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*

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.

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).

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

(for more information, see http://en.wikipedia.org/wiki/List_of_sequence_alignment_software#Genomics_analysis).

The annotation of the generated contigs or of the remaining singletons is relatively straightforward and generally based on the use of well-known homology search programs such as Blast (in its various BlastN, BlastX, BlastP and TBlast versions) or Blat (Kent, 2002). This is a very straightforward process when the reads or contigs belong to a well annotated species or a relatively close relative. However, problems increase when the sequences belong to agents that are more and more distant from the viruses present in international databases. Although automated algorithms are now routinely used, there is frequently no or little manual assessment of accuracy (Foster et al., 2012). For shotgun metagenomic sequencing, another strong limitation is the large number of contigs and singletons which have to be annotated *de novo*. This has an important impact on computing time which is likely to increase as sequence databases continue their exponential growth. The use of dedicated databases or of subsections of GenBank partially solves this problem while creating other potential misidentification pitfalls. The other problematic situation concerns the ability (or inability) to identify sequences from a virus that has no close homolog in databases. This situation is far from theoretical since, for example, a large fraction of the genes identified in marine viral metagenomes, in particular when it comes from phages or Archeae have no clear counterpart in GenBank. The recent discovery of Pandora virus, containing only 7% of the genes homologous to existing genes, is another striking example (Philippe et al., 2013). In such situations, other strategies that do not rely on homology, such as frequency analysis techniques (Palacios et al., 2010; Trifonov and Rabadan, 2010), may possibly offer an interesting option but need additional evaluation and testing.

The (meta)genome of a sample can be characterized by identifying its genomic variations or specificities as compared to a reference genome or gene catalog. For genome sequencing projects, these variations correspond to gene content, single nucleotide polymorphisms (SNP), insertion and deletion (indel), copy number variation or chromosomal re-arrangement. For metagenome sequencing projects, the characterization is also focused on defining and quantifying clusters of genes or functions or taxons sequenced in the sample, although it should be stressed that there are currently no or very few tools allowing the simple estimation of the number of taxons for viral metagenome data (Klingenberg et al., 2013).

Importantly, the bioinformatics analysis of high-throughput sequencing data is now making a transition. Initially largely cost-prohibitive and accessible only to bioinformaticians, it is becoming more and more cost-effective and biologist-friendly, in particular with the development of commercial packages such as CLC Genomics Workbench or Geneious, or of user-friendly open platforms such a Galaxy. This transition also benefits directly from the exponential growth and cost reduction of computational power and from the availability of cloud computing. The current general trend is also to simplify the use and the parameterization of these tools, making them usable without extended bioinformatics knowledge. There are now visual and user-friendly techniques to explore and analyze the data for a biologist (Foster et al., 2012). These efforts improve and facilitate the manipulation and analysis of the generated sequences. They will further speed up the diffusion of NGS and are likely to ultimately reach the diagnostics laboratory.

3. Applications of NGS approaches in plant virus detection

Conceptually, NGS can be applied to solve problems in a range of areas of virology and, in the present case, in etiology of viral diseases, in virus characterization, taxonomy, viral population genetics and in diagnostics of plant viruses. The objective of this chapter is to underline the topics which hold interest in a future application of NGS in plant virus diagnostic. A more detailed presentation of the current applications of NGS in plant virology can be found in two recent reviews (Barba et al., 2014; Prabha et al., 2013).

Moreover, NGS can also be used to study viral populations at very different scales: from single isolated host cells to organs, host plants, plant communities or even within ecosystems, giving access to the phytoviral metagenome with, for example, application in viral ecology.

3.1. Strategies for NGS analysis of plant viruses

A specificity of NGS applications for plant virus characterization is the absence in the plants of acellular fluids such as plasma, cerebrospinal fluid or lymph that are available in animals. Complex nucleic acid (NA) populations contaminated by host molecules are therefore generally used and analyzed. One of the key parameters to consider is the identity and nature of the NA molecules that are targeted by the NGS effort. Indeed, a range of NA populations have been analyzed to date, differing in the nature (DNA or RNA, single or double stranded, with or without size selection) and in the use or not of strategies to enrich the NA populations in viral sequences.

The first option is to extract total DNA or RNA from a sample and to directly sequence it in a shotgun approach. As in other approaches, bioinformatics analyses are then used to identify viral sequences. These protocols are straightforward in the laboratory but the bioinformatics analysis is more complicated as host or microflora-derived sequences generally dominate the sequence data and must be discarded. The small proportion of the generated sequences which correspond to the pathogen can however be sufficient to identify the genomes of new viruses (Wylie and Jones, 2011). It is possible to enrich DNA populations in genomic sequences of circular DNA viruses through rolling circle amplification (Wyant et al., 2012) since small enough host-derived circular molecules are generally infrequent. Likewise, total RNA preparations can be enriched in viral sequence by purification of mRNAs (to eliminate the vast amounts of host-derived ribosomal RNA) but also by subtraction or RDA (representational differential analysis) (Scott Muerhoff et al., 1997). Such an approach has for example been used to enrich NA populations for the sequencing of a novel *Cucumovirus*, the gayfeather mild mottle virus (Adams et al., 2009).

