World Journal of Virology

Submit a Manuscript: http://www.wjgnet.com/esps/
Help Desk: http://www.wjgnet.com/esps/helpdesk.aspx
DOI: 10.5501/wjv.v4.i3.265

World J Virol 2015 August 12; 4(3): 265-276
ISSN 2220-3249 (online)

REVIEW

Next-generation sequencing in clinical virology: Discovery of new viruses

Sibnarayan Datta, Raghvendra Budhauliya, Bidisha Das, Soumya Chatterjee, Vanlalhmuaka, Vijay Veer

Sibnarayan Datta, Raghvendra Budhauliya, Bidisha Das, Soumya Chatterjee, Vanlalhmuaka, Vijay Veer, Molecular Virology Laboratory, Defence Research Laboratory (DRDO), Tezpur, Assam, PIN-784001, India

Author contributions: Datta S conceptualized and designed the review; Datta S, Budhauliya R and Das B drafted the manuscript; Chatterjee S, Vanlalhmuaka and Veer V edited and critically revised the manuscript.

Supported by The author's laboratory is supported by the Defence Research and Development Organization (DRDO), Ministry of Defence, Government of India.

Conflict-of-interest statement: The authors declare no conflict of interest related to the submitted manuscript.

Correspondence to: Sibnarayan Datta, PhD, Molecular Virology Laboratory, Defence Research Laboratory (DRDO), Post bag No. 2, Tezpur, Assam, PIN-784001, India. sndatta1978@gmail.com
Telephone: +91-3712-258508
Fax: +91-3712-258534

Received: January 24, 2015
Peer-review started: January 27, 2015
First decision: March 6, 2015
Revised: March 23, 2015
Accepted: May 7, 2015
Article in press: May 8, 2015
Published online: August 12, 2015

Abstract

Viruses are a cause of significant health problem worldwide, especially in the developing nations. Due to different anthropological activities, human populations are exposed to different viral pathogens, many of which emerge as outbreaks. In such situations, discovery of novel viruses is utmost important for deciding prevention and treatment strategies. Since last century, a number of different virus discovery methods, based on cell culture inoculation, sequence-independent PCR have been used for identification of a variety of viruses. However, the recent emergence and commercial availability of next-generation sequencers (NGS) has entirely changed the field of virus discovery. These massively parallel sequencing platforms can sequence a mixture of genetic materials from a very heterogeneous mix, with high sensitivity. Moreover, these platforms work in a sequence-independent manner, making them ideal tools for virus discovery. However, for their application in clinics, sample preparation or enrichment is necessary to detect low abundance virus populations. A number of techniques have also been developed for enrichment or viral nucleic acids. In this manuscript, we review the evolution of sequencing; NGS technologies available today as well as widely used virus enrichment technologies. We also discuss the challenges associated with their applications in the clinical virus discovery.

Key words: PCR; Next-generation sequencers; Virus discovery; Sequence-independent single-primer amplification; Virus discovery based on cDNA-AFLP; Rolling circle amplification; Metagenomics

Core tip: Rapid development and commercial availability of next-generation sequencers (NGS) systems have dramatically changed almost every field of biological research, especially microbiology and metagenomics. Different NGS systems have been adapted and used for numerous applications in virology too. These systems are capable of rapidly sequencing and analyzing a complex mixture of nucleic acid templates, in a massively parallel

fashion, making them ideal tools for viral metagenomics and discovery. This manuscript reviews the prevailing NGS technologies, their application in virus discovery to serve as a guide for the readers, working in the field of virology, public health and in biothreat mitigation programs.

Datta S, Budhauliya R, Das B, Chatterjee S, Vanlalhmuaka, Veer V. Next-generation sequencing in clinical virology: Discovery of new viruses. *World J Virol* 2015; 4(3): 265-276 Available from: URL: http://www.wjgnet.com/2220-3249/full/v4/i3/265.htm DOI: http://dx.doi.org/10.5501/wjv.v4.i3.265

INTRODUCTION

Viral infections are a cause of significant health burden globally, particularly in the less developed countries. During the 20th century, methods for virus detection, characterization and taxonomical classification were established, that helped in the discovery of a number of important viruses, in prevention of viral infections and treatment. By the late-1950s, it was generally believed that most of the human pathogenic viruses had been discovered, but the emergence of a number of previously unknown viruses [Hepatitis viruses, Hantavirus, human immunodeficiency virus, Marburg virus, severe acute respiratory syndrome (SARS), Coronavirus Ebola virus] during the later part of the century strongly challenged this belief[1].

It has now become obvious that due to different anthropological activities, such as extensive globalization of travel and business, rapid unplanned urbanization, deforestation, *etc.*, epidemiology of viral diseases have changed significantly[1]. This change has led to the increased exposure of different human populations to newer pathogens, including viruses, mostly zoonotic in nature[2,3]. The emergence of Ebola virus, Nipah virus, Sin Nombre Hantavirus, SARS, Influenza viruses (H1N1, H7N9), and MERS viruses in the recent past[4], clearly signify the onset of many others in the near future. According to a recent statistical estimate, there are at least 320,000 mammalian viruses that are waiting to be discovered[5]. The World Health Organization (WHO) has correctly cautioned that, "It would be extremely naïve and complacent to assume that there will be no other disease like AIDS, Ebola, or SARS, sooner or later"[6].

Apart from natural outbreaks, the risk of pathogens, especially deadly viruses, to be used as biological weapons and agents of bioterrorism have also increased in the recent years[7]. Being exceptionally diverse, in term of etiology, morphology, nucleic acid type and sequence information, clinical manifestations, *etc.*, rapid detection and identification of viruses pose great challenge to clinical investigators. Nevertheless, during natural or deliberate outbreaks, identification and characterization of viruses in clinical samples is extremely essential to facilitate prevention and quarantine strategies, implement specific diagnostic tools, and also to determine explicit treatment strategy.

This article will review the gradual evolution and recent advances in the field of virus discovery, with special reference to the next-generation sequencing (NGS) technologies and related molecular biology methodologies.

EVOLUTION OF VIRUS DISCOVERY TECHNIQUES

Classical approaches to virus discovery

Classically, virus discovery from clinical samples was based on filtration (to remove host cells and other microbes), inoculation of the cell free filtrate in suitable cell cultures followed by purification of the viruses from cultures and their characterization[8-10]. Morphological changes in the cultured cells, collectively known as cytopathic effect, such as formation of syncytia, cell rounding, lysis, detachment, or inclusion bodies, *etc.*, indicate the presence and successful infection of the virus(es) in the cells[11]. Virus isolate(s) are purified from the cultured cells or culture supernatant using density gradient and other high speed centrifugal techniques. This is followed by structural characterization of viral particles, antigens, nucleic acids, through different biophysical and biochemical methods[4]. Although classical methods are sometimes considered as time-consuming, tedious and need significant experimental basis, but the cell inoculation method still remains an exceptional source of enriched viral particles required for serological, molecular characterization and other purposes. Nonetheless, in many cases, viruses are not readily infective to cell cultures, which severely hamper their characterization. Additionally, repeated passaging of the virus to obtain high titer could change the population of virus being sought[12].

Nucleic acid sequence-dependent amplification approaches to virus discovery

Subsequently, with the development of nucleic acid sequence-dependent techniques, such as PCR-sequencing and microarrays, the requirement of cell culture based traditional methods became obsolete to a large extent[13-16]. These techniques were comparatively much faster as compared to classical techniques and led to the discovery of several new genotypes of known viruses. Among PCR and microarray based methods, the former gained enormous popularity due to its ability to rapidly amplify very small amounts of viral sequences from clinical samples. Even though, the prior requirement of sequence information (to design primers and hybridization probe), made this technique suitable for discovery of new genotypes of known viruses, but not appropriate for absolutely novel viruses. This limitation was later addressed by the development of consensus or degenerate PCR[17,18]. Although, this PCR method was tolerant to considerable sequence variation, but it lacked its original sensitivity and was still critically dependent on prior sequence information of the virus genera/family being investigated.

