JARQ 46 (2), 123 – 128 (2012) http://www.jircas.affrc.go.jp

REVIEW
Sequencing-based Virus Hunting and Virus Detection

Kappei KOBAYASHI[1]*, Go ATSUMI[2], Naoto YAMAOKA[1] and Ken-Taro SEKINE[2]

[1] Faculty of Agriculture, Ehime University (Matsuyama, Ehime 790–8566, Japan)
[2] Iwate Biotechnology Research Center (Kitakami, Iwate 024–0003, Japan)

Abstract
Next-generation sequencers have accelerated the advancement of virus hunting or virus detection by metagenomic analysis. Conversely, earlier works enabled virus detection with classic Sanger sequencing by elaborating sample preparation/processing techniques. In this review, we introduce virus hunting technologies both with and without cutting-edge sequencing technology and subsequently discuss the possibility that combining these technologies may extend the application of sequencing-based virus detection from scientific research to everyday diagnostics. Finally, we offer an outlook for another type of technology that leads to pathogen identification using the sequence data obtained in sequencing-based virus hunting.

Discipline: Plant disease
Additional key words: infectious molecular clone, next-generation sequencer, nucleic acid extraction, pathogen identification

Introduction

Isolation and identification, or diagnosis, of pathogenic viruses have been laborious, time-consuming and sometimes unsuccessful in both animal and plant systems. Early virus isolation relied on the virus being able to infect its original and laboratory hosts. Although some viruses are readily infectious to laboratory hosts, for others it is very difficult or impossible to reproduce the infection artificially. Earlier breakthroughs in virus detection include the development of the electron microscope, the use of antibodies, the use of tissue cultures in animal systems, and the use of genetic testing, i.e. nucleic acid hybridization, PCR and so on. Despite the development of numerous technologies for virus detection, we still often encounter new viruses which take a long time to identify. Therefore, universal virus detection systems have been eagerly anticipated.

Recent progress in genome science has changed our view of life and spawned new technologies for characterizing a wide variety of organisms. What we call next-generation sequencers (NGS) has enabled us to obtain a tremendous amount of sequence data, and made us more aware of a new biological technique called metagenome analysis. As NGS became commonly used, many virologists started using it to detect viruses from metagenomic samples, resulting in the discovery of several new viruses and the advancement of successful virus hunting [13]. Although NGS contributed significantly to technical progress in virus detection, earlier studies successfully identified some new viruses using less efficient Sanger sequencing combined with specialized sample preparation/processing techniques (see below).

In this review, we first introduce some recent reports describing virus detection using NGS, focusing on the comparison of techniques rather than the findings of analyses (Table 1). Next, we mention earlier studies with special emphasis on the sample preparation/processing techniques (Table 1). Subsequently, we discuss the possible application of sequencing-based virus detection; not only to scientific research or hygienic surveillance but also to everyday diagnostics of diseases of plants, animals and humans. Finally, we discuss the weak point of sequencing-based virus detection technology: namely, the fact that the discovery of a virus sequence does not always result in the identification of the virus responsible for the disease of interest. We provide a tip to overcome this problem.

*Corresponding author: kappei@agr.ehime-u.ac.jp
Received 30 March 2011; accepted 21 July 2001.

Power of metagenomics using next-generation sequencers in virus hunting

There are several NGS available, which have already been reviewed in depth[19]. Therefore, we have not gone into detail as regards their principles and performance. Briefly, NGS reads short sequences in a massively parallel manner: the nucleotides per single read ranging from 35 to 500, and the reads per single run ranging from one million to 600 million, resulting in the generation of sequence data of 500 mega-bases to 95 giga-bases (the throughput is changing rapidly; see manufacturers' web sites: http://www.454.com/, http://www.illumina.com/systems.ilmn and http://www.appliedbiosystems.com/absite/us/en/home.html).

