Metagenomic identification of viral pathogens

Kyle Bibby

Department of Civil and Environmental Engineering, University of Pittsburgh, 709 Benedum Hall, 3700 O'Hara Street, Pittsburgh, PA 15261, USA

The target-independent identification of viral pathogens using 'shotgun' metagenomic sequencing is an emerging approach with potentially wide applications in clinical diagnostics, public health monitoring, and viral discovery. In this approach, all viral nucleic acids present in a sample are sequenced in a random, shotgun manner. Pathogens are then identified without the prerequisite of searching for a specific viral pathogen. In this opinion article, I discuss the current state and future research directions for this emerging and disruptive technology. With further technical developments, viral metagenomics has the potential to be deployed as a powerful and widely adopted tool, transforming the way that viral disease is researched, monitored, and treated.

Viral metagenomics as an improved pathogen detection method

The world contains a high diversity of human viral pathogens. There are approximately 200 recognized viral pathogen species, and additional species continue to be discovered at a rate of nearly two per year [1]. Additionally, some viruses also have the potential to mutate rapidly or jump between species, as in the cases of the recent avian [2] and swine [3] influenza epidemics. Each viral species and serotype has unique infectious, transport, and persistence characteristics, placing high value on the ability to rapidly identify viral pathogen species and novel mutations in order to treat and prevent viral disease. Despite the clinical, public health, and research importance of understanding viral type, current viral identification and diagnostic methods are limited by both our incomplete view of viral pathogen diversity and the shortcomings of modern detection methods.

Traditionally, viral pathogens are detected on cultured cell monolayers that exhibit cytopathic effects or plaques, or by antibody neutralization tests. However, many viral types are not culturable in the laboratory and antibody neutralization tests depend on the availability of quality antiserum [4], hindering identification, discovery, and research of these pathogens. Over the past few decades, molecular methods, such as PCR, have been used to detect and study unculturable or nonisolated viruses. However, established molecular methods have two major shortcomings: (i) sequence information for the target viruses must be known, making it difficult to target and study emerging viruses; and (ii) typically, an individual analytical test must be conducted to confirm or refute each pathogen, making identification of rare or unexpected pathogens difficult, and identification of previously unknown pathogens impossible.

Shotgun metagenomic sequencing of medical or environmental samples for viral pathogen identification provides a promising alternative solution that overcomes limitations of traditional methods. Metagenomics is the 'sequence-based analysis of the collective microbial genomes contained in an environmental sample' [5]. Although typically applied to understanding genomic diversity, this methodology has also been used to identify viral pathogens in both environmental [6–8] and medical [9–13] samples. Viral metagenomics-based approaches have primarily been enabled by the great increase in sequencing capacity provided by next-generation sequencing; however, Sanger sequencing methods [10,11], which provide significantly less sequence information, have also been used to identify high-titer pathogens. Finally, metagenomics has been shown to perform better than other multiple-genome detection methods, such as the microarray-based Virochip [14]. Viral metagenomics offer significant advantages over all existing methods of identifying a viral pathogen, including removing the need for targeting a specific pathogen or requiring sequence information for that pathogen, identifying multiple pathogens in a single sample, and eliminating the need for costly and often ineffective culturing or antibody laboratory tests.

The general method of constructing a viral metagenome, outlined in Figure 1, depends on the particular application and resources available to the investigators. Generally, viral metagenome construction includes the isolation of viral particles, extraction of viral DNA and RNA, reverse transcription of RNA to cDNA, fragmentation of nucleic acids, and sequencing nucleic acid fragments. Following sequence generation, short sequences may be assembled into longer sequences to assist in annotation. Raw or assembled sequences are then annotated through some combination of local or global alignment with reference genomes or genes, with further analyses typically done on sequences annotated as human pathogens.

