Published in final edited form as:

Future Virol. 2009 January ; 4(1): 47–53. doi:10.2217/17460794.4.1.47.

Advancing full length genome sequencing for human RNA viral pathogens

Appolinaire Djikeng[†] and David Spiro

Appolinaire Djikeng, Infectious Disease Group, The J Craig Venter Institute (JCVI), 9704 Medical Center Drive, Rockville, MD 20850, USA, Tel.: +1 301 795 7681; Fax: +1 301 294 3142; adjikeng@jcv.org
David Spiro, Infectious Disease Group, The J Craig Venter Institute (JCVI), 9704 Medical Center Drive, Rockville, MD 20850, USA, Tel.: +1 301 795 7826; Fax: +1 301 294 3142; dspiro@jcv.org

Abstract

In the face of numerous emerging and re-emerging viral threats, large-scale genome sequencing efforts are underway to monitor viral evolution in real-time. To fully appreciate the mechanisms of viral adaptation and evolution, and to also develop reagents and resources for a better molecular diagnosis of emerging and re-emerging viral infections, there has been an increasing effort toward producing full length viral genome sequences. To date, high-throughput platforms have been developed using traditional Sanger-based sequencing and there are currently prospects to apply next generation sequencing methods to develop an ultra high-throughput strategy for viral genome sequencing and analysis.

Keywords

emerging and re-emerging viruses; genomics; human viruses; viral genome sequencing; viral infectious diseases

Genomics & human viral pathogens

The scientific, medical and public health achievements of the 19th and 20th centuries, including vaccination and antiviral therapies, have brought many infectious diseases (e.g., polio, mumps, measles and rubella) under control [1]. However, during the last three decades we have witnessed several outbreaks of emerging and re-emerging viral diseases (e.g., SARS, influenza and HIV/AIDS) with serious consequences that have led to the weakening of public confidence in national health systems in both developed and developing countries [2,3]. Thus, the identification of viruses in circulation in the human population and understanding the dynamics of these emerging and re-emerging viruses has become an international priority, involving institutions across the globe and researchers across a broad range of disciplines [4,5]. Factors known to favor emergence and re-emergence of viral diseases are of viral (viral evolution, drug resistance, host shift and increase in virulence), societal (increased travel and exchange of goods across continents, deforestation and exploration of new agricultural lands) and environmental (climate change) origins [5,6]. With the ability to quickly and efficiently

[†]Author for correspondence: The J. Craig Venter Institute (JCVI), 9704 Medical Center Drive, Rockville, MD 20850, USA, Tel.: +1 301 795 7681, Fax: +1 301 294 3142, adjikeng@jcv.org.

Website

101. National Institute of Allergy and Infectious Diseases: microbial sequencing centers.
www3.niaid.nih.gov/research/resources/mscs/

generate new viral genome information, the rapid spread of viral pathogens during outbreaks can effectively be countered by constant surveillance and a rapid international public health response, followed by the development of new diagnostics, antiviral drugs and improved vaccines [1].

Pathogenic viruses adapt to new environments, develop drug resistance, evade existing vaccines and widen their host range by the accumulation of mutations throughout their genomes. Thus, complete genomic information is critical for the effective control of human viral disease. However, until recently, public databases, with the exception of HIV, contained limited full genome sequences for many viruses; only partial genomic sequence obtained by targeted amplification of specific genes had been available for diagnosis and analysis of emerging and re-emerging viruses [7]. To address the lack of full length viral genomic data, several large scale viral genome sequencing projects (in addition to efforts in the area of HIV genome sequencing) have been funded by agencies such as the US National Institute for Allergy and Infectious Diseases (NIAID) within a comprehensive microbial sequencing project to produce publicly available sequences for thousands of influenza virus, dengue fever virus, coronavirus, HCV and rotavirus samples [101]. Some of the expected outputs of these projects were to standardize methods for rapid and accurate viral genome sequencing and to provide a large data set for the analysis of temporal and spatial changes leading to the emergence of selected viral infectious diseases. These projects have demonstrated the importance of determining full length viral genome sequences in order to provide a comprehensive picture of viral evolution, pathogenesis and transmission between human and animal [8–12]. During the same period of time, viral metagenomics studies have also emerged and focused on examining the diversity, ecology and even discovery of viruses and the overall community of viruses that reside in a wide variety of terrestrial and aquatic environments, as well as detecting viruses that exist in human tissues, saliva and feces [13].