One of the earliest studies to use NGS to hunt pathogens was by Cox-Foster et al. (2007), who explored the causal agent(s) of honeybee colony collapse disorder (CCD)[5]. They analyzed RNA from CCD- and non-CCD-bees by sequencing randomly transcribed and amplified cDNA fragments and detected several candidates for the CCD pathogen: bacteria, fungi and viruses. Although they did not focus on viruses, they found several novel viruses, indicating that the approach is quite powerful in virus hunting (Table 1). This technology has also been applied to the analysis of clinical samples. Nakamura et al. (2009) analyzed nasal and fecal samples from patients with flu and norovirus infections, respectively[21]. They detected some viruses, such as endogenous retroviruses and plant viruses of food origin, in addition to the influenza virus and norovirus. In this study, 90% or more of the sequence reads from nasal samples were of eukaryotic origin, indicating the need to remove cellular materials from nasopharyngeal aspirates. Meanwhile however, the results also indicate the power of NGS, because the influenza virus was readily detected in those analyses (Table 1). Similar experiments were also carried out in plant systems. Adams et al. (2009) analyzed total RNA from tomatoes infected with the *Pepino mosaic virus* (PepMV) and *Gomphrena globosa*, a frequently-used tester plant, infected with an unknown pathogen[1]. They found that, of a 16.6 mega-base sequence, 20% was from PepMV whereas 70% was of host origin. In *G. globosa*, less than 50% was of host origin and 40% was from a new virus belonging to the genus Cucumovirus. These results indicate that, unlike animal viruses, it is readily possible to detect some plant viruses, which replicate to high levels, by total RNA sequencing (Table 1).

Nakamura et al. (2009) successfully removed host cellular materials and enriched viral nucleic acids by clearing the stool suspension in a high-speed centrifuge[21]. Another strategy to enrich viral RNA was adopted by Coetzee et al. (2010; Table 1). They enriched double-stranded RNA (dsRNA) using a classic technique and sequenced the entire dsRNA fraction using an NGS[4]. Although they pooled samples from 44 grapevine plants, they detected a number of grapevine-infecting viruses as well as putative fungal viruses, reaffirming the power of

Table 1. Typical methods for sequencing-based virus detection

Reference	Sample preparation[a]	Amplification[b]	Sequencing[c]
Mizutani et al. (2007)	RNA from nuclease-treated tissue culture	WGA	Selective PCR & Sanger
Yamao et al. (2009)	RNA from nuclease-treated tissue culture	Phi29	Selective PCR & Sanger
Finkbeiner et al. (2008)	RNA from filtrated stool aqueous extracts.	Wang-2	Cloning & Sanger
Nakamura et al. (2009)	Nasal sample total RNA RNA from centrifuged stool aqueous extracts	WTA	454
Cox-Foster et al. (2007)	Honeybee total RNA	Wang-1	454
Adams et al. (2009)	Total RNA from infected tester plants	Wang-1	454
Kreuze et al. (2009)	Small RNA of plants	Seq-Protocol	Illumina
Wu et al. (2010)	Small RNA of invertebrates	Seq-Protocol	Illumina
Coetzee et al. (2010)	dsRNA isolated using CF-11	Seq-Protocol	Illumina
Kobayashi et al. (2009)	dsRNA isolated using recombinant dsRNA-binding protein	WTA	Cloning & Sanger

[a] Methods to prepare viral nucleic acids. dsRNA, double-stranded RNA; CF-11, CF-11 cellulose.
[b] Methods to amplify viral nucleic acids. WGA, whole genome amplification kit (Sigma); Phi29, Phi29 DNA polymerase; WTA, whole transcriptome amplification kit (Sigma); Wang-1 and Wang-2, methods reported by Wang et al. (2002; ref 29) and (2003; ref 30).
[c] Methods for sequencing. Sanger, chain-terminator method; 454, NGS of 454 Life Science (Roch; http://roche-biochem.jp/products/454sequence/); Illumina, NGS of Illumina (http://www.illuminakk.co.jp/).

NGS.