Promise of metagenomic viral pathogen detection

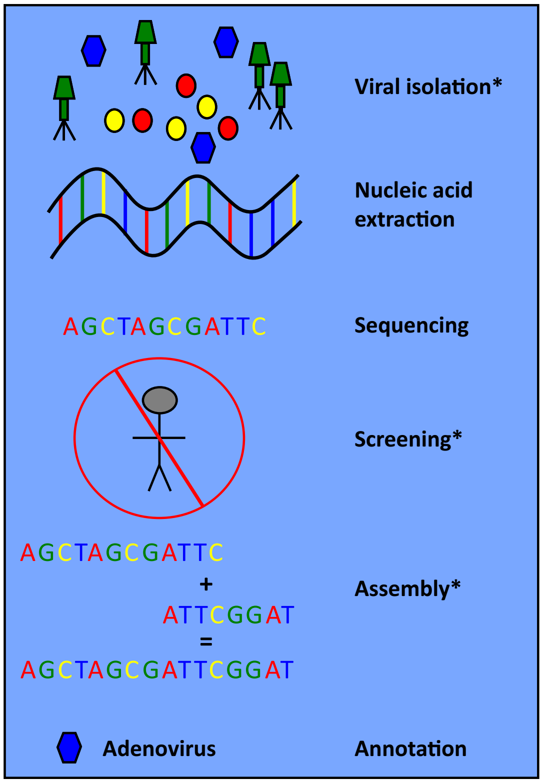
Diagnostics

Clinical diagnostics is perhaps the most desirable developmental goal of viral metagenomic pathogen detection. In many cases, clinical diagnoses of viral infections require tedious culturing and diagnostic tests; however, for rare or

Corresponding author: Bibby, K. (BibbyKJ@Pitt.edu).

Keywords: viral pathogen; metagenome; virome; viral diagnostics; viral discovery; public health; environmental monitoring.

0167-7799/$ – see front matter © 2013 Elsevier Ltd. All rights reserved. http://dx.doi.org/10.1016/j.tibtech.2013.01.016



TRENDS in Biotechnology

Figure 1. General procedure for identifying pathogens using shotgun metagenomics. First, viral particles are isolated from the sample based on size and/or density. Viral nucleic acids are then extracted, prepared, and sequenced. Sequences are then screened to remove nonviral sequences, such as those derived from human or bacterial hosts. Finally, short reads are assembled into longer contiguous sequences and annotated. *Denotes optional or application-dependent steps.

novel pathogens these tests may be insufficient or inconclusive, and even identification of common pathogens remains challenging [15,16]. Additionally, the difficulty associated with identifying viral pathogens may lead to improper clinical treatment, such as the administration of antibiotics [17]. The unambiguous and target-independent identification of all viral pathogens in a relevant clinical sample would help clinicians to direct treatment properly and avoid misdiagnoses. When applied in clinical samples, this approach also has the ability to identify coinfections [18], some of which may previously have been overlooked after the diagnosis of an initial infection. Although technical developments are necessary to deploy viral metagenomics as a stand-alone diagnostic approach, the disruptive potential of viral metagenomics is immense and has great potential to shift the way we conduct clinical diagnoses of viral infections.

Outbreak response

Another application for which metagenomics is well suited is the detection and response to viral pathogen outbreaks [19–22]. This approach has been successfully used in influenza outbreaks, to rapidly determine viral subtype in a case of influenza infection [12]. These identifications enable both the application of the correct therapeutics as well as preventative efforts against potential epidemics, such as immunization development. Another subapplication is for detection of pathogens deployed in bioterrorism attacks [23]. Pathogenic viral strains may be either naturally or engineered to be divergent from known relatives, hindering traditional diagnostics. Synthetic genes inserted into a viral host result in both a more infectious agent and an agent that can escape detection by traditional methods [24]. For example, in the case of a novel wild type Ebola virus outbreak, not all traditional tests identified Ebola virus; however, the metagenomic approach identified the causative agent in all cases [25]. Through a monitoring network enabled by viral metagenomics, previously termed ‘Public Health Metagenome Surveillance’ [26], we will be better suited to defend against and respond to all disease outbreaks, both natural and human-made.