The second option corresponds to the sequencing of NA molecules obtained from partially or completely purified viral particles. Various techniques can be used to obtain such viral particles preparations from host plants, including immune- (Wetzel et al., 1992) or print-capture (Olmos et al., 1996), simplified partial purification schemes (Muthukumar et al., 2009) or more complex purification schemes through, for example, cesium chloride gradient prior to DNA or RNA extraction. The isolated viral DNA or RNA is finally sequenced following a classical shotgun protocol. This kind of approach is technically more challenging in the laboratory and there is a risk of counter-selecting viruses with labile particles but sequences of viral origin should be enriched and may reach in some case a very high proportion of the sequencing reads (Thapa et al., 2012).

The third option relies on the isolation and subsequent high-throughput sequencing of small interfering RNAs (siRNA). siRNA are intermediate molecules as well as end products in the antiviral defense pathway called RNA interference in plants and animals. A consequence of this defense reaction is the production of 21–24 nt long siRNAs representing most if not all of the viral genomes accumulated during infection, irrespective of the genome structure of the targeted virus (Ding and Voinnet, 2007). This strategy has thus the potential to be truly polyvalent and to allow the simultaneous detection of DNA and RNA viruses as well as viroids (Itaya et al.,

2001). Given the high number of sequencing reads needed, siRNAs are generally size selected by gel purification before sequencing. The small reads are further assembled *de novo* to yield long enough contigs for homology-based annotation. Starting from the seminal work of Kreuze et al. (2009), profiling of siRNAs using NGS has identified numerous plant viruses that had never been reported previously.

Another option, which is limited to the analysis of RNA viruses and viroids, is the purification of the double-stranded RNA (dsRNA) molecules that accumulate during RNA viruses replication. These molecules can readily be purified through their selective affinity for cellulose under appropriate conditions, providing for an important enrichment in viral sequences (Dodds et al., 1984; Roossinck et al., 2010).

The last option represents a targeted approach based on *a priori* and involves the use of primers targeting a specific genomic region of the virus under study. The primers are used during a (RT)-PCR reaction to produce specific amplicons which can then be sequenced at very high depth to analyze the structure of the viral population (Fabre et al., 2012). This approach is limited to population genetics studies at various ecological scales (plant, field, region) but is not suitable to identify novel viruses unless primers with broad polyvalence are used (James et al., 2006).

3.2. Application of NGS for genome characterization

3.2.1. Whole genome characterization

During the last two years, there has been an exponential growth in publications describing the determination, using NGS approaches, of the genome sequences of novel viruses or of new genome sequences for known viruses. Currently more than fifty peer-reviewed papers describing the identification of new plant viruses have thus been published (Barba et al., 2014). Briefly, these results have been obtained in a very wide range of crop species, including temperate, Mediterranean or tropical species, herbaceous or woody species, annual or perennial crops. Among the described protocols, the siRNA sequencing approach is by far the most popular one. Interestingly, in case of mixed infection, the NGS approaches have proven able to detect and characterize all the viral genomes present in the samples, whether they belong to different strains of a viral species (Li et al., 2012) or to different species (Alabi et al., 2012; Li et al., 2012; Sela et al., 2013).

3.2.2. Viral population genetics

Viruses have very high mutation rates, short generation times, and large population sizes (Duffy et al., 2008). Under these conditions, genetic variants are produced constantly and in an infected host the virus population can display a high degree of genetic diversity. Because of their diversity, intra-host virus populations are often referred to as mutant clouds, swarms, or viral quasi-species (Beerenwinkel et al., 2012). Given the sequencing depth that they achieve, the use of NGS technologies offers unprecedented views at the intra-isolate or intra-host diversity of viruses, allowing to better assess the impact of various evolutionary forces on viral diversity (Fabre et al., 2012). For example, in a pioneering paper in plant virology (Simmons et al., 2012), the high-throughput sequencing of 5 regions of the zucchini yellow mosaic virus genome in a wild gourd, *Cucurbita pepo* ssp. *texana*, was carried out under two inoculation conditions: aphid vectored and mechanically inoculated. The results suggest that the vector or host-imposed bottlenecks (associated respectively with transmission and systemic infection) are less stringent than previously supposed. These results also showed that some mutations were fixed in the aphid-vectored populations but remained at low frequency in the mechanically inoculated plants.