Two studies independently demonstrated that the sequencing analysis of small interfering RNA (siRNA) was useful for detecting viruses in plants and invertebrates (insects and nematodes)[12, 31]. Although not applicable to vertebrates, who can establish strong adaptive immunity to invading viruses, it is well known that RNA silencing is a major antiviral defense mechanism in plants and invertebrates. Kreuze et al. (2009) analyzed small RNA profiles in severe synergistic sweet potato viral disease and, by chance, found two novel DNA viruses belonging to the genus *Badnavirus* and the genus *Mastrevirus*[12], reporting that 30,000 reads of 22 nucleotide siRNA could lead to a reliable diagnosis. Based on this finding, they claimed that 100 samples could be analyzed by a single run of the Illumina genome analyzer, which could give rise to 3–4 million reads/run. However, 22 nt siRNA is not always the most abundant molecular species of small RNA in plant, fungal and invertebrate cells. In *Caenorhabditis elegans*, carrying the *Flock house virus* genome, Wu et al. showed that 0.6% of the small RNA was of viral origin, with 23 nt molecular species as the most abundant viral siRNA[31]. They mapped 6% of small RNA reads (13.4% of assembled contigs) from *Drosophila* S2 cells to five viruses, four of which were newly identified. Although the viral siRNAs constitute a somewhat large population, they do not represent the majority of small RNA populations: e.g. in S2 cells, 62% of assembled contigs were mapped to transposons. Therefore, successful virus detection by the analysis of small RNA would rely on the power of NGS (Table 1).

Sample preparation and processing in virus hunting with Sanger sequencing

Although NGS is a powerful tool in genomics and related research fields, its use is not easy: the machines are expensive, with high running costs, and the analysis of huge nucleotide sequence data requires high performance computers and informatics skills, hence NGS is only accessible to a limited number of scientists. The need for universal virus detection methods pushed virologists in another direction to establish more efficient methods for the same.

Mizutani et al. (2007) reported a sample processing method, which they called rapid determination of viral RNA sequence[20]. They treated tissue culture supernatant with DNase and RNase to minimize the co-extraction of cellular materials with an encapsidated viral genome. They then extracted RNA and amplified cDNA using a commercial whole genome amplification kit. The ingenious procedure they established is cloning-free sequencing, which was enabled by an adopter-mediated selective PCR (Table 1). Finkbeiner et al. (2008) adopted a similar strategy to analyze diarrhea samples but cloned their cDNAs[7], because Sanger sequencing requires the separation of each cDNA from the library and cloning using a plasmid vector is the simplest way to achieve this (Table 1). The efficient protocol used by Mizutani et al. (2007; Table 1) accelerated the analysis by omitting the cloning step[20]. Another advancement achieved by the same group is the optimization of a new exhaustive amplification technique. Yamao et al. (2009) used Phi29 DNA polymerase, which exhibits strong strand displacement activity and thus amplifies the DNA sequence *in vitro* quite efficiently, to amplify a tiny amount of viral nucleic acid. Employing a ligation step with a supplemental oligonucleotide, they successfully improved the analytical sensitivity[32]. As a result, they successfully detected a number of viruses, which had previously been difficult or impossible to detect. Although the use of NGS would have eliminated the need for adopter-mediated selective PCR, the Phi29-mediated amplification technique would still be useful. Indeed, Yamao et al. (2009) also analyzed their amplification products using NGS to obtain more sequence information for a virus they newly identified[32] (Table 1).

Enrichment of viral nucleic acids would be the key factor in making the analysis more efficient. RNA viruses, regardless of genome configuration — single- or double-stranded, and positive-, negative- or ambi-sense — produce double-stranded RNA (dsRNA), either as a genome or replication intermediate. The enrichment of dsRNA has been the strategy of choice, especially for plant virologists. A classic technique for dsRNA extraction uses an ion-exchange cellulose called CF-11[28]. However, because the CF-11 technique is not efficient when nucleic acid extract is viscous and not easy for beginners, we developed a recombinant dsRNA-binding protein (DRBP) as an alternative tool[11]. The isolation of dsRNA using DRBP is easy and compatible with subsequent analyses, such as gel electrophoresis and reverse transcription. Inspired by the studies of Mizutani and colleagues, we established a protocol for universal plant virus detection and named it dsRNA-isolation, exhaustive amplification, cloning and sequencing (DECS; Table 1)[11]. Using DECS with Sanger sequencing, we detected two novel viruses from gentian and one each from *Chrysanthemum*, Japanese basil and *Eustoma* (unpublished results). The results indicate that isolation of dsRNA, regardless of the method used, is effective in enriching viral RNA sequences for sequencing-based virus detection. For DNA viruses, however, this is not effective and a novel approach to concentrate virus particles is needed.

