranded RNA
Metagenomics
siRNA
virus-like particles

ABSTRACT

The advent of next generation sequencing (NGS), or deep sequencing, has allowed great advances to be made in discovery, diagnostics, and evolutionary studies in plant viruses. Various methods have been used for enrichment for virus-specific nucleic acids, each of which have some drawbacks. Many novel viruses have been discovered in plants by NGS technologies, and there is a good deal of promise for more comprehensive studies in virus evolution. However, each aspect of using NGS has its caveats that need to be considered, and there is still a need for better tools of analysis, as well as method for validation of sequence variation.

Contents

1. Introduction

Like many areas of science, plant virus research has been heavily impacted by the development of deep sequencing methods, or next generation sequencing (NGS). NGS tools have been used for virus discovery, diagnostics and analyzing the population structures and evolution of plant viruses and viroids. Metagenomics, defined as the analysis of microbial communities in environmental samples through sequence analysis, was first described in 1998 (Handelsman et al., 1998). Although early metagenomic studies were done using shotgun cloning and sequencing, the advent of NGS dramatically increased the efficiency of such studies for all microbes, including viruses. For plant viruses the picture of virus diversity from metagenomics is vastly different from that of recognized viral species, with the surprising discovery that viruses with persistent lifestyles are the most common type of viruses (Roossinck, 2012).

E-mail address: mjr25@psu.edu

http://dx.doi.org/10.1016/j.virusres.2016.11.019
0168-1702/

While the new sequencing tools are powerful, they are not without limitations. Bioinformatics tools were slow to develop, partly because the amount of data generated by NGS was so much more than any of us imagined in the beginning. Various methods of sequencing generate different read lengths, and different levels of sequencing error, making some methods more appropriate than others, depending on the applications (Wu et al., 2015). Here I will point out some of the areas of concern in drawing conclusions from NGS data, and areas where improved methods will likely address these concerns.

2. Virus discovery and diagnostics

The first use of NGS for virus discovery was for viruses in aquatic systems (Suttle, 2007); these methods have expanded to many terrestrial systems and have included wide-scale surveys as well as analyses of individual hosts. Plant viruses are found in many environments besides plants: sea water; fresh water; waste water; insects; and the feces of many animals (Culley et al., 2006; Djigeng et al., 2009; Kim et al., 2008; Li et al., 2010; Mehle and Ravnikar, 2012; Ng et al., 2014; Phan et al., 2011; Rosario et al., 2009; Tamaki et al., 2012; Victoria et al., 2009; Zhang et al., 2006). Most of these viruses are probably due to the consumption of plants, with some highly stable plant viruses passing intact through the guts of animals. The details of plant virus discovery through NGS have been reviewed in depth recently (Roossinck et al., 2015).

2.1. Discovery methods

Since there are no universal genes for viruses, deep sequencing methods often use random-priming for reverse transcription (RT) or PCR to obtain virus sequences. Using total RNA or DNA as a starting material is one approach (Dayaram et al., 2012; Rwahnih et al., 2009), but these methods yield a majority of sequence data that is not related to viruses, so most studies have used enrichment for viral nucleic acids (Table 1). For plant viruses there are no known giant viruses except in algae, and all of the currently known viruses from vascular plants have either RNA genomes, or small circular DNA genomes. One reason for the relatively small size of plant viruses is because they must move between plant cells through the size-restricted plasmodesmata. Many RNA viruses generate double-stranded RNA (dsRNA) during replication, so enrichment for dsRNA is a common strategy to obtain RNA viral nucleic acids (Roossinck et al., 2010). Extraction and processing of dsRNA is a labor intensive procedure, although a modification of this method was published recently that may improve the throughput (Blouin et al., 2016).

Another strategy is to use virus-like particles (VLPs), obtained through filtration or differential centrifugation, followed by extraction of total nucleic acids (Melcher et al., 2008). In plants this methods has had variable results. For example, when duplicate samples were tested using dsRNA or VLPs as a method of enrichment, many more viruses were found using dsRNA (unpublished observation).

The plant adaptive immune response involves the generation of small interfering RNAs (siRNAs) that target viral sequences for degradation (Ghoshal and Sanfaçon, 2015). These siRNAs have been used for discovery of RNA and DNA viruses in plants (Donaire et al., 2009; Kreuze et al., 2009). More recently this method was used to recover viroid sequences, including a novel viroid with structural, but no apparent sequence similarity to known viroids (Wu et al., 2012).

Finally *in silico* methods for virus discovery by examining databases such as expressed sequence tag (EST), siRNA or RNAseq libraries have proved to be an effective means of discovering new viruses and virus-like entities (Liu et al., 2012; Mambole et al., 2014). For the EST and RNAseq libraries this requires that the virus have a poly-A tail, so many plant viruses are missed. However, this does not require any wet lab work, and there are extensive databases of this type to be scanned.

2.2. Discovery versus diagnostics

When SARS emerged in 2003, it was considered groundbreaking that the complete genome of the virus was determined in just a few weeks (Marra et al., 2003). The advent of NGS has dramatically accelerated surveillance and discovery, while reducing costs, and new viruses are assessed in 48 h or less. However, the presence of a virus in a sample does not prove it is the causative agent, and with mounting evidence of benign and beneficial viruses (Roossinck, 2015), evidence of a virus sequence alone cannot be considered as definitive.

There have been attempts to use NGS for diagnostics in plant virus diseases (Mumford et al., 2016). This can be a powerful tool, and new bioinformatic tools can assist in finding viruses (Stobbe et al., 2013), although it can be difficult to determine the significance of viruses that are found this way. For example, in a transmissible lethal necrotic disease of corn in Kenya two viruses were found by NGS, even though they could not be detected by ELISA or electron microscopy, and were presumed to be causative (Adams et al., 2013). A novel closterovirus was found in wild roses exhibiting a severe disease phenotype through siRNA deep sequencing, along with three additional known viruses. Primers were developed for RT-PCR and in 20 collected samples 12 were positive for the novel virus, three were positive for one or more of the other viruses, but only seven had the disease. Experimental transmission was not discussed (He et al., 2015). Hence the causative agent is still unclear. In another example, deep sequencing of siRNAs recovered the complete sequence of the *Bell pepper endornavirus* (Sela et al., 2012), but this virus does not cause any plant disease. In the case of NGS for diagnostics, Koch's postulates, or any semblance of them, have largely been abandoned. NGS can detect viruses in diseased plants, but this does not mean that the virus is the causative agent (Castrignano and Nagasse-Sugahara, 2015).

2.3. Koch's postulates and modifications

The original postulates to determine a microbe as a causative agent of disease were established by Henle and Koch in the 19th century. They require that the parasite occurs in every case of the disease, and only in disease cases, and that it can be isolated, cultured and introduced to new hosts to cause disease. Koch recognized that in some cases causing disease anew in an experimental host was not possible, and also that in some cases culturing in the lab was not possible. He acknowledged that in some cases even the criterion that it be found in every case of disease but in no cases where disease was absent also could not hold; presumably some disease agents were asymptomatic in some hosts. The discovery and characterization of viruses as agents of disease added another level of complications to Henle-Koch's postulates. They could not be cultivated in a cell-free media (Evans, 1976).

Rivers (Rivers, 1937) recognized many of the difficulties of Koch's postulates for viruses and established different criteria. One aspect he discussed was diseases caused by a mixed infections, first described in plants (Dickson, 1925). He also recognized that the same disease can be caused by different viruses, that the manifestation of disease can vary from host to host, that individuals may suffer from more than one disease at a time, and that experimen-