Viral discovery

An emerging role of viral metagenomics is viral discovery [27,28]. There are estimated to be between 36 and 562 viral pathogens that remain to be discovered [1]. For example, in cases of acute diarrhea, an etiological agent is identified in approximately only 60% of cases [10], with many of these unidentified cases expected to be caused by viruses. Also, zoonosis of animal pathogens continually generates novel pathogens [26,29], and may be accelerated by concentrated agricultural operations [30]. Already, metagenomic sequencing of diarrhea of unknown etiology has led to the discovery of novel potentially pathogenic human-associated viruses, including cosavirus [31,32] and klassevirus [33]. Recent environmental metagenomic surveys have shown klassevirus to be widespread globally [34], highlighting the shortcomings of traditional viral detection methodologies. Additionally, novel types of known human viral pathogens have also been discovered, such as a novel rhinovirus [35], a novel bocavirus [36], a novel arenavirus [37], and novel parechoviruses [38]. Solely metagenomic identification of viral pathogens is challenged by the inability to confirm the infectious nature of the virus. These limitations, and potential solutions, are discussed in Box 1. Despite challenges, continued metagenomic sequencing of suspected viral infections will improve our understanding of viral pathogen diversity, help to direct research efforts studying novel pathogens, and inform future metagenomic pathogen surveys.

Environmental monitoring

In complex environmental samples such as sewage-polluted water, the potential for pathogen diversity is high. Typically, areas such as recreational beaches are monitored for concentrations of fecal coliform bacteria [39]; however, it has been long recognized that these bacteria are not accurate indicators of viral pathogens [40]. Pathogen diversity is likely high in environments affected by pollution. Target-independent metagenomic viral pathogen detection has the potential to direct research efforts, risk assessment, and regulation more effectively at pathogens that are highly enriched in the environment [41,42]. Initial surveys have indeed shown high and unexpected pathogen diversity [6,8], with identification of emerging viruses being far more prevalent than for viruses that are typically monitored. Identifying actual pathogen presence

Box 1. Limitations in viral pathogen discovery through metagenomics

Metagenomic sequencing for the identification of novel viral pathogens is inherently challenged by the fact that viruses identified by shotgun metagenomic approaches typically have not been isolated or cultured, limiting future studies of the viruses. Also, in cases where metagenomics is the only tool used to identify a pathogen, the inability to confirm this identification using traditional methods may leave some researchers skeptical. Even in cases where a novel virus is identified in a symptomatic patient, the lack of a pure isolate creates a challenge in satisfying Koch's Postulates – co-occurrence of a virus with a disease symptom does not confirm causation. Koch's Postulates have typically served as the standard by which disease causation is determined; however, satisfaction of these postulates requires isolation of the microbe, which is not accomplished in metagenomic approaches. There have been cases where causation has been demonstrated using molecular methods, but these cases require other demonstrations that depend upon viral particle isolation, such as protein expression and reactivity to anti-serum [50]. In an attempt to work within these guidelines to prove disease causation using metagenomics, the 'Metagenomic Koch's Postulates' have previously been introduced [26]. However, this adaptation still requires the introduction of disease to prove causation; a request that is unlikely to be approved by an institutional review board (IRB) for a human pathogen causing yet-undetermined symptoms. Given technical and ethical challenges, confirming the causative nature of novel human pathogens remains a major technical challenge in confirming the pathogenic nature of novel, human-associated viruses identified using metagenomics.

and diversity through metagenomics, rather than solely the presence of indicator organisms, represents a significant improvement over current environmental monitoring practices and regulation.