3.2.3. Plant virus ecology

Deep sequencing technologies have now been applied to the non-targeted discovery of viruses in a wide range of environments, a process called viral metagenomics or viromics. These approaches have so far been applied only in a limited fashion to plant viruses, possibly because of the perceived limited importance of the free phase (*i.e.* outside of host plants and vectors) of plant viruses and of the limited information available on this phase. Remarkably, although not intentionally targeting plant viruses, several studies primarily targeting phages or animal/human viruses have identified a range of known or of novel plant viruses in diverse environments, including continental fresh water (Djikeng et al., 2009), raw sewage (Cantalupo et al., 2011), reclaimed water (Rosario et al., 2009), human or rodent feces (Nakamura et al., 2009; Victoria et al., 2009; Zhang et al., 2006), bat guano (Li et al., 2010) or even the near-surface atmosphere (Whon et al., 2012). There have been few efforts to study the phytoviral metagenome directly from plant populations but the few publications available illustrate, as expected, a wide diversity of viruses associated with plants. A large fraction of these viruses appear to be novel, in particular when it comes to double-stranded RNA viruses (Coetzee et al., 2010; Giampetruzzi et al., 2012; Min et al., 2012; Muthukumar et al., 2009; Quito-Avila et al., 2011; Roossinck et al., 2010; Scheets, 2013; Thapa et al., 2012; Wren et al., 2006; Wylie et al., 2013), with important implications for the way we envision and develop research efforts in plant virus ecology and diagnostic (Malmstrom et al., 2011; Roossinck, 2011, 2012).

4. View on the future developments and bottlenecks

Until very recently the prospect of using NGS-based approaches for plant virus diagnostics, potentially down to plant clinics level would have seemed extremely remote if not improbable, in particular due to time constraints and to the costs involved. Recent developments in sequencing machines (see above) have drastically changed the situation. Moreover, sequencing machine is more and more coupled with robots allowing automation and high throughput sample processing. The ability to perform a general viral and viroid indexing on a plant sample, including the detection of distant variants of known viruses or of unknown or novel agents is an extremely attractive possibility in a range of diagnostic situations (quarantined plant material, certification programs, quality assurance of seed lots. . .). It should be stressed that no other technique currently has the potential to deliver such broad-spectrum diagnostics. Given the pace of research and the above mentioned developments in sequencing technologies it is clear that the use of these approaches will increase in the coming years, including in diagnostic. In a pioneering experiment, a virus diagnostic protocol based on high-throughput sequencing was tested (Hagen et al., 2012), eliminating the need to develop and apply targeted real-time PCR protocols or of antibody development for ELISA. The sensitivity of the technique was similar to PCR. This breakthrough concept, although limited to dsDNA viruses, combined amplification of circular viral DNA (without previous knowledge), high-throughput sequencing of the amplified DNA and bioinformatic analysis of the obtained sequence data

Besides technical and budgetary aspects, there are however today a range of issues that need to be addressed and solved before a wide acceptance of these approaches in the plant virus diagnostics field. In a sense, this situation is not novel, since essentially the same questions have to be asked for each newly introduced diagnostic technique.

A first question that has to be considered is the sequencing strategy used (*i.e.* which target, which sequencing platform. . .) in terms of advantages, limitations and costs, and whether a truly

complete viral indexing can indeed be achieved. Currently two strategies appear to have the potential to detect all viral agents, irrespective of the nature (DNA or RNA) and structure (linear, circular, single or double-stranded) of their genome: total RNA (or mRNA) (Al Rwahnih et al., 2009) or siRNAs (Kreuze et al., 2009) sequencing. Other targets like circular DNA (Wyant et al., 2012), dsRNA sequencing sequences (Roossinck et al., 2010) or virus particle isolation (Muthukumar et al., 2009) target only a fraction of the viral diversity (in this case, respectively circular DNA viruses, RNA viruses and viroids, or agents with stable enough particles to withstand the purification procedure). On the other hand, both the total RNA and the siRNA approaches require a higher number of reads (and hence a higher cost per sample) than the strategies involving an enrichment step like dsRNA sequencing or virus particle isolation as indicated by the few direct comparisons available (Adams et al., 2009). Moreover, advances in the metagenomic field are also bringing innovative and comprehensive preparation protocols which might hold future interest in plant virus diagnostic (Roume et al., 2013). Thus, further direct comparison of these approaches is clearly needed in order for the diagnostician to select the most appropriate technique based on sound data rather than on educated guesswork. Ultimately this raises the question of deciding whether there exists an optimal and universal strategy for diagnostics purposes.

The second very practical question concerns the sensitivity of these techniques in comparison to the currently used diagnostic techniques and, in particular, to the current gold standard of real-time (RT)-PCR. Although sensitivity may not be an issue in some applications, it will clearly be a crucial question in many other diagnostics situations. Remarkably, there is extremely little information available on the sensitivity of the NGS-based approaches, in any field of virology. The results of Cheval and co-workers (Cheval et al., 2011) with human viruses suggest however that a sensitivity similar to optimized real-time PCR protocols can be reached with sufficient sequencing depth (and therefore costs) of total DNA or RNA on the Illumina platform. On the other hand, the detection of novel agents not represented in the databases necessitated *de novo* assembly of contigs, which could only be achieved at higher viral titers. Almost no corresponding data is currently available for plant viruses, with the exception of the work of Hagen (Hagen et al., 2012), who showed that siRNA sequencing allowed the detection of tomato spotted wilt virus (TSWV) in tomato plants as early as 4 days post-inoculation, a time at which no symptoms are observed and virus detection by molecular hybridization is at best very difficult. Comparative efforts, assessing the efficiency of NGS-based approaches on spiked samples close to the limits of detection of current serological (ELISA) or molecular [(RT)-PCR] approaches are therefore clearly needed, in particular with plant matrices, for the diagnostician to have a clear idea of the sensitivity of these novel technologies. It is however obvious that even with current sequencing technologies, detection of some low titer agents may rest on the identification of very few or even of a single read (Tan et al., 2013), which will clearly raise questions about detection threshold, as in any other detection technique.