For most of the large scale RNA virus sequencing projects, methods have been developed based on knowledge of existing sequences. However, recent technological developments have focused on sequencing viral genomes *de novo* without previous knowledge of their sequences [14–16]. Technical approaches used and challenges faced to determine full length viral genome sequences of RNA viruses are discussed in the sections below. We also attempt to provide an outlook of what lies ahead and to underscore the importance of viral genomics in the context of the overall control of human viral diseases.

Sequence-dependent approach for viral genome sequencing

The most commonly used strategy for high-throughput sequencing of RNA viruses involves the design of tiled, overlapping amplicons that span the entire genome followed by the targeted amplification of approximately 1000 bp genomic regions by reverse transcription PCR in 96-well plates. Available sequences for specific RNA viruses are downloaded from public domain databases then aligned using publicly available software. The output of the alignment (Figure 1A) is manually inspected for regions of sequence conservation and primers are designed manually or using an automated primer design pipeline. In the absence of perfect sequence conservation in a region, a degenerate primer is designed. During synthesis, each primer is usually linked to a M13 tag to facilitate high-throughput direct sequencing of amplicons. Sequence reads are then assembled and areas of low coverage are targeted for genome closure (Figure 1C). Ideally, this procedure aims to achieve four-to-sixfold sequencing coverage for any region of the genome; in cases of poor sequence quality or low amplification leading to low coverage (i.e., less than fourfold coverage by two separate amplicons) the corresponding area is selected for direct PCR amplification followed by sequencing of the PCR product to increase the coverage level and validate the consensus sequence.

The advantages of the sequence-dependent approach include its application to both RNA and DNA viruses, its specificity in the presence of large contaminating amounts of host nucleic acids and sensitivity for small amounts of viral nucleic acids. In the case of influenza, Epstein–Barr virus, varicella-zoster virus and enteroviruses full genome sequences have been obtained from total RNA prepared directly from patient samples [17] thus eliminating the need to grow viruses with the associated risk of selecting for fast-growing mutants. By contrast, this method may not be useful for obtaining genomic information from viruses with little or no available sequence data or from viral species that diverge greatly from isolates with published sequence information.

Sequence-independent approach for viral genome sequencing

Viral genome sequencing methods have also been tested and optimized using procedures that do not require prior sequence knowledge of the virus to be characterized. These methods depend on the concentration of viral particles and reduction of the amount of host nucleic acid present (by chemical or nuclease treatment) so as to enable the use of random amplification of viral genomes. Initial sequence-independent approaches focused mostly on viral DNA by using linkers to ligate and amplify viral DNA followed by construction of shotgun libraries (Linker Amplification Shotgun Library [LASL]; [18]) and some modified versions such as the expressible-LASLs [19]. For practical purposes, the LASL method contains three steps: fragmentation of DNA using the Tsp509I restriction endonuclease; ligation of a single linker at either end of the fragmented DNA; and PCR amplification of linker-associated DNA with a linker-targeted primer. The resulting DNA, ranging from 0.5 to more than 4 kb, is cloned in a TOPO TA vector. The decision to use a specific vector depends on whether there is need to select for expressing clones or simply recombinant clones. For functional genomics and metagenomic screening, the choice of the pBAD TOPO TA vector for the selection of expressing clones led to the identification of anti-Bacillus lysins, and an aerolysin from a bacteriophage genome and a community of genomes [19–21]. Notwithstanding the success of the use of LASL for sequencing partial or full length viral genomes it is not well suited to use for RNA viruses, which cannot be detected using this method.

