Virus Research 188 (2014) 90–96

Contents lists available at ScienceDirect

Virus Research

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

Review

Current impact and future directions of high throughput sequencing in plant virus diagnostics

Sebastien Massart[a,*], Antonio Olmos[b], Haissam Jijakli[a], Thierry Candresse[c,d]

[a] Laboratory of Phytopathology, University of Liège, Gembloux Agro-BioTech, Passage des déportés, 2, 5030 Gembloux, Belgium
[b] Centro de Protección Vegetal, Instituto Valenciano de Investigaciones Agrarias (IVIA), Apartado Oficial, 46113 Moncada, Valencia, Spain
[c] UMR 1332 de Biologie du fruit et Pathologie, INRA, CS20032, 33882 Villenave d'Ornon cedex, France
[d] UMR 1332 de Biologie du fruit et Pathologie, Université de Bordeaux, CS20032, 33882 Villenave d'Ornon cedex, France

ARTICLE INFO

Article history:
Received 24 January 2014
Received in revised form 27 March 2014
Accepted 28 March 2014
Available online 6 April 2014

Keywords:
Plant virus
Virus detection
Virus diagnostic
High throughput sequencing
Next Generation Sequencing
Trends
Plant certification

ABSTRACT

The ability to provide a fast, inexpensive and reliable diagnostic for any given viral infection is a key parameter in efforts to fight and control these ubiquitous pathogens. The recent developments of high-throughput sequencing (also called Next Generation Sequencing – NGS) technologies and bioinformatics have drastically changed the research on viral pathogens. It is now raising a growing interest for virus diagnostics. This review provides a snapshot vision on the current use and impact of high throughput sequencing approaches in plant virus characterization. More specifically, this review highlights the potential of these new technologies and their interplay with current protocols in the future of molecular diagnostic of plant viruses. The current limitations that will need to be addressed for a wider adoption of high-throughput sequencing in plant virus diagnostics are thoroughly discussed.

Contents

1. Introduction

The ability to provide a fast, inexpensive and reliable diagnostic for any given viral infection is a key parameter in efforts to fight and control these ubiquitous pathogens. The past 40 years have seen tremendous progress in this area of virology, with the successive introduction of simple serological assays like the ELISA test, molecular hybridization, PCR in its various forms and real-time PCR (Martin et al., 2000; Mumford et al., 2006; Wetzel et al., 1991). Each of these techniques has improved our ability to efficiently diagnose viral infection, in particular in terms of sensitivity, specificity and reproducibility (López et al., 2009). However, the application of these techniques is largely restricted to known and decently well

* Corresponding author. Tel.: +32 81 622 431.
E-mail address: sebastien.massart@ulg.ac.be (S. Massart).

http://dx.doi.org/10.1016/j.virusres.2014.03.029
0168-1702/

characterized viral agents for which serological reagents and/or sequence information are available. For unknown agents or those still too poorly characterized, the diagnostician still faces very complex challenges that are only very partially met by the use of polyvalent serological or molecular assays or by the use of biological indexing. As a consequence a full virological indexing, *i.e.* the identification of all viruses present in a given sample, was until recently essentially an unattainable goal, as witnessed by the constant discovery of novel viruses. Recent developments in high-throughput sequencing (or Next Generation Sequencing – NGS) technologies and in bioinformatic analyses of the vast amount of sequence data thus generated have changed this situation drastically. Indeed, it is now conceptually feasible to detect any viral agent by high-throughput sequencing of the nucleic acids from a host and the identification of viral sequences of known or unknown agents in the generated sequences. Such developments, reviewed in details elsewhere (Prabha et al., 2013), have already produced key advances in the etiology of diseases (identifying the causal agent and allowing its characterization) and viral ecology (metagenomics) but also have the potential to strongly modify the way we see and perform virus diagnostics in the coming years. After briefly discussing recent developments of general interest, this review provides a snapshot vision on the current use of those approaches in plant virology, underlining the most relevant information for a diagnostician. The future developments as well as the current limitations that will need to be addressed for a wider adoption of these approaches in plant virus diagnostics are then extensively discussed.

2. Impact of sequencing trends and bioinformatics developments on virus discovery

2.1. Technological changes

Many NGS technologies have been developed so far and new technologies are currently being developed. These technologies and their performances have been reviewed in details elsewhere (Shokralla et al., 2012) and will not be specifically addressed here. It is worth to mention that during the past 10-years, the exponential growth in sequencing throughput has halved every 6 month the price per sequenced base, largely surpassing the evolution pace of any other technological field (see http://www.genome.gov/sequencingcosts/).