Moreover, this method could only amplify small fractions of viral genome, which were sometimes not enough for further analysis.

Nucleic acid sequence-independent amplification approaches to virus discovery

The limitations of sequence dependent techniques prompted the investigators to resort to "metagenomics", a technique that does not presume any knowledge about the organisms being investigated[19]. Metagenomics is the study of total genetic material present in a given sample, without culturing the organisms present in it. Conventional metagenomics analyses involved direct amplification of the nucleic acids through PCR, cloning and sequencing, *etc.*[15,20]. At the outset, this technique was intensely used for assessing the bacterial diversity within highly diverse samples ranging from soil, oceans, and lakes to human gut and stool, which demonstrated the power of this technique to discover genetic materials of unknown origin[21]. Subsequently, early virus discovery investigators developed a number of random amplification techniques for viral metagenomics, such as sequence-independent single-primer amplification (SISPA), virus discovery based on cDNA-AFLP (VIDISCA), rolling circle amplification (RCA), *etc.*, to amplify viral genetic materials for cloning and sequencing[15,20,22]. Extensive use of these viral metagenomic techniques, led to the discovery of different viruses, including human T-cell lymphotropic virus type-1, Torque Teno virus, different Parvoviruses, Coronaviruses, Polyomaviruses, Hepatitis C virus, Sin Nombre virus, Human Herpesviruses 6 and 8, and West Nile virus *etc.* in clinical samples[23-26].

NGS-based metagenomic approaches to virus discovery

In all the above nucleic acid sequence based virus discovery approaches, the Sanger sequencing method played a very significant role. However, with the commercial availability of high throughput sequencing technologies in 2005, a gradual shift in generation of sequencing technologies became evident. These massively parallel sequencing technologies evolved rapidly and entirely transformed almost every field of biological research including clinical research laboratories[27]. NGS is presently the most attractive approach towards metagenomics, including viral metagenomics, due to its independence from the requirement of prior sequence information. Furthermore, being highly sensitive, NGS can rapidly recuperate nearly full genome sequences of viruses, with relatively less amount of starting material as compared to conventional cloning based approaches[28,29]. Moreover, the large dynamic detection range of the NGS has established it as the most powerful technology available till date, which has catalyzed the rate of virus discovery[30-33]. In combination with conventional methods such as SISPA, VIDISCA, RCA, *etc.*, NGS can dramatically augment turnaround time and sensitivity of virus discovery[23]. Additionally, NGS has enormous, exciting applications in virology, including analysis of viral evolution and quasispecies analysis, antiviral resistance, vaccine, *etc.*[23,33-35].

A comparison of different virus discovery approaches, their advantages and limitations, applicability in different scenarios, *etc.*, is presented in Table 1.

EVOLUTION OF SEQUENCING TECHNOLOGIES

First-generation sequencers

Originally, two different DNA sequencing methods were described almost simultaneously, the Sanger's method, and the Maxam-Gilbert's method[36,37], both considered as the first-generation of sequencing methods. Sanger's method was based on DNA sequencing with chain-terminating inhibitors, while Maxam-Gilbert's method was based on base-specific chemical modification and cleavage of the DNA backbone[38]. Due to its ease and possibility of automation, Sanger's method became instantly popular and was successfully commercialized into DNA sequencing machines. As a result, for almost last 3 decades, the Sanger's method dominated as the gold standard for DNA sequencing[39]. This sequencing method was primarily accomplished by amplification of templates with fluorescently labeled chain-terminating nucleotides, followed by capillary electrophoresis of the amplicons and reading the fluorescence signals, which can provide consistent sequence information of templates up to 1000 bp. Despite its wide use for sequencing pure templates, this sequence method was constrained by its low throughput, higher cost, time and labor involved in sequencing larger genomes. Furthermore, complete dependence on specific primers, inability to sequence the genetic material from a mix of diverse organisms severely restricted its use for direct metagenomic applications.

Second-generation sequencers

To overcome the technological constraints of the Sanger sequencers, second-generation or the NGS technologies were developed, based on a large number of innovations in the amplification technology, sequencing chemistry, microfluidics, imaging technologies, and Bioinformatics, *etc.*[40]. These novel sequencing technologies, initially commercialized by two companies, namely Roche and Illumina, and later by Life Technologies have spectacularly high throughput and high sensitivity, making them more appropriate for direct application in metagenomic studies. As compared to Sanger sequencers, currently available 2nd-generation NGS platforms are capable of generating only short sequence reads, but the true magnificence of NGS lies in their capability to sequence and analyze complex mixes of DNA in a massively parallel manner, generating millions to billions of sequence reads in a single run. Consequently, these technologies are often referred to as "short read" technologies and are distinguished by "third generation" sequencing technologies (or "long read") that provide significantly longer reads (kilobases). However, at present, these long read technologies have, on the whole, lower throughput and accuracy[41-43].

Table 1 A comparative evaluation of the different virus discovery approaches showing advantages and disadvantages associated with them

	Classical approaches (Cell culture and infection based)	Nucleic acid sequence-dependent amplification approaches	Nucleic acid sequence-independent amplification approaches	Next-generation sequencers-based metagenomic approaches
Requirement of cell culture systems	Yes, required for virus particle enrichment	Not required	Not required	Not required
Information about the cytopathic effects of the virus	Yes, could be achieved through cell changes	No information could be achieved	No information could be achieved	No information could be achieved
Requirement of special equipments for purification	Yes, Ultracentrifuge/high speed centrifuges, density gradient is required for preparing pure virus	Not necessary, semi pure preparations obtained through low speed centrifuges are suitable	Not necessary, semi pure preparations obtained through low speed centrifuges are suitable	Not necessary, semi pure preparations obtained through low speed centrifuges are suitable
Information about detailed morphological/structural features of the virus	Yes, could be achieved through Electron/Atomic Force microscopy	No information on virus morphology/structure could be achieved directly	No information on virus morphology/structure could be achieved directly	No information on virus morphology/structure could be achieved directly
Time required for virus identification	Long time is required for identification, ranging from days to weeks	Comparatively faster, days required if cloning and sequencing is involved. Faster with microarray based approaches	Comparatively faster, virus could be identified within few days	Fastest available approach, identification could be done within days and even some times within hours
Requirement of prior knowledge about the virus	Not required	Some information is required regarding genus/family to design primers/probes	Being sequence independent technique, no information is required	Being sequence independent technique, no information is required
Dynamic detection range	Very narrow	Narrow	Wide	Extremely wide
Tolerance to non-viral materials	Vulnerable to other pathogens capable of infecting cell	Being sequence dependent, less vulnerable to other sequences from host and other pathogens	Being sequence in-dependent, more vulnerable to other sequences from host and other pathogens. Virus enrichment techniques required before analysis	Being sequence in-dependent, more vulnerable to other sequences from host and other pathogens. Virus enrichment techniques required before analysis
Suitability for discovery of new viruses	Yes	Less suitable, good at discovery of genotypes/variants of known viruses	Yes	Yes
Suitability during outbreaks	Not suitable due to requirement of long time	Not suitable due to requirement of prior sequence information	Yes, but still considerable time is required during outbreaks	Being fast, very much suitable in detecting pathogens in an outbreak scenario

Even though, widely distinct in their sequencing chemistry and detection technology, NGS platforms are common in terms of massively parallel sequencing of clonally amplified or single DNA molecules. On these platforms, sequencing is executed by repetitive cycles of polymerase-mediated nucleotide extension (Roche-454, Illumina GA) or oligonucleotide ligation (SOLiD). Using a "wash-and-scan" technique, sequence data is acquired as large sets of fluorescence or luminescence images of the flow-cell surface, subsequent to each repetitive sequencing cycle step[44]. This data is later compiled by using a computer-intensive pipeline for image integration, quality assessment, storage, processing and analysis. A typical NGS run generates several hundred megabases (Mb) to gigabases (Gb) of nucleotide sequence data, depending on the platform.