A research agenda for metagenomic viral pathogen detection

Continued improvement in sequencing technologies and computational capabilities will certainly assist in the development of metagenomic pathogen detection. However, in order to be deployed in clinical and public health settings, metagenomics requires additional technical development. The great potential and broad utility of viral metagenomics, paired with clear technical and developmental goals, provide an area ripe for research development in many biotechnological disciplines.

Isolation of viral particles

An important initial step in the construction of viral metagenomes is the isolation of viral particles through filtration, flocculation, and density dependent centrifugation [21,43]. These steps serve to concentrate viral particles for more efficient nucleic acid extraction and also help to remove contamination by nonviral cells, maximizing the amount of viral sequence obtained. However, by using charge, density, or size to select for viral particles, each method of concentration potentially biases which viral particles are selected and sequenced. An additional challenge in methods development is that for some sample types, such as clinical swabs or blood samples, the sample size is small. To avoid selection bias, and assist in nucleic acid extraction, some studies of clinical samples have forgone viral concentration (e.g., [44]), presumably because viral pathogen enrichment is expected to be high enough to allow sequencing without concentration. Additionally, there are metagenomic experiments, such as the identification of the Merckel cell polyomavirus [45], that are designed to identify viral infections *in situ*, eliminating the necessity for viral particle isolation. Studies that forego viral particle isolation typically employ bioinformatic-screening methods post-sequencing to remove contaminating DNA – however this means filtering much more than 90% of sequence data, reducing the coverage of pathogen sequences. Improving laboratory methods for the isolation and concentration of mixed viral particles will reduce the amount of sequencing necessary to identify viral pathogens and result in higher sequencing coverage of viral pathogens – assisting in other areas of development, namely, bioinformatic techniques and verification of pathogen identifications.

Bioinformatic techniques

Computational handling of the large amount of sequence data generated in viral metagenome sequencing is another area of necessary research focus. The majority of programs utilized to assemble metagenomic data were originally developed to assemble single genomes, which contain less complexity and have higher sequence coverage. Recognized artifacts in metagenomic assembly include chimeras (genomes artificially combined in assembly) [46] and artificial repeats [47]. There are tools for detecting and sorting these artifacts; however, their effectiveness in detecting false assemblies in metagenomes assembled for pathogen detection and discovery are unknown, suggesting an area for further research and development.

Several annotation schemes may be used to identify pathogens; the most common are local alignments (such as BLAST [48] or the more rapid USEARCH [49]) with reference databases. In cases of novel virus identification where there is no representative in the database, additional bioinformatic assessment of the genome, such as gene calling and alignment with related proteins, is typically necessary. These approaches have been well developed and are commonly accepted for classifying short reads. A previous *in silico* study has indicated that pathogen annotation from short metagenomic sequence reads could be >99% accurate [6]. However, the accuracy of annotation approaches for identifying viral pathogens from real metagenomic datasets, both in avoiding false positives and false negatives, is unknown. This challenge is likely even greater in complex samples such as feces or environmental waters, because these samples have a higher concentration of bacteriophages and non-human eukaryotic viruses, the majority of which have not been characterized and are not in any database. This research area will require the integration of computational biology, clinical microbiology, and biotechnology, and will yield deep insights into not only viral pathogen diversity, but also the ecology of all viruses.

Verification

Along with continued development of laboratory and computational methods for viral pathogen identification, an important step in the development of metagenomic viral pathogen identification is the verification of pathogen identification using well-accepted methods, such as culturing or

PCR. Previous studies have confirmed the presence of pathogens using molecular methods; however, these results have not shown complete agreement between PCR results and metagenome annotation [18]. Additionally, sequence abundance only weakly correlates with pathogen concentration [18]. The lack of correlation between metagenomic methods and PCR may be due to many causes, including lack of necessary sequencing depth, sequencing biases, errors in assembly or annotation, or a false PCR result. Further research of these issues is essential to the continued development and improvement of metagenomic viral pathogen detection.