Indeed, the determination of a suitable detection threshold is a key aspect in essentially all diagnostic techniques. Whether the detection of a single viral read should be considered sufficient to assess as positive a sample analyzed by deep sequencing remains to be investigated but, more importantly, debated. Given that the current NGS strategies all involve one or more PCR steps, it is easy to see that NGS-based approaches will be as susceptible to contamination problems as (notoriously) are PCR-based assays. This is in fact a frequently observed but rarely reported observation in many if not all laboratories that have invested in NGS approaches. In many cases, these problems, frequently limited to a low proportion of reads, are of little consequences but the specifics of the diagnostics field clearly give more impact to such potential contamination problems. A direct consequence is that the same precautions and standards currently implemented in (RT)-PCR-based diagnostics will almost surely need to be also implemented for diagnostics uses of NGS technologies.

Other strategies may help to resolve contamination/low representation and detection threshold issues, for example the detailed analysis of the coverage of the viral genomes by the sequences obtained, since a low but distributed coverage is likely to reflect the presence of a virus while the same number of reads mapping to a single genomic region are more likely to reflect a contamination problem. Associated with these aspects will come questions about the need to validate the NGS-based approaches and therefore about their reproducibility and repeatability both within a given laboratory and between laboratories using the same protocol (Massart et al., 2008, 2009).

Another question concerns the adaptation of the protocols and techniques to particular plant species or tissues or to particular target virus(es) or diagnostic needs. A number of plant species are notorious for containing substance that interfere with nucleic acids extraction or with their subsequent enzymatic manipulation (reverse transcription, PCR amplification). Simple and robust extraction protocols have been developed in many cases for (RT)-PCR-based assays and will likely prove appropriate for NGS-based approaches but will first need to be validated. Also, initial results in the authors laboratories indicate that sequencing depth will likely need to be adapted in some cases since, for example, the fraction of siRNAs of viral origin appears to be much lower in woody plant samples (Olmos, unpublished results) than reported in herbaceous hosts like sweet potato (Kreuze et al., 2009).

Lastly, if these NGS-based approaches are to take their place in diagnostics laboratories, there is a clear need for the further development of simple, reliable and user-friendly bioinformatic tools that will allow the rapid and simple identification of the viruses present in a sample. Although much progress in that direction has been made, there is still much room for improvement and for the development of faster and more efficient programs and algorithms. Of particular concern, if complete viral indexing is to be achieved, is the question of the identification of viruses too divergent to be identified by the currently used homology-based approaches. Homology-independent approaches can be envisioned in some cases, but with exceptions (Wu et al., 2012) are still in their infancy.

In conclusion, these novel NGS-based approaches have a huge potential in the diagnostics field. This potential is already amply illustrated by the monthly harvest of newly discovered and described human, animal and plant viruses published in the scientific literature. Other more practical and diagnostics-oriented uses of these novel technologies will likely take more time to develop and spread and will require specific efforts for the optimization, comparison and validation of these approaches before they can replace for specific applications the currently validated and used diagnostic techniques. The pace of adoption of these technologies will also vary depending on the origin of the sample: it is likely that NGS will be applied first in the certification of high-value plant material, as mother trees for *in vitro* multiplication. The new and currently used techniques should however not be seen as competitors, the development of one signaling the demise of the other. Rather, as has almost always been the case in the diagnostics field, these techniques will continue parallel lives, the diagnostician selecting to use one or the other, relying on a cost-benefit analysis to decide which technique best suits his precise needs.

Another emerging question, which will become more acute as more results accumulate, is the biological significance and the implications of the identification of a novel virus. This question is particularly acute in the case of well characterized or high value