Although NGS platforms commercially available today, provide massive parallel sequencing, but due to their technological features and data output capabilities, every platform is suitable for certain specific applications. Hence, as per explicit requirements, NGS platform needs to be carefully selected. In cases of virus discovery, which is the scope of this review, NGS platforms capable of generating longer sequence reads are preferable over the others. Long reads are extremely useful for *de novo* read assembly and generation of longer contigs, which endow with improved statistical power of finding related sequences in nucleotide database searches[45]. Conversely, for characterization and analysis of virus variants and quasispecies, platforms providing high quality reads, *i.e.*, less error and increased depth became the choice, over longer read lengths. In this review, we will discuss briefly the most popular NGS technologies (Illumina and Roche 454), widely used in virology. The details of the technologies, sequencing chemistries and other applications have been reviewed elsewhere in details[10,31,34].

The most widely used NGS is the Illumina sequencing technology, where clonal amplification of the template is attained to form DNA clusters, using primers attached to solid surface and sequencing is achieved *via* reversible dye-terminator technology. Although Illumina sequencing has higher sequence yield at a relatively low cost per base, this platform has a characteristic systematic base calling bias, exhibit differences in sequence quality, a higher sequencing error rate and increased single-base errors associated with GGC motifs[46-49].

On the other hand, 454 sequencing platforms are based on parallel pyrosequencing, utilizing sequencing-by-synthesis chemistry and chemiluminescence is detected to achieve nucleotide sequence. This method amplifies DNA through an emulsion PCR, generating

clones of DNA using a single template. The main benefit of this technology is its ability to produce long reads, while restricted by its high error rate in homopolymers containing regions, and a high rate of artificial amplification[50-52]. The error rates of NGS are higher relative to the Sanger sequencers, and also require advanced computational tools and statistical calculations before further data processing and assembly[53]. Due to the NGS platform specific errors, presently, use of barcoding strategies, simultaneous sequencing of the samples by two different NGS platforms or high coverage sequencing have been recommended to counteract the effects of errors[54-56]. Nevertheless, these issues are being continually addressed and resolved in the newer versions of these platforms to make them more robust, both in terms of quality and quantity.

With the advancement in instrumentations, NGS platforms are now available as benchtop sequencing instruments in the form of the 454 GS Junior (Roche) and MiSeq (Illumina) which, despite having a small footprint, offer exciting NGS capabilities for clinical settings, at modest running costs[45]. MiSeq includes the Nextera, TruSeq, and reversible terminator-based sequencing by synthesis chemistry and has highest data integrity with broader range of application, including amplicon sequencing, clone checking, small genome sequencing etc. The MiSeq provides maximum throughput per run with lowest error rates, while the 454 GS Junior generates longer reads (approximately 600 bases) with better assemblies, but is limited by lower throughput and homopolymer-associated errors.

Apart from the two most widely used NGS technologies, another technology known as the SOLiD technology (by Life Technologies) is commercially available, but its representation in the scientific literature is limited compared to Roche 454 and Illumina, which might be attributable to its recent availability or complexity of data processing and assembly[57]. Nevertheless, SOLiD is slowly but gradually being accepted as a very reliable platform and has recently been used for *de novo* sequencing of a large mammalian genome[58].

Technical details of the NGS technologies have been extensively reviewed earlier[23,45,59]. A comparison of the currently available NSG systems is also available at the Genohub website (https://genohub.com/ngs-instrument-guide/).

Third-generation sequencers

The third generation of the sequencers has evolved lately, that include the Ion Torrent (Life Technologies), Single-Molecule Real-Time technology SMRT (Pacific Biosciences), and the Nanopore sequencing technology (Oxford Nanopore Technologies). Third-generation sequencers are distinct from their predecessors in two primary features: (1) template amplification is not needed prior to sequencing, which cuts down template preparation time; and cost (2) the signal is registered in real time, directly, during the enzymatic reaction. Apart from the Ion Torrent, rest of the third-generation sequencing technologies is quite recent, and still in the evaluation stages. Moreover, data on their application in the field of virus discovery is extremely scanty. Hence, all these will be discussed only briefly in this review.

The Ion Torrent Personal Genome Machine is based on a semiconductor based sequencing technology and does not require a fluorescence or chemiluminescence based image scanning, resulting in high speed, low cost sequencing system within small size equipment. Cyclically, the semiconductor microfluidic chip is flooded with each nucleotide, and a voltage is generated if it is incorporated, and no voltage is generated when not incorporated. This is based on the fact that every time a nucleotide is incorporated into the DNA molecules, a proton is released, causing a change in voltage, which is subsequently detected and registered by the chip[45,60].

Using the SMRT, single large DNA molecules can be sequenced with high processivity of up to 7 kb, with average read lengths of 3-4 kb[23,61]. On a SMRT cell, numerous Zero-Mode Waveguides are embedded with single set of enzymes and DNA template. During the reaction, enzyme incorporates a nucleotide into the complementary strand, cleaving off fluorescent dye linked with the nucleotide, and this fluorescent signal is captured[61].

Nanopore sequencing is another recently developed method of the third-generation sequencing[62,63]. Nanopore is a tiny biopore with diameter in nanoscale, and involves a heptameric transmembrane channel α-haemolysin (αHL) from *Staphylococcus aureus.* This protein has the ability to tolerate extraordinary voltage and current conditions (up to 100 mV, 100 pA). Under a standard condition of ionic flow, when a DNA molecule is passed through the channel of, *etc*., HL, current is modulated according to the size difference between every deoxyribonucleoside monophosphate (dNMP). This current modulation is detected by standard electrophysiological techniques and the dNMP is identified[62]. Nanopore sequencers are extremely small (size of a USB drive), can sequence long read faster (> 5 kb at a rate of 1 bp/ns), free of fluorescence/chemiluminescence and other enzymes, less sensitive to temperature and other conditions. These benefits make it fit as an extremely rapid sequencing device for field conditions, but the requirement of highly purified DNA needs to be addressed for their wide application in virus discovery.

Among the different NGS platforms available today, choosing the right one for correct application is extremely essential before embarking on a metagenomic project. In case of absence of a reference genome, or where highly divergent sequences are expected, such as in case of virus discovery, *de novo* sequencing and assembly is necessary. Such an assembly requires extensive computational power and datasets containing longer reads with higher coverage are preferable[64-66]. When reference genomes for assembly are available, technologies that generate short reads could also be used to have a high coverage of the metagenomes[53].

When compared in terms of publications, Illumina technology is the most widely used platform, irrespective of application. Earlier the use of this platform was not suited for virus discovery or *de novo* sequencing projects due to its short reads. However, regular augmentation in read length for Illumina platforms has made it suitable for *de novo* assembly of genomes, at a sensitivity, comparable to specific PCR[53,67,68]. However, according to the number of publications, specifically for metagenomic studies, pyrosequencing technology (Roche 454) is preferred over the other NGS approaches producing shorter reads. Of late, Roche has announced the discontinuation of its 454 technology by the mid-2016, which leaves the new investigators with alternative NGS platforms available today.

SAMPLE PREPARATION FOR VIRAL METAGENOMICS AND DISCOVERY

NGS has emerged as the most promising tool for the detection and discovery of novel infectious agents in clinical specimens[23]. However, being unbiased method of sequencing, NGS is greatly affected by very low virus-to-host genome ratios in clinical samples[69-71]. Hence, enrichment of pathogen genetic material or depletion of host genetic materials is essential to maximize sensitivity for discovery of novel pathogens, including viruses in clinical samples[23,72,73]. A schematic representation of the different steps involved in NGS based virus metagenomics and discovery is depicted in Figure 1.