In order to develop a method applicable to both DNA and RNA viruses, several research groups have developed and adapted modifications of the sequence-independent single primer amplification (SISPA) of viral genomes and coupled it to the construction of a plasmid library, mass sequencing of random library clones, and assembly of sequences into full or near full-length viral genomes [14,16]. We propose an adaptation of the original ‘random-PCR’ [22] that we refer to as the random priming (RP)-mediated SISPA or RP-SISPA. The RP-SISPA method is based on random reverse transcription of viral RNA followed by annealing of excess primers to cDNA and conversion into dsDNA by Klenow polymerase and PCR amplification. The reverse transcription is performed using a chimeric primer containing a fixed sequence of 20 nt and a random hexamer attached to its 3′-end. The PCR is subsequently conducted using a single primer derived from the fixed portion of the chimeric primer used for the reverse transcription (Figure 2). For viral DNA (ssDNA and dsDNA), the RP-SISPA procedure begins with the denaturation of DNA followed by annealing of random primers onto ssDNA and the use of the Klenow enzyme to fill the gaps, thus generating a dsDNA. The product of the Klenow reaction is subsequently used as a template for PCR amplification using a single primer. The PCR products, usually ranging in size from 300 to 2600 nt, are purified and cloned into a plasmid vector. Random clones are selected from the library and then sequenced and assembled to build full length viral genomes. We have successfully used the RP-SISPA method to obtain the complete genome sequence of RNA viruses ranging in size from 3000–15,000 kb including the full length genome sequences of a total of 78 rhinovirus isolates available at the American Type Culture Collection (ATCC) [Palmenberg *et al.*, Sequencing and analyses of all known human rhinoviruses genomes. *Science* (2008); Submitted manuscript]. Alternatively, PCR

products can also be processed differently without cloning. We have size-selected PCR products and use fragments of 300 to 600 nt in length for 454 pyrosequencing. In doing so, we have explored the possibility for use of the 454 pyrosequencing for ultra high-throughput sequencing in parallel of several viral genomes as described in the next section.

Viral genome sequencing & massively-parallel sequencing methods

Recent advances in sequencing referred to as the next generation sequencing (NGS) methods [23–25] have revolutionized many aspects of genomics and are rapidly replacing the conventional Sanger sequencing strategy for many applications by offering greatly increased data output with reasonable costs [26]. Current NGS methods can generate approximately 100 Mbp (for 454 with a read size of ~240 nt), 800 Mbp (for Solexa with a read size of 25–40 nt) and 1200 Mbp (for the sequencing platform SOLiD™ with a read size of 25–30 nt) of DNA sequence from a single run. Parameters such as read length, the ability to generate mate pairs, and resolution of homopolymer runs significantly influence the choice of any of the three NGS methods for either *de novo* sequencing or a re-sequencing project. Based on read length and sequence accuracy, the 454 method has arguably emerged as the best candidate for large scale sequencing of RNA viruses. However, due to the high cost of one 454 sequencing run and the relatively small sizes of RNA viruses, optimal use of the 454 requires multiplexing of several viral isolates. Given the output of the 454 pyrosequencing (currently at 100 Mbp bases) it is in principle possible to multiplex hundreds of RNA viral genomes in a single run with adequate coverage to resolve sequence polymorphisms. There have been several successful attempts to barcode different samples and demonstrate the use of 454 for multiplex genome sequencing [27–31]. All these barcoding attempts used 2, 4 and 10-nt barcodes and targeted the sequencing of fractions of genomes such as mtDNA, rDNA, a specific class of RNAs, small interfering RNAs and a section of the HIV genome for mutations in the context of drug resistance. Despite these successful attempts, there is still a need to develop a much more robust barcoding strategy coupled to 454 pyrosequencing or any of the next-generation sequencing methods for multiplex viral genome sequencing. Given the small sizes and relatively low complexity of viral genomes, it is anticipated that the level of throughput provided by 454 pyrosequencing may effectively allow the complete genome sequencing of several viruses in a single run.