Sequencing-based virus detection for everyday diagnostics?

As mentioned above, NGS alone is useful for virus detection and hunting. The technology would also be useful for everyday diagnostics of plants, animals and human beings, if a solution to the prohibitive cost could be found. One way to minimize the analysis cost of NGS is to increase the sample numbers analyzed in parallel. Sequencing of multiple samples in a single run has already proved feasible using kits provided by the NGS manufacturers. However, the simultaneous sequencing of multiple samples reduces the number of sequence reads from each sample, making it important to amplify and enrich viral sequences to a level detectable by sequencing, i.e. to reduce the levels of sequences from host and co-existing non-pathogenic microorganisms such that they do not push away the viral sequences from the analysis. The sample preparation/processing technology mentioned above is summarized in Fig. 1. The dotted line indicates the workflow not reported to date, while thicker lines indicate the workflows we consider suitable for sequencing-based virus diagnosis. As mentioned above, virus detection by small RNA profiling relies on NGS sequencing power, meaning the scale of analysis should not be reduced if possible. Furthermore, some viruses would code for the inhibitor of Dicer[24, 27], the enzyme responsible for the siRNA production. Therefore, it is preferable that viral genomes or their replication intermediates be enriched. Amplification using phi29 DNA polymerase was shown to be advantageous in terms of improving the detection sensitivity[32]. Recently, compact models with reduced NGS running costs have been released by a few companies (see the aforementioned manufacturers' web sites) and sequencing-based virus detection in daily diagnosis has now been realized.

From sequences to diseases: how to prove the pathogenesis of a "virus" discovered by sequencing-based methods

Sequencing-based virus detection often results in the discovery of novel viruses. For human diseases, etiological surveillance is performed to examine the relevance of the virus in the pathogenesis of interest, while in some cases, animal experiments might prove its pathogenicity[8]. In many cases, a retrospective epidemiological study would be regarded as sufficient to identify a particular virus as the pathogen of a particular disease. Although this is an exception from Koch's postulates, it is natural that ethics restricts etiological study. In contrast, in plant systems, there is always a need to prove pathogenesis in the original host and artificially reproduce the disease of interest, because there are no ethical restrictions in plant disease research. Unlike the majority of plant viruses, which can establish infection after mechanical inoculation onto host plants, some plant viruses cannot establish infection by mechanical inoculation. For example, the *Rice dwarf virus* (RDV) infection in rice can only be established by insect transmission[22]. In addition to viruses that have limited transmission pathways, it should be noted that many viruses are unrelated to pathogenesis.