Physical enrichment of virus particles

A number of virus enrichment protocols involving physical and enzymatic techniques have been successfully applied for clinical samples. These include virus capsid purification through freeze/thaw cycles of cell disruption, filtration through appropriate pore membranes (0.45 μm and 0.22 μm), centrifugation, prior nuclease digestion of host genome, *etc.*, followed by extraction of capsid-protected viral nucleic acids, their conversion to cDNA (in case of RNA virus) and non-specific PCR amplification[15]. The efficiency of enrichment in NGS-mediated virus discovery, especially the prior nuclease digestion has been clearly demonstrated by different studies[12,72,74]. Recently Hall *et al*[74] reviewed literatures available on methods for enrichment of viral nucleic acids from clinical samples for NGS-based studies. They found that both ultracentrifugation-mediated enrichment and low-speed centrifugation together with filtration and a nuclease digestion step is widely used for enrichment of viral nucleic acids.

Alternatively, approaches to deplete host genetic materials include use of methylation-specific DNase activity, host ribosomal RNA removal, duplex-specific nuclease normalization methods[75-77]. Such techniques on one hand increase the detection sensitivity of the NGS platform, circumventing the cost and time involved in generating and analyzing huge amounts of background data on the other hand. Ideally, in a clinical setting virus enrichment methods are required to be rapid, standardized and undemanding in terms of cost, manpower or instrumentation facility.

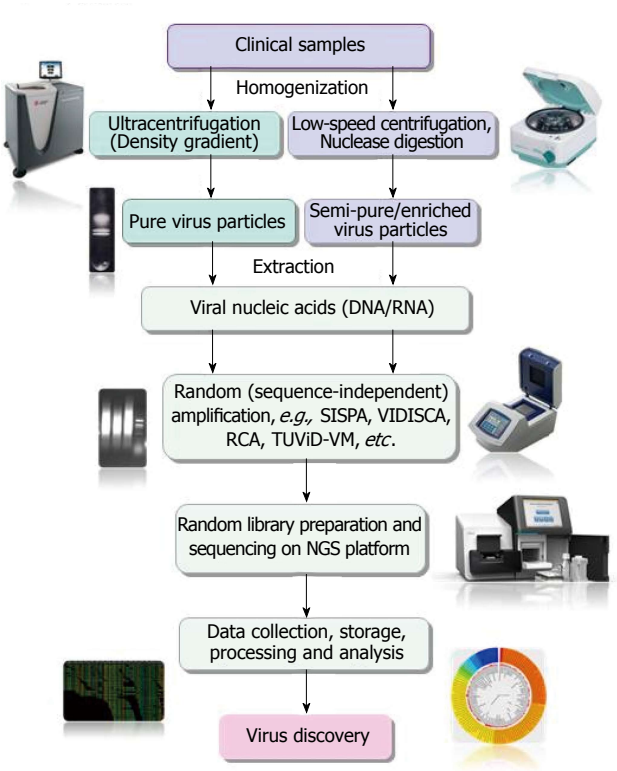

Figure 1 Diagrammatic representation of main steps of clinical virus discovery by next-generation sequencer based technologies.

Enrichment of viral nucleic acids through non-specific amplification techniques

A number of virus enrichment methods have been applied successfully for NGS studies of different clinical samples. Of them, the sequence-independent single primer amplification (SISPA), developed by Reyes and Kim[78], was modified for successful amplification of viral sequences from serum by Allander *et al*[79] and later by others for identification of novel viruses through Sanger sequencing[80-83]. Recently, SISPA was used in combination with NGS and shown to be successful in detection of Hepatitis B and C viruses (HBV, HCV) in solid tissue samples[72]. In a recent study, SISPA-NGS strategy was found to be helpful in detection of Schmallenberg virus (SBV) in veterinary samples[84], suggesting the utility of this technique in screening of field animals that are intermediate hosts to many human viruses. In some of the recent studies no specific physical enrichment of virus particles was applied, but NGS was done on SISPA generated random PCR products, that also resulted in rapid detection of hemorrhagic fever-associated Yellow Fever Virus (YFV), Lujo virus (LUJV), and a new Arenavirus (related to lymphocytic choriomeningitis virus, LCMV) in diverse clinical samples[85-87].

Likewise, another well-established sequence-independent amplification technique is the virus discovery

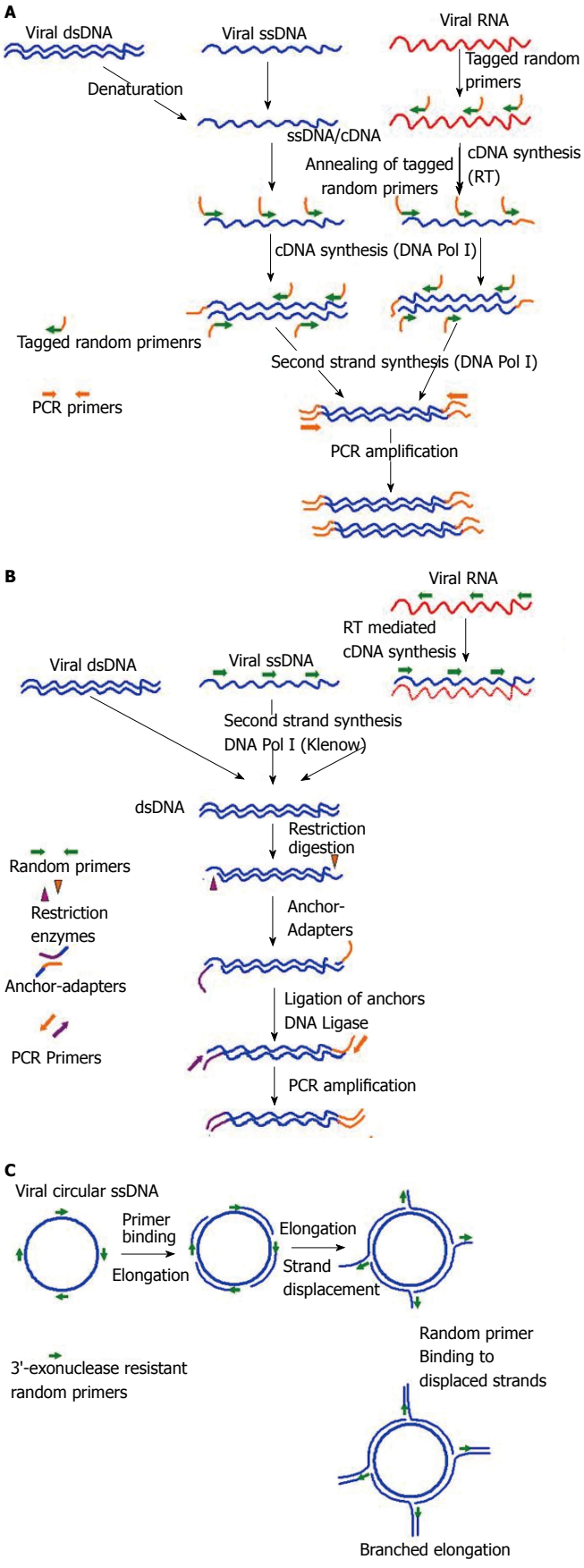

Figure 2 Different virus nucleic acid enrichment techniques. A: Sequence-independent single-primer amplification. Initially viral RNA and ssDNA is transcribed into complementary DNA (cDNA) using reverse transcriptase (RT) and DNA Pol I respectively, with the help of tagged-primers having defined sequence at the 5' end while random nucleotides at the 3' end. Subsequently, second strand synthesis is performed using DNA Pol I (Klenow) to make the cDNA double stranded (dsDNA). Now all the nuceic acids present in the reaction are dsDNA fragments have tagged sequence at their ends. Finally, anchored dsDNA is amplified with primers annealing to the adapter specific sequences, PCR product are checked and ready for analysis through cloning-sequencing or direct sequencing through next-generation sequencers (NGS); B: Virus discovery based on cDNA-AFLP. Initially viral RNA is reverse transcribed into complementary DNA (cDNA) using RT and random primers. Subsequently, second strand synthesis is performed using DNA Pol I (Klenow) to make the cDNA double stranded (dsDNA). In this step, other viral single stranded DNA (ssDNA) viral is also converted to dsDNA. Now all the nuceic acids present in the reaction are dsDNA. In the next step dsDNA are digested with a set of frequent cutter restriction endonucleases, which produce asymmetric cuts. Now specially designed matching anchor-adapters are ligated ends of the restriction fragments using DNA Ligase. Finally, anchored dsDNA is amplified with primers annealing to the adapter specific sequences, PCR product are checked and ready for analysis through cloning-sequencing or direct sequencing through NGS; C: Rolling circle amplification. Amplification of multiply primed single stranded circular viral genomes. 3'-exonuclease resistant primers randomly bind the genome and are elongated by the Phi29 polymerase. The growing strand subsequently displaces the preceding strand of the DNA, making the strand available for binding of random primers and further elongation. This cyclic displacement and elongation leads to a highly branched structure of growing DNA, which is linear in topology. Rolling circle amplification has the capability to specifically enrich the circular ssDNA genomes in an environment of other genetic materials, and could then be characterized by NGS.