Concluding remarks

Recent and significant improvements in DNA sequencing capability and cost, as well as quality and availability of cloud computing, suggest that the routine sequencing of viral metagenomes for clinical diagnoses or environmental monitoring could be the norm in the foreseeable future. Such an approach would allow detailed and location-specific information on the spread and ecology of viral disease, facilitating vaccine development and proper clinical treatment. It is somewhat surprising then, when considering the great upside of this technology, that viral metagenomics for pathogen identification is not an area of greater research focus. Although several technical developments are necessary prior to widespread adoption of this method, these are concrete and achievable goals. To encourage the further development of viral metagenomics for clinical applications, the biggest hurdles to overcome include the forces of perception and inertia. Current methods of viral identification are well established and understood by clinicians and microbiologists. Conversely, metagenomic approaches have yet to be fully embraced by the research community, let alone the entire medical and public health fields. Continued development of metagenomic viral pathogen identification is justified by both the broad importance of identifying viral pathogens and the disruptive potential of this approach. Further technical development, including the directions noted in this article, will demonstrate the utility of viral metagenomics for pathogen detection and serve to catalyze the widespread adoption of this approach.

References

1 Woolhouse, M.E.J. *et al.* (2008) Temporal trends in the discovery of human viruses. *Proc. R. Soc. B: Biol. Sci.* 275, 2111–2115

2 Ungchusak, K. *et al.* (2005) Probable person-to-person transmission of avian influenza A (H5N1). *N. Engl. J. Med.* 352, 333–340

3 Smith, G.J.D. *et al.* (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 459, 1122–1125

4 Wang, D. *et al.* (2002) Microarray-based detection and genotyping of viral pathogens. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15687–15692

5 Riesenfeld, C.S. *et al.* (2004) METAGENOMICS: Genomic analysis of microbial communities. *Annu. Rev. Genet.* 38, 525–552

6 Bibby, K. *et al.* (2011) Viral metagenome analysis to guide human pathogen monitoring in environmental samples. *Lett. Appl. Microbiol.* 52, 386–392

7 Cantalupo, P.G. *et al.* (2011) Raw sewage harbors diverse viral populations. *mBio* 2, e00180–e211

8 Rosario, K. *et al.* (2009) Metagenomic analysis of viruses in reclaimed water. *Environ. Microbiol.* 11, 2806–2820

9 Allander, T. *et al.* (2005) Cloning of a human parvovirus by molecular screening of respiratory tract samples. *Proc. Natl. Acad. Sci. U.S.A.* 102, 12891–12896

10 Finkbeiner, S.R. *et al.* (2008) Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathog.* 4, e1000011

11 Svraka, S. *et al.* (2010) Metagenomic sequencing for virus identification in a public-health setting. *J. Gen. Virol.* 91, 2846–2856

12 Yongfeng, H. *et al.* (2011) Direct pathogen detection from swab samples using a new high-throughput sequencing technology. *Clin. Microbiol. Infect.* 17, 241–244

13 Zhang, T. *et al.* (2005) RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.* 4, e3

14 Yozwiak, N.L. *et al.* (2012) Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl. Trop. Dis.* 6, e1485

15 Hammitt, L.L. *et al.* (2012) Specimen collection for the diagnosis of pediatric pneumonia. *Clin. Infect. Dis.* 54 (Suppl 2), S132–S139

16 Talbot, H.K. and Falsey, A.R. (2010) The diagnosis of viral respiratory disease in older adults. *Clin. Infect. Dis.* 50, 747–751

17 Pavia, A.T. (2011) Viral infections of the lower respiratory tract: old viruses, new viruses, and the role of diagnosis. *Clin. Infect. Dis.* 52 (Suppl 4), S284–S289