One approach to multiplexing viral genome sequencing is outlined in Figure 3. It utilizes 20 nt-long barcodes that are integrated into the viral genome during the RP-SISPA and then are used to differentiate RNA isolates by bioinformatics analysis of the sequence output. Our preliminary experiments [Djikeng *et al.*, Unpublished data] have demonstrated the possibility of mass sequencing of up to 33 different influenza and 33 human rhinovirus genomes per half-454 pyrosequencing run. Detailed analysis of the data output indicated that the sequence space from half of a 454 run could even produce full or near full length of approximately 100 viral genomes (genome size ranging from 6 to 15 kb) at approximately 10× coverage. This approach may, in the near future, allow the full length or near full length sequencing of several hundred isolates of viruses, thus allowing a good understanding of viral evolution and also to develop a rapid disease agent identification and intervention during an outbreak.

Most clinical specimens and tissue culture samples to be used for viral genome sequencing are usually contaminated with human cells, other microorganisms and naked DNA and RNA from disrupted cells. Such samples require enrichment for viral particles. Several approaches have been designed for the enrichment for viral particles. They include nuclease (DNase and RNase) treatment, size selection filtration through commercially available filters for the removal of cellular debris, bacterial cell and other large cells [14,32], and sucrose density gradient and ultracentrifugation [33]. In some instances the contamination is postviral DNA and RNA preparation and in such instances hybridization-based subtractive methods, such as

the representational difference analysis, has been used to enrich for viral nucleic acids [34]. Such enriched viral nucleic acids can then be processed using the RP-SISPA and sequencing.

In addition to the 454 method other NGS methods such as the Solexa and SOLiD sequencing platforms should, in principle, be applicable to viral genome sequencing. However, at the moment none of these methods have so far been used for viral genome sequencing. This may mainly be due to the amount of data output that cannot justify the use for such small genomes unless a multiplex strategy is also designed to allow parallel sequencing of several viruses in a single SOLiD or Solexa run.

Future perspective

At this time, the field of viral genomics is providing thousands of sequences of viral isolates per year. The current bottleneck is caused by the relatively low sequence space provided by Sanger sequencing methods and the difficulties inherent in closing highly divergent viral genomes. With the advent of the NGS methods, sequence data generation will be limited only by the rate of collection, culture and extraction of viral RNA. It is expected that within the next 3–5 years the methodology will exist for mass sequencing of hundreds to thousands of viral genomes in a single reaction. Potential applications of such capability could be for the analysis of thousands of viral isolates during an outbreak or during a season. In conjunction with increased collection and association of epidemiological and ecological metadata this will enable a rapid examination of viral evolutionary processes in near real time. Additionally, large scale coordinated studies using the NGS to generate sequence data from individual human hosts and their viral pathogens will greatly increase our understanding of viral evolution within the context of selective pressures in the human host.

Acknowledgments

Financial & competing interests disclosure

This work was supported in part by the US National Institute of Allergy and Infectious Diseases (NIAID) and by the J. Craig Venter Institute (JCVI). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Bibliography

Papers of special note have been highlighted as: ▪ of interest ▪▪ of considerable interest

1. Rappuoli R. From Pasteur to genomics: progress and challenges in infectious diseases. *Nat Med* 2004;10(11):1177–1185. [PubMed: 15516917]
2. Zessin KH. Emerging diseases: a global and biological perspective. *J Vet Med B Infect Dis Vet Public Health* 2006;53(Suppl 1):7–10. [PubMed: 17123357]
- 3. Woolhouse ME, Howey R, Gaunt E, Reilly L, Chase-Topping M, Savill N. Temporal trends in the discovery of human viruses. *Proc Biol Sci* 2008;275(1647):2111–2115. [PubMed: 18505720]
Summary of human viruses discovered since the beginning of the last century
- 4. Anderson NG, Gerin JL, Anderson NL. Global screening for human viral pathogens. *Emerging Infect Dis* 2003;9(7):768–774. [PubMed: 12890315]A perspective paper on methods, approaches and the importance of the screening for new human viral pathogens
5. Woolhouse M, Gaunt E. Ecological origins of novel human pathogens. *Crit Rev Microbiol* 2007;33(4):231–242. [PubMed: 18033594]
6. Hui EK. Reasons for the increase in emerging and re-emerging viral infectious diseases. *Microbes Infect* 2006;8(3):905–916. [PubMed: 16448839]