cDNA-amplified fragment length polymorphism (VIDISCA), used for discovery of a novel human SARS-associated coronavirus, HCoV-NL63[88,89]. Later this technique was successfully used in combination with Sanger sequencing to discover other novel viruses in clinical samples[90,91]. To late, the utility of this technique in combination with NGS for virus discovery has been demonstrated in veterinary samples, as well as in clinical samples[92,93]. Additionally, Shaukat *et al*[92] modified the VIDISCA method at the reverse transcription step by using specially designed mix of random hexamers that do not anneal to ribosomal RNA, further increasing the specificity of the assay. Apart from SISPA and VIDISCA, multiply-primed RCA has also been demonstrated to enrich circular viral genomes, suitable for sequencing through NSG platforms[94-96]. A diagrammatic representation of SISPA, VIDISCA and RCA is depicted in Figure 2 respectively. Recently, Kohl *et al*[97] reported an ultra-centrifugation and DNA digestion based enrichment protocol followed by SISPA for detection of known and new viruses in human tissue samples. This technique, termed as tissue-based universal virus detection for viral metagenomics was demonstrated to complete within 28 h, making it suitable for discovery of zoonotic and biothreat agents of viral origin during outbreaks[97].

Alternatively, in another study, the authors used a barcoding strategy to carry out unbiased deep sequencing in multiple clinical samples and removed human and other low-quality sequences through bioinformatic filtering pipeline and identified viruses belonging to the Herpesviridae, Flaviviridae, Circoviridae, Anelloviridae, Asfarviridae, and Parvoviridae families in serum samples from tropical febrile illness[2].

Apart from virus discovery and detection in clinical samples, analysis of quasispecies, drug-resistant viral

Table 2 Important bioinformatics challenges associated with application of next-generation sequencers in viral diagnostics action taken or proposed to overcome challenges

Bioinformatics challenges associated with application of NGS in viral diagnostics	Action taken or proposed to overcome challenges
Generation of huge volumes of data by NGS platforms-"data deluge"	Advancement in storage and computation facilities, availability of computer with greater storage and highly powerful processors, cluster/grid computing and cloud computing. Computation facilities needs to be updated with emergence of newer platforms delivering larger datasets
Challenges in uploading data for submission to databases and supercomputing servers for analysis	Requirement of uninterrupted and extremely fast networks
Challenges in storage, public archival and ease of access	Creation of specialized data archive such as the Sequence Read Archive by NIH and ENA (European nucleotide Archive) by EBI. Sharing of data within the three major databases (NIH, EBI and DDBJ) for public accessibility
Challenges in analysis and visualization of large volumes of data, beyond the scope of computation facilities available in molecular biology laboratories	Creation of metagenomic or NGS data analysis pipelines and integrated tool kits, such as those available at NIH-NCBI, EMBL-EBI, MGRAST, CASAVA, MetaVir, Megan, UCSC Genome Browser, BioLinux, *etc*., availability of cloud computing based servers such as Galaxy
Challenges in alignment, *de novo* assembly, gene prediction and phylogenetic analyses NGS datasets, especially short read datasets	Availability of alignment algorithms/programs such as ABySS, ELAND, SOAP, Bowtie, Cloudburst, Zoom, BWA, SHRiMP, MOM, SeqMap, Metagene, Velvet, QSRA, ALLPATHS, EDENA, VCAKE, FragGeneScan, BLAST, GLIMMER, EULER-SR, Avadis, Eagle View, *etc.*
Interpretation of huge amount of data generated in metagenomic analyses by NGS platforms	Proper interpretation of analyzed data is of utmost importance to identify newer pathogens as well as their clinical significance

NGS: Next-generation sequencers.

variants and monitoring of genetic consistency of live viral vaccines there are numerous applications of NGS, which are directly associated with human viral diseases. NGS-based virus detection technique has also been shown to be useful in surveillance of vector-borne and zoonotic viruses[23]. This possibility of detecting arthropod-borne viruses was demonstrated using Dengue virus-infected mosquito pools *(*Aedes aegypti*)*, where, use of NGS resulted in highly sensitive detection of mosquito pools containing infected vectors[98]. Similarly, in a surveillance study focused on the discovery of bat-transmitted pathogens, using coronavirus consensus PCR and unbiased NGS, a new coronavirus related to SARS-CoV was documented[99].

BIOINFORMATICS CHALLENGES ASSOCIATED WITH NGS

Regardless of the field of applications and platforms used, ever-increasing capacities of NGS platforms and their wide usage have resulted in extremely unprecedented volumes of data. This is commonly referred to as "data deluge", and is represented by huge NGS datasets deposited in specialized data archive such as the SRA, a primary archive of NIH, dedicated for submission and storage of raw data and alignment information, generated by all major NGS platforms. Being part of the International Nucleotide Sequence Database Collaboration at the National Center for Biotechnology Information, data submitted to either of the databases SRA, ENA (European nucleotide Archive of European Bioinformatics Institute, EBI) and the DDBJ (DNA Database of Japan) are shared amongst them. SRA serves as an initial point for downstream analysis of NGS data and also provide access to data from human clinical samples to authorized users. According to a recent comparison of GenBank statistics (Release 197, 8/2013 *vs* Release 203, 8/2014), total nucleotide entries to the GenBank represent an annual growth of more than 43%, and annual growth exclusively for virus sequence entries is 21%[100]. This data deluge has posed significant hardware, software and bioinformatics challenges towards storing, transfer, analysis and interpretation of the data[101].

All NGS platforms are advancing towards the capability to sequence longer DNA fragments, and to generate even larger volume of data sets[53]. To analyze such gigantic volumes of data, exceptionally massive computational facilities are also required, which has entirely revolutionized the field of Bioinformatics[60,102]. Once NGS sequence has been generated, the biggest of the challenges comes, *i.e.*, computational requirements for storage and analysis of the massive data sets. Although a detailed description of bioinformatic processes involved in metagenomics data analysis is beyond the scale of this review, the key processes involved in the NGS data analysis are quality assessment, sequence assembly and annotation of the dataset against a database of nucleotide or protein sequences[34]. Quality assessment and data cleaning involves filtering out of low-quality sequences from the dataset, followed by alignment and error correction to separate true variance from the experimental noise[23]. After sequencing and quality assessment, there are two approaches for assembly of the reads. The sequence reads are then mapped to the available reference genome, or individual sequencing reads are assembled *de novo,* using different assembly servers[34,103]. The *de novo* approach is generally followed for discovery of viruses, considering the fact that reference genomes or related sequences may not be available in the databases. To determine the affinity of the assembled reads or the contigs, Basic Local Alignment Search Tool (BLAST) is used, that computes regions of similarity and statistical significance of possible

matches between a query sequence and GenBank submissions[104]. Despite the availability of the BLAST, analyzing a viral metagenome may still be a challenging task in case of highly divergent or novel viral families, which are not represented in the database.

In the Table 2, we have summarized the challenges associated with handling and analysis of NGS generated data, their solutions presently available or suggested.