Natural transmission of the *Rice tungro bacilliform*

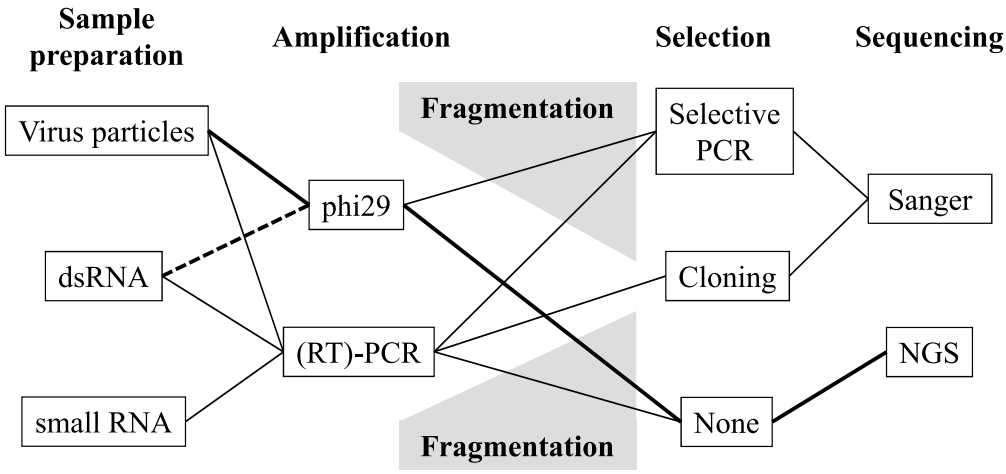

Fig. 1. Workflow of sequencing-based virus hunting and virus detection
Typical methods are shown for four successive steps of sequencing-based virus detection: sample preparation, amplification, selection and sequencing. Some selection and sequencing techniques require DNA fragmentation, as shown by gray trapezoids. Solid lines indicate the workflows that have been reported, and dotted lines show those unreported. The thicker lines indicate workflow that we think suitable for sequencing-based diagnosis.

virus (RTBV) is also restricted to insect transmission. In the case of RTBV, however, it was shown that artificial infection could be established by agroinfection, i.e. the *Agrobacterium*-mediated introduction of viral genome into host cells[6]. Although the agroinfection technique was first established in DNA viruses[9], it is also used in a number of RNA viruses, including those with a divided genome[16, 25]. In many cases, pBin-based binary vectors have been used to construct agroinfection constructs. Those vectors are known as low-copy number plasmids, but we have also experienced the instability of some viral sequences. It is often observed that viral sequences in plasmid vectors affect the growth of *Escherichia coli*. In some cases, viral sequences are rejected by *E. coli*: only the plasmids that lose partial or entire viral sequences could be maintained in the cells of the latter. Therefore, a binary vector is needed, in which any viral genome can be maintained, to establish a system for testing viral pathogenesis. Single copy plasmids might be useful for this purpose. Another important feature of agroinfection constructs is the infectivity on plant hosts. Some additional sequences in both 3' and 5' ends abolish the infectivity in some viruses, but not others (e.g. Refs 2 and 26). Ribozyme sequences, which have been widely used in expressing RNA viral genomes from DNA (e.g. Refs 10 and 17), should be included when constructing a versatile agroinfection vector.

Concluding remarks

The usefulness of sequencing-based virus detection is unquestioned. Moreover, the rapid drop in the cost of DNA sequencing helps make this method more applicable for everyday diagnostics. Accordingly, future studies in this area should focus on developing cost-effective methods, which would change along with the evolution of sequencing technology. The technique of sequencing ribosomal RNA and internal transcribed spacer (ITS) has already become key in the classification and identification of bacteria and fungi[14, 15, 18, 23]. Future sequencing-based plant disease diagnosis should not be limited to viral diseases but would also deal with plant diseases caused by bacteria, fungi or nematodes.