plant materials such as mother plants or 'top tier' plants in clean plant programs that serve or have served as the source of plants for international trade. The finding of a novel virus in such plant material raises very practical questions that the diagnostician will need to address, in particular the impact of the virus identified on the plant concerned and, ultimately, the potential need to halt the distribution of the plant(s) concerned (MacDiarmid et al., 2013). At the same time, we must acknowledge the fact that very little information will be available, besides the complete or even partial sequence of the novel agent, to assist in reaching a decision. It is already clear that some of the agents identified will not be plant viruses, but will in fact be mycoviruses infecting fungal agents associated with the analyzed plant. Mycoviruses within plants are associated with the presence of plant pathogens or endophytes and are poorly studied so far (Pearson et al., 2009). In grasses, it has been demonstrated that mycovirus incidence reached 22% among endophytic fungi isolates (Herrero et al., 2009). As numerous endophyte species, more than one hundred in some cases, can be associated with a plant species, mycoviruses are frequently detected and can represent more than 50% of the viral sequences generated in a NGS experiment (Al Rwahnih et al., 2011). Any new viral sequence discovered must therefore be carefully analyzed, for example through blastn or blastx algorithms, to try to determinate its plant or fungal origin. In some case, additional experiments may be needed to determine the host of the virus discovered, for example trying to isolate (Al Rwahnih et al., 2011) or detect fungi from the studied plant.

We can also reasonably anticipate that the pace of discovery of novel viral agents, either mycovirus or plant virus, will slow down in the future. Indeed, the virome of crop species will progressively be better characterized through the increasing number of NGS experiments carried out. As the viral sequence databases become more and more complete, the discovery of new viral agents during a certification process will become less frequent but the detection of poorly characterized viruses could remain frequent. This strengthens the importance to also develop efforts aimed at the understanding the biological significance of all these newly characterized virus species, in order to be able to make the most relevant diagnostic decisions in the future. However, it should be stressed that although the discovery of such novel or poorly characterized agents may today complicate the decision-making process, it will clearly lead, in the long run, to an improved situation of better informed decisions and to the safer movement and trade of plant propagation material.

A side question raised about the rapid pace of discovery of novel viral agents for which very little if any, biological data is available is that of the impact on viral taxonomy. Indeed, the Executive Committee of the International Committee for the Taxonomy of Viruses (ICTV) recently agreed to accept species or higher taxa proposals based only on sequence data from (meta)genomic studies, with certain safeguards. These include evidence that the sequences are effectively complete, that correct assembly has been verified and that the sequence is indeed viral in origin (Gorbalenya, 2014). It is therefore probably just a matter of time before the first viral species described solely on the basis of sequence data become accepted at the international level.

It should also be stressed that with their remarkable potential when it comes to viral genome characterization, NGS-based techniques are likely to have a large and long-lasting impact on (RT)-PCR-based assays, since the availability of a wider diversity of genomic sequences will allow us to develop more broad-specificity primers, allowing the reliable detection of all isolates of a given virus. Such a situation is for example illustrated by the recent identification by NGS of isolates of plum bark necrosis stem pitting associated virus (PBNSPaV) that are not or only poorly detected by currently used RT-PCR assays and by the use of the sequence of such isolates to design new, more polyvalent PBNSPaV-specific PCR primers (Marais et al., 2014).

References

Adams, I.P., Glover, R.H., Monger, W.A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M., Boonham, N., 2009. Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol. Plant Pathol. 10 (4), 537–545.

Al Rwahnih, M., Daubert, S., Golino, D., Rowhani, A., 2009. Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. Virology 387 (2), 395–401.

Al Rwahnih, M., Daubert, S., Urbez-Torres, J.R., Cordero, F., Rowhani, A., 2011. Deep sequencing evidence from single grapevine plants reveals a virome dominated by mycoviruses. Arch. Virol. 156 (3), 397–403.

Alabi, O.J., Zheng, Y., Jagadeeswaran, G., Sunkar, R., Naidu, R.A., 2012. High-throughput sequence analysis of small RNAs in grapevine (*Vitis vinifera* L.) affected by grapevine leafroll disease. Mol. Plant Pathol. 13 (9), 1060–1076.

Barba, M., Czosnek, H., Hadidi, A., 2014. Historical perspective, development and applications of next-generation sequencing in plant virology. Viruses 6 (1), 106–136.

Beerenwinkel, N., Gunthard, H.F., Roth, V., Metzner, K.J., 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbiol. 3, 329.

Cantalupo, P.G., Calgua, B., Zhao, G., Hundesa, A., Wier, A.D., Katz, J.P., Grabe, M., Hendrix, R.W., Girones, R., Wang, D., Pipas, J.M., 2011. Raw sewage harbors diverse viral populations. mBio 2 (5).

Cheval, J., Sauvage, V., Frangeul, L., Dacheux, L., Guigon, G., Dumey, N., Pariente, K., Rousseaux, C., Dorange, F., Berthet, N., Brisse, S., Moszer, I., Bourhy, H., Manuguerra, C.J., Lecuit, M., Burguiere, A., Caro, V., Eloit, M., 2011. Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. J. Clin. Microbiol. 49 (9), 3268–3275.

Coetzee, B., Freeborough, M.J., Maree, H.J., Celton, J.M., Rees, D.J.G., Burger, J.T., 2010. Deep sequencing analysis of viruses infecting grapevines: virome of a vineyard. Virology 400 (2), 157–163.

Ding, S.W., Voinnet, O., 2007. Antiviral immunity directed by small RNAs. Cell 130 (3), 413–426.