7. Egloff MP, Coutard B, Campanacci V, et al. Structural genomics on viral replicative proteins: a tool for antiviral drug discovery. *Antiviral Res* 2005;65(3):A72.
- 8. Ghedin E, Sengamalay NA, Shumway M, et al. Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature* 2005;437(7062):1162–1166. [PubMed: 16208317] First global analysis of influenza virus evolution based on full-length genome sequences
- 9. King CC, Chao DY, Chien LJ, et al. Comparative analysis of full genomic sequences among different genotypes of dengue virus type 3. *J Virol* 5(2008):63. Same as for [8], but also providing a better understanding of different dengue virus genes under selection pressure, with consequences on epidemic prevention and vaccine development
10. Nelson MI, Viboud C, Simonsen L, et al. Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. *PLoS Pathog* 2008;4(2):e1000012. [PubMed: 18463694]
11. Rabadan R, Levine AJ, Robins H. Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes. *J Virol* 2006;80(23):11887–11891. [PubMed: 16987977]
12. Salzberg S, Ghedin E, Spiro D. Shared data are key to beating threat from flu. *Nature* 2006;440(7084):605. [PubMed: 16572147]
13. Delwart EL. Viral metagenomics. *Rev Med Virol* 2007;17(2):115–131. [PubMed: 17295196]
14. Djikeng A, Halpin RA, Kuzmickas R, et al. Viral genome sequencing by random priming methods. *BMC Genomics* 2008;9(1):5. [PubMed: 18179705]
15. Mizutani T, Endoh D, Okamoto M, et al. Rapid genome sequencing of RNA viruses. *Emerging Infect Dis* 2007;13(2):322–324. [PubMed: 17479903]
- 16. Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc Natl Acad Sci USA* 2001;98(20):11609–11614. [PubMed: 11562506] Seminal paper on the discovery of new viruses
17. Sefers SE, Rickmyre J, Blackman A, Li HJ, Edwards K, Tang YW. QIAamp MinElute Virus kit effectively extracts viral nucleic acids from cerebrospinal fluids and nasopharyngeal swabs. *J Clin Virol* 2006;35(2):141–146. [PubMed: 16039902]
18. Breitbart M, Salamon P, Andresen B, et al. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* 2002;99(22):14250–14255. [PubMed: 12384570]
19. Schmitz JE, Daniel A, Collin M, Schuch R, Fischetti VA. Rapid DNA library construction for functional genomic and metagenomic screening. *Appl Environ Microbiol* 2008;74(5):1649–1652. [PubMed: 18083885]
20. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* 2004;68(4):669–685. [PubMed: 15590779]
21. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 2004;38:525–552. [PubMed: 15568985]
- 22. Froussard P. A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. *Nucleic Acids Res* 1992;20(11):2900. [PubMed: 1614887] Seminal paper on the amplification of viral sequences without prior knowledge of their sequences
23. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005;437(7057):376–380. [PubMed: 16056220]
24. Bennett ST, Barnes C, Cox A, Davies L, Brown C. Toward the 1000 dollars human genome. *Pharmacogenomics* 2005;6(4):373–382. [PubMed: 16004555]
25. Gupta PK. Ultrafast and low-cost DNA sequencing methods for applied genomics research. *Proc Natl Acad Sci India B Biol Sci* 2008;78:91–102.
26. Schuster SC. Next-generation sequencing transforms today's biology. *Nat Methods* 2008;5(1):16–18. [PubMed: 18165802]
27. Parameswaran P, Jalili R, Tao L, et al. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res* 2007;35(19):e130. [PubMed: 17932070]
28. Binladen J, Gilbert MT, Bollback JP, et al. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* 2007;2(2):e197. [PubMed: 17299583]