CONCLUSION

During the last decade, numerous innovations in virus enrichment techniques, sequencing chemistry and signal detection technologies, availability of high end dedicated bioinformatic servers for analysis of the NGS data has greatly accelerated the discovery of viral pathogens in clinical samples. Apart from its increasing applications in virus discovery, NGS has been successfully used in monitoring of antiviral drug resistance, investigation of viral evolution, diversity and quasispecies, and evaluation of the human virome. The supreme advantage of the NGS platforms is their ability to characterize hundreds of different pathogens simultaneously that are not otherwise cultivable using conventional approaches. Nevertheless, there are a number of challenges that need to be overcome for these technologies to become routine in clinical settings. The initial cost of set-up, turnaround time, requirement of powerful computational facilities along with the requirement of a highly skilled group of people are the major barriers to their wide application in resource-limited countries, where the cases of emerging viruses are the highest.

Despite the broad utility of NGS in virus discovery, extremely high sensitivity of this technique also makes it prone to unintentional contamination. The use of random primers for enrichment and the deep sequencing may result in significant potential for carryover contamination from laboratory reagents. Simultaneous analyses of blinded controls may be one approach towards excluding such possibilities, but it will also double the cost of sequencing. Another outcome of the NGS data is the rapid rate of discovery of viruses. However, the absence of appropriate cell culture systems or animal models limit the possibility of experimental studies on these new viruses, thereby the clinical significance of these new viruses remains to be properly understood.

ACKNOWLEDGMENTS

We thankfully acknowledge the Defence Research and Development Organization (DRDO), Ministry of Defence, Government of India for funding and support. We also thank the editor and three anonymous reviewers for their constructive comments, which helped us immensely to improve this manuscript.

REFERENCES

1 **Bichaud L**, de Lamballerie X, Alkan C, Izri A, Gould EA, Charrel RN. Arthropods as a source of new RNA viruses. *Microb Pathog* 2014; **77**: 136-141 [PMID: 25239874 DOI: 10.1016/j.micpath.2014.09.002]

2 **Yozwiak NL**, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl Trop Dis* 2012; **6**: e1485 [PMID: 22347512 DOI: 10.1371/journal.pntd.0001485]

3 **Jones KE**, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P. Global trends in emerging infectious diseases. *Nature* 2008; **451**: 990-993 [PMID: 18288193 DOI: 10.1038/nature06536]

4 **Dong J**, Olano JP, McBride JW, Walker DH. Emerging pathogens: challenges and successes of molecular diagnostics. *J Mol Diagn* 2008; **10**: 185-197 [PMID: 18403608 DOI: 10.2353/jmoldx.2008.070063]

5 **Anthony SJ**, Epstein JH, Murray KA, Navarrete-Macias I, Zambrana-Torrelio CM, Solovyov A, Ojeda-Flores R, Arrigo NC, Islam A, Ali Khan S, Hosseini P, Bogich TL, Olival KJ, Sanchez-Leon MD, Karesh WB, Goldstein T, Luby SP, Morse SS, Mazet JA, Daszak P, Lipkin WI. A strategy to estimate unknown viral diversity in mammals. *MBio* 2013; **4**: e00598-e00513 [PMID: 24003179 DOI: 10.1128/mBio.00598-13]

6 World Health Annual Report 2007. [accessed 2015 Jan 1]. Available from: http://www.who.int/whr/2007/overview/en/index2.html

7 **Bronze MS**, Huycke MM, Machado LJ, Voskuhl GW, Greenfield RA. Viral agents as biological weapons and agents of bioterrorism. *Am J Med Sci* 2002; **323**: 316-325 [PMID: 12074486]

8 **Todaro GJ**, Zeve V, Aaronson SA. Cell culture techniques in the search for cancer viruses of man. *In Vitro* 1971; **6**: 355-361 [PMID: 4360734]

9 **Herrmann EC**. New concepts and developments in applied diagnostic virology. *Prog Med Virol* 1974; **17**: 221-289 [PMID: 4138170]

10 **Lipkin WI**, Firth C. Viral surveillance and discovery. *Curr Opin Virol* 2013; **3**: 199-204 [PMID: 23602435 DOI: 10.1016/j.coviro.2013.03.010]

11 **Friedman RM**, Ramseur JM. Mechanisms of persistent infections by cytopathic viruses in tissue culture. Brief review. *Arch Virol* 1979; **60**: 83-103 [PMID: 226039]

12 **Neill JD**, Bayles DO, Ridpath JF. Simultaneous rapid sequencing of multiple RNA virus genomes. *J Virol Methods* 2014; **201**: 68-72 [PMID: 24589514 DOI: 10.1016/j.jviromet.2014.02.016]

13 **Haagmans BL**, Andeweg AC, Osterhaus AD. The application of genomics to emerging zoonotic viral diseases. *PLoS Pathog* 2009; **5**: e1000557 [PMID: 19855817 DOI: 10.1371/journal.ppat.1000557]

14 **Ansorge WJ**. Next-generation DNA sequencing techniques. *N Biotechnol* 2009; **25**: 195-203 [PMID: 19429539 DOI: 10.1016/j.nbt.2008.12.009]

15 **Delwart EL**. Viral metagenomics. *Rev Med Virol* 2007; **17**: 115-131 [PMID: 17295196 DOI: 10.1002/rmv.532]

16 **Wang D**, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL. Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci USA* 2002; **99**: 15687-15692 [PMID: 12429852 DOI: 10.1073/pnas.242579699]

17 **Rose TM**. CODEHOP-mediated PCR - a powerful technique for the identification and characterization of viral genomes. *Virol J* 2005; **2**: 20 [PMID: 15769292 DOI: 10.1186/1743-422X-2-20]

18 **Kricka LJ**. Nucleic acid detection technologies -- labels, strategies, and formats. *Clin Chem* 1999; **45**: 453-458 [PMID: 10102903]

19 **Handelsman J**. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* 2004; **68**: 669-685 [PMID: 15590779 DOI: 10.1128/MMBR.68.4.669-685.2004]

20 **Bexfield N**, Kellam P. Metagenomics and the molecular identification of novel viruses. *Vet J* 2011; **190**: 191-198 [PMID: 21111643 DOI: 10.1016/j.tvjl.2010.10.014]

21 **Schloss PD**, Handelsman J. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol* 2005; **6**: 229 [PMID: 16086859 DOI: 10.1186/gb-2005-6-8-229]

22 **Thurber RV**, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. *Nat Protoc* 2009; **4**: 470-483 [PMID: 19300441 DOI: 10.1038/nprot.2009.10]

23 **Barzon L**, Lavezzo E, Militello V, Toppo S, Palù G. Applications of next-generation sequencing technologies to diagnostic virology. *Int J Mol Sci* 2011; **12**: 7861-7884 [PMID: 22174638 DOI: 10.3390/ijms12117861]

24 **Chiu CY**. Viral pathogen discovery. *Curr Opin Microbiol* 2013; **16**: 468-478 [PMID: 23725672 DOI: 10.1016/j.mib.2013.05.001]

25 **Schelhorn SE**. Going viral- An integrated view on virological data analysis from basic research to clinical applications. PhD Dissertation. 2013, Saarland University, Germany. Available from: http://d-nb.info/1053980728/34

26 **Boheemen SV**. Virus Discovery and Characterization using Next-Generation Sequencing. PhD Dissertation. 2014, Erasmus Medical Center, Rotterdam. Available from: http://repub.eur.nl/pub/76063/

27 **Lecuit M**, Eloit M. The human virome: new tools and concepts. *Trends Microbiol* 2013; **21**: 510-515 [PMID: 23906500 DOI: 10.1016/j.tim.2013.07.001]

28 **Marston DA**, McElhinney LM, Ellis RJ, Horton DL, Wise EL, Leech SL, David D, de Lamballerie X, Fooks AR. Next generation sequencing of viral RNA genomes. *BMC Genomics* 2013; **14**: 444 [PMID: 23822119 DOI: 10.1186/1471-2164-14-444]