Djikeng, A., Kuzmickas, R., Anderson, N.G., Spiro, D.J., 2009. Metagenomic analysis of RNA viruses in a fresh water lake. PLoS ONE 4 (9).

Dodds, J.A., Morris, T.J., Jordan, R.L., 1984. Plant viral double stranded RNA. Annu. Rev. Phytopathol. 22, 18.

Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9 (4), 267–276.

Fabre, F., Montarry, J., Coville, J., Senoussi, R., Simon, V., Moury, B., 2012. Modelling the evolutionary dynamics of viruses within their hosts: a case study using high-throughput sequencing. PLoS Pathog. 8 (4), e1002654.

Foster, J.A., Bunge, J., Gilbert, J.A., Moore, J.H., 2012. Measuring the microbiome: perspectives on advances in DNA-based techniques for exploring microbial life. Brief. Bioinform. 13 (4), 420–429.

Giampetruzzi, A., Roumi, V., Roberto, R., Malossini, U., Yoshikawa, N., La Notte, P., Terlizzi, F., Credi, R., Saldarelli, P., 2012. A new grapevine virus discovered by deep sequencing of virus- and viroid-derived small RNAs in Cv *Pinot gris*. Virus Res. 163 (1), 262–268.

Gorbalenya, A.E., 2014. Taxonomic proposals based on metagenomic and genome-only studies. ICTV Newslett. 11.

Hagen, C., Frizzi, A., Gabriels, S., Huang, M., Salati, R., Gabor, B., Huang, S., 2012. Accurate and sensitive diagnosis of geminiviruses through enrichment, high-throughput sequencing and automated sequence identification. Arch. Virol. 157 (5), 907–915.

Hamilton, J.P., Neeno-Eckwall, E.C., Adhikari, B.N., Perna, N.T., Tisserat, N., Leach, J.E., Levesque, C.A., Buell, C.R., 2011. The Comprehensive Phytopathogen Genomics Resource: a web-based resource for data-mining plant pathogen genomes. Database (Oxford) 2011, bar053.

Herrero, N., Sánchez Márquez, S., Zabalgogeazcoa, I., 2009. Mycoviruses are common among different species of endophytic fungi of grasses. Arch. Virol. 154 (2), 327–330.

Itaya, A., Folimonov, A., Matsuda, Y., Nelson, R.S., Ding, B., 2001. Potato spindle tuber viroid as inducer of RNA silencing in infected tomato. Mol. Plant Microbe Interact. 14 (11), 1332–1334.

James, D., Varga, A., Pallas, V., Candresse, T., 2006. Strategies for simultaneous detection of multiple plant viruses. Can. J. Plant Pathol. 28 (1), 16–29.

Kent, W.J., 2002. BLAT – the BLAST-like alignment tool. Genome Res. 12 (4), 656–664.

Klingenberg, H., Aßhauer, K.P., Lingner, T., Meinicke, P., 2013. Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. Bioinformatics 29 (8), 973–980.

Kreuze, J.F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., Simon, R., 2009. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388 (1), 1–7.

Li, L., Victoria, J.G., Wang, C., Jones, M., Fellers, G.M., Kunz, T.H., Delwart, E., 2010. Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. J. Virol. 84 (14), 6955–6965.

Li, R., Gao, S., Hernandez, A.G., Wechter, W.P., Fei, Z., Ling, K.S., 2012. Deep sequencing of small RNAs in tomato for virus and viroid identification and strain differentiation. PLOS ONE 7 (5).

López, M.M., Llop, P., Olmos, A., Marco-Noales, E., Cambra, M., Bertolini, E., 2009. Are molecular tools solving the challenges posed by detection of plant pathogenic bacteria and viruses? Curr. Issues Mol. Biol. 11, 13–46.

MacDiarmid, R., Rodoni, B., Melcher, U., Ochoa-Corona, F., Roossinck, M., 2013. Biosecurity implications of new technology and discovery in plant virus research. PLoS Pathog. 9 (8), e1003337, http://dx.doi.org/10.1371/journal.ppat.1003337.

Malmstrom, C.M., Melcher, U., Bosque-Pérez, N.A., 2011. The expanding field of plant virus ecology: historical foundations, knowledge gaps, and research directions. Virus Res. 159 (2), 84–94.

Marais, A., Faure, C., Couture, C., Bergey, B., Gentit, P., Candresse, T., 2014. Characterization by deep sequencing of divergent Plum bark necrosis stem pitting associated virus isolates and development of a polyvalent PBNSPaV-specific detection assay. Phytopathology, http://dx.doi.org/10.1094/PHYTO-08-13-0229-R.

Martin, R.R., James, D., Levesque, C.A., 2000. Impacts of molecular diagnostic technologies on plant disease management. Annu. Rev. Phytopathol. 38, 207–239.