18 Yang, F. *et al.* (2011) Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J. Clin. Microbiol.* 49, 3463–3469

19 Rosario, K. and Breitbart, M. (2011) Exploring the viral world through metagenomics. *Curr. Opin. Virol.* 1, 289–297

20 Patowary, A. *et al.* (2012) De novo identification of viral pathogens from cell culture hologenomes. *BMC Res. Notes* 5, 11

21 Delwart, E.L. (2007) Viral metagenomics. *Rev. Med. Virol.* 17, 115–131

22 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

23 Fricke, W.F. *et al.* (2009) The role of genomics in the identification, prediction, and prevention of biological threats. *PLoS Biol.* 7, e1000217

24 Clem, A. *et al.* (2007) Virus detection and identification using random multiplex (RT)-PCR with 3′-locked random primers. *Virol. J.* 4, 65

25 Towner, J.S. *et al.* (2008) Newly discovered Ebola virus associated with hemorrhagic fever outbreak in Uganda. *PLoS Pathog.* 4, e1000212

26 Mokili, J. *et al.* (2012) Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2, 63–67

27 Tang, P. and Chiu, C. (2010) Metagenomics for the discovery of novel human viruses. *Future Microbiol.* 5, 177–189

28 Bexfield, N. and Kellam, P. (2011) Metagenomics and the molecular identification of novel viruses. *Vet. J.* 190, 191–198

29 Taylor, L.H. *et al.* (2001) Risk factors for human disease emergence. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 356, 983–989

30 Ramirez, A. *et al.* (2006) Preventing zoonotic influenza virus infection. *Emerg. Infect. Dis.* 12, 996–1000

31 Holtz, L. *et al.* (2008) Identification of a novel picornavirus related to cosaviruses in a child with acute diarrhea. *Virol. J.* 5, 159

32 Kapoor, A. *et al.* (2008) A highly prevalent and genetically diversified Picornaviridae genus in South Asian children. *Proc. Natl. Acad. Sci. U.S.A.* 105, 20482–20487

33 Greninger, A. *et al.* (2009) The complete genome of klassevirus – a novel picornavirus in pediatric stool. *Virol. J.* 6, 82

34 Holtz, L. *et al.* (2009) Klassevirus 1, a previously undescribed member of the family Picornaviridae, is globally widespread. *Virol. J.* 6, 86

35 Lysholm, F. *et al.* (2012) Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS ONE* 7, e30875

36 Kapoor, A. *et al.* (2009) A newly identified bocavirus species in human stool. *J. Infect. Dis.* 199, 196–200

37 Palacios, G. *et al.* (2008) A new arenavirus in a cluster of fatal transplant-associated diseases. *N. Engl. J. Med.* 358, 991–998

38 Linlin, L. *et al.* (2009) Genomic characterization of novel human parechovirus type. *Emerg. Infect. Dis.* 15, 288–291

39 United States Environmental Protection Agency (1986) *Ambient Water Quality Criteria for Bacteria*,

40 Gerba, C.P. *et al.* (1979) Failure of indicator bacteria to reflect the occurrence of enteroviruses in marine waters. *Am. J. Public Health* 69, 1116–1119

41 Wong, K. *et al.* (2012) Application of enteric viruses for fecal pollution source tracking in environmental waters. *Environ. Int.* 45, 151–164

42 Aw, T.G. and Rose, J.B. (2012) Detection of pathogens in water: from phylochips to qPCR to pyrosequencing. *Curr. Opin. Biotechnol.* 23, 422–430

43 Thurber, R.V. *et al.* (2009) Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* 4, 470–483

44 Wylie, K.M. *et al.* (2012) Sequence analysis of the human virome in febrile and afebrile children. *PLoS ONE* 7, e27735

45 Feng, H. *et al.* (2008) Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* 319, 1096–1100

46 Mende, D.R. *et al.* (2012) Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE* 7, e31386

47 Schmieder, R. and Edwards, R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864

48 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

49 Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461

50 Fredericks, D.N. and Relman, D.A. (1996) Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin. Microbiol. Rev.* 9, 18–33