29. Hoffmann C, Minkah N, Leipzig J, et al. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* 2007;35(13):e91. [PubMed: 17576693]
30. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* 2008;5(3):235–237. [PubMed: 18264105]
31. Kasschau KD, Fahlgren N, Chapman EJ, et al. Genome-wide profiling and analysis of *Arabidopsis* siRNAs. *PLoS Biol* 2007;5(3):e57. [PubMed: 17298187]
32. Victoria JG, Kapoor A, Dupuis K, Schnurr DP, Delwart EL. Rapid identification of known and new RNA viruses from animal tissues. *PLoS Pathog* 2008;4(9):e1000163. [PubMed: 18818738]
33. Gabrenaite-Verkhovskaya R, Andreev IA, Kalinina NO, Torrance L, Talianky ME, Makinen K. Cylindrical inclusion protein of potato virus A is associated with a subpopulation of particles isolated from infected plants. *J Gen Virol* 2008;89(Pt 3):829–838. [PubMed: 18272775]
34. Endoh D, Mizutani T, Kirisawa R, et al. Species-independent detection of RNA virus by representational difference analysis using non ribosomal hexanucleotides for reverse transcription. *Nucleic Acids Res* 2005;33(6):e65. [PubMed: 15817564]

Figure 1. Strategy for amplicons-based influenza genome sequencing
(A) Multiple alignments of existing sequences for primer selection. **(B)** Primer and amplicons coverage. **(C)** Sequence assembly and evaluation of sequence quality and coverage of the PB2 genome segment.

Figure 2. Application of the random priming-sequence-independent single primer amplification for viral genome amplification and sequencing

The random priming-sequence-independent single primer amplification begins with the design of two primers. Primer 26 is a chimeric oligonucleotide with a 20 nt sequence fused to a random hexamers sequence. Primer 20 is made up of the 20 nt of the chimeric primer. Primer 26 is used for reverse transcription of viral RNA followed by a Klenow reaction after the excess primer 26 is annealed to the single stranded cDNA. For viral DNA, primer 26 is annealed to denatured viral DNA followed by a Klenow reaction. The dsDNA product of the Klenow reaction is then used for a PCR amplification using primer 20. The PCR product is made up of a heterogeneous mix of amplicons with sizes ranging from 200 to 2500 nt. A fraction of the

amplicons mix is used for sequencing using either Sanger (after cloning into the TOPO TA vector) or 454 pyrosequencing.

Figure 3. Outline of the multiplexing of the RP-SISPA method for ‘ultra’ high-throughput viral genome sequencing

RP-SISPA: Random priming-sequence-independent single primer amplification.

Executive summary

Genomics & human viral pathogens

- Emerging and re-emerging viruses remain a major threat to global health.
- Viral genomics is expected to contribute to a better control of human viral diseases.
- Full length viral genome sequence determination provides complete information on mutations accumulated by the virus.

Sequence-dependent methods for viral genome sequencing

- An approach used for sequencing viruses belonging to a group whose sequences are conserved.

Sequence-independent methods for viral genome sequencing

- An approach used for sequencing both unknown and known viruses.
- Sequence-independent viral genome sequencing is a relatively cost effective method for generating viral genome sequence information.
- Sequence-independent viral genome sequencing allows the discovery of new viruses.

Viral genome sequencing & massively-parallel sequencing methods

- The non-Sanger-based sequencing methods would significantly accelerate the sequencing of viruses.

Future perspective

- Many aspects of viral research, especially the control and surveillance of known and unknown viral infections, would greatly benefit from full-length genome sequences.
 - The increase in the throughput level of viral genome sequencing is anticipated based on the availability of new sequencing methods.
-