Massart, S., Brostaux, Y., Barbarossa, L., César, V., Cieslinska, M., Dutrecq, O., Fonseca, F., Guillem, R., Laviña, A., Olmos, A., Steyer, S., Wetzel, T., Kummert, J., Jijakli, M.H., 2008. Inter-laboratory evaluation of a duplex RT-PCR method using crude extracts for the simultaneous detection of Prune dwarf virus and Prunus necrotic ringspot virus. Eur. J. Plant Pathol. 122 (4), 539–547.

Massart, S., Brostaux, Y., Barbarossa, L., Batlle, A., Cesar, V., Dutrecq, O., Fonseca, F., Guillem, R., Komorowska, B., Olmos, A., Steyer, S., Wetzel, T., Kummert, J., Jijakli, M.H., 2009. Interlaboratory evaluation of two reverse-transcriptase polymeric chain reaction-based methods for detection of four fruit tree viruses. Ann. Appl. Biol. 154 (1), 133–141.

Min, B.E., Feldman, T.S., Ali, A., Wiley, G., Muthukumar, V., Roe, B.A., Roossinck, M., Melcher, U., Palmer, M.W., Nelson, R.S., 2012. Molecular characterization, ecology, and epidemiology of a novel tymovirus in *Asclepias viridis* from Oklahoma. Phytopathology 102 (2), 166–176.

Minot, S., Grunberg, S., Wu, G.D., Lewis, J.D., Bushman, F.D., 2012. Hypervariable loci in the human gut virome. Proc. Natl. Acad. Sci. U. S. A. 109 (10), 3962–3966.

Mumford, R., Boonham, N., Tomlinson, J., Barker, I., 2006. Advances in molecular phytodiagnostics – new solutions for old problems. Eur. J. Plant Pathol. 116 (1), 1–19.

Muthukumar, V., Melcher, U., Pierce, M., Wiley, G.B., Roe, B.A., Palmer, M.W., Thapa, V., Ali, A., Ding, T., 2009. Non-cultivated plants of the Tallgrass Prairie Preserve of northeastern Oklahoma frequently contain virus-like sequences in particulate fractions. Virus Res. 141 (2), 169–173.

Nakamura, S., Yang, C.S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., Goto, N., Takahashi, K., Yasunaga, T., Ikuta, K., Mizutani, T., Okamoto, Y., Tagami, M., Morita, R., Maeda, N., Kawai, J., Hayashizaki, Y., Nagai, Y., Horii, T., Iida, T., Nakaya, T., 2009. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PLoS ONE 4 (1).

Olmos, A., Dasí, M.A., Candresse, T., Cambra, M., 1996. Print-capture PCR: a simple and highly sensitive method for the detection of Plum pox virus (PPV) in plant tissues. Nucleic Acids Res. 24 (11), 2192–2193.

Palacios, G., Lovoll, M., Tengs, T., Hornig, M., Hutchison, S., Hui, J., Kongtorp, R.T., Savji, N., Bussetti, A.V., Solovyov, A., Kristoffersen, A.B., Celone, C., Street, C., Trifonov, V., Hirschberg, D.L., Rabadan, R., Egholm, M., Rimstad, E., Lipkin, W.I., 2010. Heart and skeletal muscle inflammation of farmed salmon is associated with infection with a novel reovirus. PLoS ONE 5 (7).

Pearson, M.N., Beever, R.E., Boine, B., Arthur, K., 2009. Mycoviruses of filamentous fungi and their relevance to plant pathology. Mol. Plant Pathol. 10 (1), 115–128.

Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Garin, J., Claverie, J.M., Abergel, C., 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science 341 (6143), 281–286.

Prabha, K., Baranwal, V.K., Jain, R.K., 2013. Applications of next generation high throughput sequencing technologies in characterization, discovery and molecular interaction of plant viruses. Indian J. Virol. 24 (2), 157–165.

Quito-Avila, D.F., Jelkmann, W., Tzanetakis, I.E., Keller, K., Martin, R.R., 2011. Complete sequence and genetic characterization of Raspberry latent virus, a novel member of the family Reoviridae. Virus Res. 155 (2), 397–405.

Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F., Gordon, J.I., 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466 (7304), 334–338.

Roossinck, M.J., 2011. The big unknown: plant virus biodiversity. Curr. Opin. Virol. 1 (1), 63–67.

Roossinck, M.J., 2012. Plant virus metagenomics: biodiversity and ecology. Annu. Rev. Genet. 46, 359–369.

Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., Lai, H., Chavarría, F., Shen, G., Roe, B.A., 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol. Ecol. 19 (Suppl. 1), 81–88.

Rosario, K., Nilsson, C., Lim, Y.W., Ruan, Y., Breitbart, M., 2009. Metagenomic analysis of viruses in reclaimed water. Environ. Microbiol. 11 (11), 2806–2820.

Roume, H., Heintz-Buschart, A., Muller, E.E.L., Wilmes, P., 2013. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. Methods Enzymol. 531, 219–236.

Scheets, K., 2013. Infectious transcripts of an asymptomatic panicovirus identified from a metagenomic survey. Virus Res. 176 (1–2), 161–168.