29 **Boonham N**, Kreuze J, Winter S, van der Vlugt R, Bergervoet J, Tomlinson J, Mumford R. Methods in virus diagnostics: from ELISA to next generation sequencing. *Virus Res* 2014; **186**: 20-31 [PMID: 24361981 DOI: 10.1016/j.virusres.2013.12.007]

30 **Wang D**. Fruits of virus discovery: new pathogens and new experimental models. *J Virol* 2015; **89**: 1486-1488 [PMID: 25410872 DOI: 10.1128/JVI.01194-14]

31 **Capobianchi MR**, Giombini E, Rozera G. Next-generation sequencing technology in clinical virology. *Clin Microbiol Infect* 2013; **19**: 15-22 [PMID: 23279287 DOI: 10.1111/1469-0691.12056]

32 **Moore RA**, Warren RL, Freeman JD, Gustavsen JA, Chénard C, Friedman JM, Suttle CA, Zhao Y, Holt RA. The sensitivity of massively parallel sequencing for detecting candidate infectious agents associated with human tissue. *PLoS One* 2011; **6**: e19838 [PMID: 21603639 DOI: 10.1371/journal.pone.0019838]

33 **Svraka S**, Rosario K, Duizer E, van der Avoort H, Breitbart M, Koopmans M. Metagenomic sequencing for virus identification in a public-health setting. *J Gen Virol* 2010; **91**: 2846-2856 [PMID: 20660148 DOI: 10.1099/vir.0.024612-0]

34 **Radford AD**, Chapman D, Dixon L, Chantrey J, Darby AC, Hall N. Application of next-generation sequencing technologies in virology. *J Gen Virol* 2012; **93**: 1853-1868 [PMID: 22647373 DOI: 10.1099/vir.0.043182-0]

35 **Lipkin WI**. Microbe hunting. *Microbiol Mol Biol Rev* 2010; **74**: 363-377 [PMID: 20805403 DOI: 10.1128/MMBR.00007-10]

36 **Sanger F**, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977; **74**: 5463-5467 [PMID: 271968]

37 **Maxam AM**, Gilbert W. A new method for sequencing DNA. 1977. *Biotechnology* 1992; **24**: 99-103 [PMID: 1422074]

38 **Friedmann T**. Rapid nucleotide sequencing of DNA. *Am J Hum Genet* 1979; **31**: 19-28 [PMID: 373426]

39 **Grada A**, Weinbrecht K. Next-generation sequencing: methodology and application. *J Invest Dermatol* 2013; **133**: e11 [PMID: 23856935 DOI: 10.1038/jid.2013.248]

40 **Buermans HP**, den Dunnen JT. Next generation sequencing technology: Advances and applications. *Biochim Biophys Acta* 2014; **1842**: 1932-1941 [PMID: 24995601 DOI: 10.1016/j.bbadis.2014.06.015]

41 **Flaherty P**, Natsoulis G, Muralidharan O, Winters M, Buenrostro J, Bell J, Brown S, Holodniy M, Zhang N, Ji HP. Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Res* 2012; **40**: e2 [PMID: 22013163 DOI: 10.1093/nar/gkr861]

42 **Xu X**, Hou Y, Yin X, Bao L, Tang A, Song L, Li F, Tsang S, Wu K, Wu H, He W, Zeng L, Xing M, Wu R, Jiang H, Liu X, Cao D, Guo G, Hu X, Gui Y, Li Z, Xie W, Sun X, Shi M, Cai Z, Wang B, Zhong M, Li J, Lu Z, Gu N, Zhang X, Goodman L, Bolund L, Wang J, Yang H, Kristiansen K, Dean M, Li Y, Wang J. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* 2012; **148**: 886-895 [PMID: 22385958 DOI: 10.1016/j.cell.2012.02.025]

43 **Navin N**, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, Wigler M. Tumour evolution inferred by single-cell sequencing. *Nature* 2011; **472**: 90-94 [PMID: 21399628 DOI: 10.1038/nature09807]

44 **Schadt EE**, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet* 2010; **19**: R227-R240 [PMID: 20858600 DOI: 10.1093/hmg/ddq416]

45 **Loman NJ**, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012; **30**: 434-439 [PMID: 22522955 DOI: 10.1038/nbt.2198]

46 **Nakamura K**, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* 2011; **39**: e90 [PMID: 21576222 DOI: 10.1093/nar/gkr344]

47 **Schröder J**, Bailey J, Conway T, Zobel J. Reference-free validation of short read data. *PLoS One* 2010; **5**: e12681 [PMID: 20877643 DOI: 10.1371/journal.pone.0012681]

48 **Erlich Y**, Mitra PP, delaBastide M, McCombie WR, Hannon GJ. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat Methods* 2008; **5**: 679-682 [PMID: 18604217 DOI: 10.1038/nmeth.1230]

49 **Dolan PC**, Denver DR. TileQC: a system for tile-based quality control of Solexa data. *BMC Bioinformatics* 2008; **9**: 250 [PMID: 18507856 DOI: 10.1186/1471-2105-9-250]

50 **Quince C**, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 2009; **6**: 639-641 [PMID: 19668203 DOI: 10.1038/nmeth.1361]

51 **Gomez-Alvarez V**, Teal TK, Schmidt TM. Systematic artifacts in metagenomes from complex microbial communities. *ISME J* 2009; **3**: 1314-1317 [PMID: 19587772 DOI: 10.1038/ismej.2009.72]

52 **Margulies M**, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; **437**: 376-380 [PMID: 16056220 DOI: 10.1038/nature03959]

53 **Escalante AE**, Lev Jardón Barbolla, Santiago Ramírez-Barahona, Luis E. Eguiarte. The study of biodiversity in the era of massive sequencing. *Rev Mex de Biodivers* 2014; **85**: 1249-1264 [DOI: 10.7550/rmb.43498]

54 **Fox EJ**, Bayliss KSR, Emond MJ, Loeb LA. Accuracy of Next Generation Sequencing Platforms. *Next Generat Sequenc Applic* 2014; **1**: 1 [DOI: 10.4172/jngsa.1000106]

55 **Dalloul RA**, Long JA, Zimin AV, Aslam L, Beal K, Blomberg Le Ann, Bouffard P, Burt DW, Crasta O, Crooijmans RP, Cooper K, Coulombe RA, De S, Delany ME, Dodgson JB, Dong JJ, Evans C, Frederickson KM, Flicek P, Florea L, Folkerts O, Groenen MA, Harkins TT, Herrero J, Hoffmann S, Megens HJ, Jiang A, de Jong P, Kaiser P, Kim H, Kim KW, Kim S, Langenberger D, Lee MK, Lee T, Mane S, Marcais G, Marz M, McElroy AP, Modise T, Nefedov M, Notredame C, Paton IR, Payne WS, Pertea G, Prickett D, Puiu D, Qioa D, Raineri E, Ruffier M, Salzberg SL, Schatz MC, Scheuring C, Schmidt CJ, Schroeder S, Searle SM, Smith EJ, Smith J, Sonstegard TS, Stadler PF, Tafer H, Tu ZJ, Van Tassell CP, Vilella AJ, Williams KP, Yorke JA, Zhang L, Zhang HB, Zhang X, Zhang Y, Reed KM. Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biol* 2010; **8**: pii: e1000475 [PMID: 20838655 DOI: 10.1371/journal.