Briefly, current technologies are based on three fundamental steps: (i) preparation of the library of nucleic acids to be sequenced, (ii) the clonal amplification of the prepared libraries to produce a detectable quantity of DNA and (iii) the massive parallel sequencing of millions or billions DNA fragments in a single experiment. Current developments are focused on the simplification of library preparation and on the suppression of the clonal amplification step. For example, new technologies like PacBio RS II and Oxford Nanopore Technology do not need an amplification step. One major trend has also been to shorten the run time from weeks to a single day or a couple of hours. The classical detection method by fluorescence emission is also currently challenged by the rise of electronic detection strategies which eliminate the need for expensive scanning systems. Another trend is the development of cheaper bench-sequencers like the Roche GS Junior (discontinued in 2016), the MiSeq (Illumina) or the Ion Torrent PGM (Life Technologies), cutting prices and making them affordable for a growing number of laboratories.

2.2. Sequenced host genomes

To date, a number of plant species have their complete genome finished, including *Arabidopsis thaliana*, *Glycine max*, *Medicago truncatula*, *Oryza sativa*, *Populus trichocarpa*, *Solanum lycopersicum*, *Sorghum bicolor*, *Vitis vinifera*, *Musa acuminata*, *Zea mays*, *etc.* For many other plants or crops a high-quality draft genome is available, like *Carica papaya*, *Helianthus annuus*, *Manihot esculenta* or *Solanum tuberosum* (see http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html). From the practical diagnostics point of view, the access to more and more complete genome sequences of host plants opens the way to *in silico* subtraction approaches as already used for human pathogens, in which sequencing reads are first screened for homology to the host genome so that further analysis efforts are concentrated on non-host sequences.

2.3. Sequenced viruses and viromes

Part of the challenge in the bioinformatics analysis of NGS data for virus identification is that this step largely relies on the identification of homologies with already known agents (see below). The growing availability in public databases of genomic sequences for a wide diversity of viruses is therefore a key element in a successful diagnostic. More than 3500 reference sequences of virus (and viroid) genomes are now available at NCBI (see http://www.ncbi.nlm.nih.gov/genome) and 623 plant virus genomes are also available in Comprehensive Phytopathogen Genomics Resource (Hamilton et al., 2011).

Besides targeted sequencing efforts in individual hosts that allow the characterization of individual agents, a wide range of novel viruses have been identified in metagenomic efforts aimed at the characterization of virus populations in various environments like feces (Minot et al., 2012; Reyes et al., 2010), or fresh (Djikeng et al., 2009; Rosario et al., 2009) or saline aquatic environments (Williamson et al., 2008). These projects have already greatly expanded the viral genes and genomes catalogs and will continue to do so at an increasing pace in the coming years, contributing to an improved ability to identify viral sequences among NGS data. Nevertheless, the recent discovery and sequencing of giant viruses (~2 Mb) with only 7% of their genes with matches in databases also shows the limitations of our current knowledge (Philippe et al., 2013).

2.4. Bioinformatics development

Bioinformatics developments impact the four steps of any high-throughput sequencing project: quality control, sequence assembly into contigs, contigs annotation and identification of variations between samples.

The quality control is dependent on the sequencing technology used. Standard parameters and thresholds are usually provided by the manufacturer. It is now a very standardized process on “older” technologies like Roche pyrosequencing or Illumina Sequencing by synthesis. Given the increase in throughput of sequencing machines, an extra step of demultiplexing combined samples is more and more frequently used before the second step of sequences assembly.

The assembly of sequences to generate contigs can be done in two ways: *de novo* assembly or mapping of reads on (a) reference sequence(s). For *de novo* assembly, the (meta)genome is reconstructed by matching all the generated sequences to each other. This is considered the current gold standard for bacteria or virus genome sequencing. It is also the only feasible approach to characterize novel viral agents for which no reference genome is available. Each NGS platform developed its own bioinformatic tool, such as the Consensus Assessment of Sequence and Variation (CASAVA – Illumina) or the GS De Novo Assembler, Reference mapper and Variant Analyzer (Roche). In parallel, many softwares were developed specifically for *de novo* assembly or alignment