References

1. Adams, I. A. N. P. et al. (2009) Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Mol. Plant Pathol.*, **10**, 537–545.
2. Ahlquist, P. et al. (1984) Multicomponent RNA plant virus infection derived from cloned viral cDNA. *Proc. Natl. Acad. Sci. U. S. A.*, **81**, 7066–7070.
3. Blanco, L. et al. (1989) Highly efficient DNA synthesis by the phage phi 29 DNA polymerase, symmetrical mode of DNA replication. *J Biol. Chem.*, **264**, 8935–8940.
4. Coetzee, B. et al. (2010) Deep sequencing analysis of viruses infecting grapevines: virome of a vineyard. *Virology*, **400**, 157–163.
5. Cox-Foster, D. L. et al. (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. *Science*, **318**, 283.
6. Dasgupta, I. et al. (1991) Rice tungro bacilliform virus DNA independently infects rice after Agrobacterium-mediated transfer. *J. Gen. Virol.*, **72**, 1215.
7. Finkbeiner, S. R. et al. (2008) Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathog.*, **4**, e1000011.
8. Fredericks, D. & Relman, D. A. (1996) Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clinic. Microbiol. Rev.*, **9**, 18–33.
9. Grimsley, N. et al. (1986) Agroinfection, an alternative route for viral infection of plants by using the Ti plasmid. *Proc. Natl. Acad. Sci. U. S. A.*, **83**, 3282–3286.
10. Ishikawa, M. et al. (1997) In vivo DNA expression of functional brome mosaic virus RNA replicons in Saccharomyces cerevisiae. *J. Virol.*, **71**, 7781–7790.
11. Kobayashi, K., Tomita, R. & Sakamoto, M. (2009) Recombinant plant dsRNA-binding protein as an effective tool for the isolation of viral replicative form dsRNA and universal detection of RNA viruses. *J. Gen. Plant Pathol.*, **75**, 87–91.
12. Kreuze, J. F. et al. (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**, 1–7.
13. Kristensen, D. M. et al. (2009) New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.*, **18**, 11–19.
14. Kuninaga, S. et al. (1997) Sequence variation of the rDNA ITS regions within and between anastomosis groups in Rhizoctonia solani. *Current Genet.*, **32**, 237–243.
15. Kusaba, M. & Tsuge, T. (1995) Phologeny of Alternaria fungi known to produce host-specific toxins on the basis of variation in internal transcribed spacers of ribosomal DNA. *Current Genet.*, **28**, 491–498.
16. Liu, L. & Lomonossoff, G. P. (2002) Agroinfection as a rapid method for propagating Cowpea mosaic virus-based constructs. *J. Virol. Methods*, **105**, 343–348.
17. Liu, Y., Schiff, M. & Dinesh-Kumar, S. (2002) Virus-induced gene silencing in tomato. *Plant J.*, **31**, 777–786.
18. Maes, M., Garbeva, P. & Crepel, C. (1996) Identification and sensitive endophytic detection of the fire blight pathogen Erwinia amylovora with 23S ribosomal DNA sequences and the polymerase chain reaction. *Plant pathol.*, **45**, 1139–1149.
19. Mardis, E. R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.*, **24**, 133–141.
20. Mizutani, T. et al. (2007) Rapid genome sequencing of RNA viruses. *Emerg. Infect. Dis.*, **13**, 322–324.
21. Nakamura, S. et al. (2009) Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS One*, **4**, e4219.
22. Omura, T. & Yan, J. (1999) Role of outer capsid proteins in

transmission of Phytoreovirus by insect vectors. *Adv. Virus Res.*, **54**, 15–43.
23. Pastrik, K. H., Elphinstone, J. & Pukall, R. (2002) Sequence analysis and detection of Ralstonia solanacearum by multiplex PCR amplification of 16S-23S ribosomal intergenic spacer region with internal positive control. *Eur. J. Plant Pathol.*, **108**, 831–842.
24. Qu, F., Ren, T. & Morris, T. J. (2003) The coat protein of turnip crinkle virus suppresses posttranscriptional gene silencing at an early initiation step. *J. Virol.*, **77**, 511–522.
25. Ratcliff, F., Martin-Hernandez, A. M. & Baulcombe, D. C. (2001) Technical advance: tobacco rattle virus as a vector for analysis of gene function by silencing. *Plant J.*, **25**, 237–245.
26. Sarnow, P. (1989) Role of 3'-end sequences in infectivity of poliovirus transcripts made in vitro. *J. Virol.*, **63**, 467–470.
27. Sullivan, C. S. & Ganem, D. (2005) A virus-encoded inhibitor that blocks RNA interference in mammalian cells. *J. Virol.*, **79**, 7371–7379.
28. Valverde, R. A., Nameth, S. T. & Jordan, R. L. (1990) Analysis of double-stranded RNA for plant virus diagnosis. *Plant Dis.*, **74**, 255–258.
29. Wang, D. et al. (2002) Microarray-based detection and genotyping of viral pathogens. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 15687–15692.
30. Wang, D. et al. (2003) Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol.*, **1**, E2.
31. Wu, Q. et al. (2010) Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 1606–1611.
32. Yamao, T. et al. (2009) Novel virus discovery in field-collected mosquito larvae using an improved system for rapid determination of viral RNA sequences (RDV ver4. 0). *Arch. Virol.*, **154**, 153–158.