Scott Muerhoff, A., Leary, T.P., Desai, S.M., Mushahwar, I.K., 1997. Amplification and subtraction methods and their application to the discovery of novel human viruses. J. Med. Virol. 53 (1), 96–103.

Sela, N., Lachman, O., Reingold, V., Dombrovsky, A., 2013. A new cryptic virus belonging to the family Partitiviridae was found in watermelon co-infected with Melon necrotic spot virus. Virus Genes 47 (2), 382–384.

Shokralla, S., Spall, J., Gibson, J., Hajibabaei, M., 2012. Next-generation sequencing technologies for environmental DNA research. Mol. Ecol. 21, 11.

Simmons, H.E., Dunham, J.P., Stack, J.C., Dickins, B.J., Pagan, I., Holmes, E.C., Stephenson, A.G., 2012. Deep sequencing reveals persistence of intra- and inter-host genetic diversity in natural and greenhouse populations of zucchini yellow mosaic virus. J. Gen. Virol. 93 (Pt 8), 1831–1840.

Tan, L.V., van Doorn, H.R., Nghia, H.D.T., Chau, T.T.H., Tu, L.T.P., de Vries, M., Canuti, M., Deijs, M., Jebbink, M.F., Baker, S., Bryant, J.E., Tham, N.T., Bkrong, N.T.T.C., Boni, M.F., Loi, T.Q., Phuong, L.T., Verhoeven, J.T.P., Crusat, M., Jeeninga, R.E., Schultsz, C., Chau, N.V.V., Hien, T.T., van der Hoek, L., Farrar, J., de Jong, M.D., 2013. Identification of a new cyclovirus in cerebrospinal fluid of patients with acute central nervous system infections. mBio 4 (3).

Thapa, V., Melcher, U., Wiley, G.B., Doust, A., Palmer, M.W., Roewe, K., Roe, B.A., Shen, G., Roossinck, M.J., Wang, Y.M., Kamath, N., 2012. Detection of members of the Secoviridae in the Tallgrass Prairie Preserve, Osage County, Oklahoma, USA. Virus Res. 167 (1), 34–42.

Trifonov, V., Rabadan, R., 2010. Frequency analysis techniques for identification of viral genetic data. mBio 1 (3).

Victoria, J.G., Kapoor, A., Li, L., Blinkova, O., Slikas, B., Wang, C., Naeem, A., Zaidi, S., Delwart, E., 2009. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J. Virol. 83 (9), 4642–4651.

Wetzel, T., Candresse, T., Ravelonandro, M., Dunez, J., 1991. A polymerase chain reaction assay adapted to plum pox potyvirus detection. J. Virol. Methods 33 (3), 355–365.

Wetzel, T., Candresse, T., Macquaire, G., Ravelonandro, M., Dunez, J., 1992. A highly sensitive immunocapture polymerase chain reaction method for plum pox potyvirus detection. J. Virol. Methods 39 (1–2), 27–37.

Whon, T.W., Kim, M.S., Roh, S.W., Shin, N.R., Lee, H.W., Bae, J.W., 2012. Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. J. Virol. 86 (15), 8221–8231.

Williamson, S.J., Rusch, D.B., Yooseph, S., Halpern, A.L., Heidelberg, K.B., Glass, J.I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C.S., Sutton, G., Frazier, M., Venter, J.C., 2008. The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS ONE 3 (1), e1456.

Wren, J.D., Roossinck, M.J., Nelson, R.S., Scheets, K., Palmer, M.W., Melcher, U., 2006. Plant virus biodiversity and ecology. PLoS Biol. 4 (3).

Wu, Q., Wang, Y., Cao, M., Pantaleo, V., Burgyan, J., Li, W.X., Ding, S.W., 2012. Homology-independent discovery of replicating pathogenic circular RNAs by deep sequencing and a new computational algorithm. Proc. Natl. Acad. Sci. U. S. A. 109 (10), 3938–3943.

Wyant, P.S., Strohmeier, S., Schäfer, B., Krenz, B., Assunção, I.P., Lima, G.S.D.A., Jeske, H., 2012. Circular DNA genomics (circomics) exemplified for geminiviruses in bean crops and weeds of northeastern Brazil. Virology 427 (2), 151–157.

Wylie, S.J., Jones, M.G., 2011. The complete genome sequence of a Passion fruit woodiness virus isolate from Australia determined using deep sequencing, and its relationship to other potyviruses. Arch. Virol. 156 (3), 479–482.

Wylie, S.J., Li, H., Dixon, K.W., Richards, H., Jones, M.G.K., 2013. Exotic and indigenous viruses infect wild populations and captive collections of temperate terrestrial orchids (Diuris species) in Australia. Virus Res. 171 (1), 22–32.

Zhang, T., Breitbart, M., Lee, W.H., Run, J.Q., Wei, C.L., Soh, S.W.L., Hibberd, M.L., Liu, E.T., Rohwer, F., Ruan, Y., 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol. 4 (1), 0108–0118.