pbio.1000475]
56 **Aury JM**, Cruaud C, Barbe V, Rogier O, Mangenot S, Samson G, Poulain J, Anthouard V, Scarpelli C, Artiguenave F, Wincker P. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* 2008; **9**: 603 [PMID: 19087275 DOI: 10.1186/1471-2164-9-603]
57 **Flicek P**, Birney E. Sense from sequence reads: methods for alignment and assembly. *Nat Methods* 2009; **6**: S6-S12 [PMID: 19844229 DOI: 10.1038/nmeth.1376]
58 **Rubin CJ**, Megens HJ, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, Wang C, Carlborg Ö, Jern P, Jørgensen CB, Archibald AL, Fredholm M, Groenen MA, Andersson L. Strong signatures of selection in the domestic pig genome. *Proc Natl Acad Sci USA* 2012; **109**: 19529-19536 [PMID: 23151514 DOI: 10.1073/pnas.1217149109]
59 **Metzker ML**. Sequencing technologies - the next generation. *Nat Rev Genet* 2010; **11**: 31-46 [PMID: 19997069 DOI: 10.1038/nrg2626]
60 **Mardis ER**. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 2008; **9**: 387-402 [PMID: 18576944 DOI: 10.1146/annurev.genom.9.081307.164359]
61 **Quail MA**, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012; **13**: 341 [PMID: 22827831 DOI: 10.1186/1471-2164-13-341]
62 **Branton D**, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich SB, Krstic PS, Lindsay S, Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss JA. The potential and challenges of nanopore sequencing. *Nat Biotechnol* 2008; **26**: 1146-1153 [PMID: 18846088 DOI: 10.1038/nbt.1495]
63 **Laszlo AH**, Derrington IM, Ross BC, Brinkerhoff H, Adey A, Nova IC, Craig JM, Langford KW, Samson JM, Daza R, Doering K, Shendure J, Gundlach JH. Decoding long nanopore sequencing reads of natural DNA. *Nat Biotechnol* 2014; **32**: 829-833 [PMID: 24964173 DOI: 10.1038/nbt.2950]
64 **Cahais V**, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, Chiari Y, Belkhir K, Ranwez V, Galtier N. Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Mol Ecol Resour* 2012; **12**: 834-845 [PMID: 22540679 DOI: 10.1111/j.1755-0998.2012.03148]
65 **Glenn TC**. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* 2011; **11**: 759-769 [PMID: 21592312 DOI: 10.1111/j.1755-0998.2011.03024.x]
66 **Martin JA**, Wang Z. Next-generation transcriptome assembly. *Nat Rev Genet* 2011; **12**: 671-682 [PMID: 21897427 DOI: 10.1038/nrg3068]
67 **Shokralla S**, Spall JL, Gibson JF, Hajibabaei M. Next-generation sequencing technologies for environmental DNA research. *Mol Ecol* 2012; **21**: 1794-1805 [PMID: 22486820 DOI: 10.1111/j.1365-294X.2012.05538.x]
68 **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 CJ, Lecuit M, Burguiere A, Caro V, Eloit M. Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. *J Clin Microbiol* 2011; **49**: 3268-3275 [PMID: 21715589 DOI: 10.1128/JCM.00850-11]
69 **He B**, Li Z, Yang F, Zheng J, Feng Y, Guo H, Li Y, Wang Y, Su N, Zhang F, Fan Q, Tu C. Virome profiling of bats from Myanmar by metagenomic analysis of tissue samples reveals more novel Mammalian viruses. *PLoS One* 2013; **8**: e61950 [PMID: 23630620 DOI: 10.1371/journal.pone.0061950]
70 **Baker KS**, Leggett RM, Bexfield NH, Alston M, Daly G, Todd S, Tachedjian M, Holmes CE, Crameri S, Wang LF, Heeney JL, Suu-Ire R, Kellam P, Cunningham AA, Wood JL, Caccamo M, Murcia PR. Metagenomic study of the viruses of African straw-coloured fruit bats: detection of a chiropteran poxvirus and isolation of a novel adenovirus. *Virology* 2013; **441**: 95-106 [PMID: 23562481 DOI: 10.1016/j.virol.2013.03.014]
71 **Nakamura S**, Yang CS, 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. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS One* 2009; **4**: e4219 [PMID: 19156205 DOI: 10.1371/journal.pone.0004219]
72 **Daly GM**, Bexfield N, Heaney J, Stubbs S, Mayer AP, Palser A, Kellam P, Drou N, Caccamo M, Tiley L, Alexander GJ, Bernal W, Heeney JL. A viral discovery methodology for clinical biopsy samples utilising massively parallel next generation sequencing. *PLoS One* 2011; **6**: e28879 [PMID: 22216131 DOI: 10.1371/journal.pone.0028879]
73 **Whon TW**, Kim MS, Roh SW, Shin NR, Lee HW, Bae JW. Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. *J Virol* 2012; **86**: 8221-8231 [PMID: 22623790 DOI: 10.1128/JVI.00293-12]
74 **Hall RJ**, Wang J, Todd AK, Bissielo AB, Yen S, Strydom H, Moore NE, Ren X, Huang QS, Carter PE, Peacey M. Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J Virol Methods* 2014; **195**: 194-204 [PMID: 24036074 DOI: 10.1016/j.jviromet.2013.08.035]
75 **Oyola SO**, Gu Y, Manske M, Otto TD, O'Brien J, Alcock D, Macinnis B, Berriman M, Newbold CI, Kwiatkowski DP, Swerdlow HP, Quail MA. Efficient depletion of host DNA contamination in malaria clinical sequencing. *J Clin Microbiol* 2013; **51**: 745-751 [PMID: 23224084 DOI: 10.1128/JCM.02507-12]
76 **He S**, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG, Wang Z, Chen F, Lindquist EA, Sorek R, Hugenholtz P. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat Methods* 2010; **7**: 807-812 [PMID: 20852648 DOI: 10.1038/nmeth.1507]
77 **Shagina I**, Bogdanova E, Mamedov IZ, Lebedev Y, Lukyanov S, Shagin D. Normalization of genomic DNA using duplex-specific nuclease. *Biotechniques* 2010; **48**: 455-459 [PMID: 20569220 DOI: 10.2144/000113422]
78 **Reyes GR**, Kim JP. Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Mol Cell Probes* 1991; **5**: 473-481 [PMID: 1664049 DOI: 10.1016/S0890-8508(05)80020-9]
79 **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**: 11609-11614 [PMID: 11562506 DOI: 10.1073/pnas.211424698]
80 **Cheng WX**, Li JS, Huang CP, Yao DP, Liu N, Cui SX, Jin Y, Duan ZJ. Identification and nearly full-length genome characterization of novel porcine bocaviruses. *PLoS One* 2010; **5**: e13583 [PMID: 21049037 DOI: 10.1371/journal.pone.0013583]
81 **Abed Y**, Boivin G. Molecular characterization of viruses from clinical respiratory samples producing unidentified cytopathic effects in cell culture. *Viruses* 2009; **1**: 84-90 [PMID: 21994539 DOI: 10.3390/v1020084]
82 **Victoria JG**, Kapoor A, Li L, Blinkova O, Slikas B, Wang C, Naeem A, Zaidi S, Delwart E. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J Virol* 2009; **83**: 4642-4651 [PMID: 19211756 DOI: 10.1128/JVI.02301-08]
83 **Kirkland PD**, Frost MJ, Finlaison DS, King KR, Ridpath JF, Gu X. Identification of a novel virus in pigs--Bungowannah virus: a possible new species of pestivirus. *Virus Res* 2007; **129**: 26-34 [PMID: 17561301 DOI: 10.1016/j.virusres.2007.05.002]
84 **Rosseel T**, Scheuch M, Höper D, De Regge N, Caij AB, Vandenbussche F, Van Borm S. DNase SISPA-next generation sequencing confirms Schmallenberg virus in Belgian field samples and identifies genetic variation in Europe. *PLoS One* 2012; **7**: e41967 [PMID: 22848676 DOI: 10.1371/journal.pone.0041967]
85 **McMullan LK**, Frace M, Sammons SA, Shoemaker T, Balinandi S, Wamala JF, Lutwama JJ, Downing RG, Stroeher U, MacNeil A, Nichol ST. Using next generation sequencing to identify yellow fever virus in Uganda. *Virology* 2012; **422**: 1-5 [PMID: 21962764 DOI: 10.1016/j.virol.2011.08.024]

86 **Briese T**, Paweska JT, McMullan LK, Hutchison SK, Street C, Palacios G, Khristova ML, Weyer J, Swanepoel R, Egholm M, Nichol ST, Lipkin WI. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. *PLoS Pathog* 2009; **5**: e1000455 [PMID: 19478873 DOI: 10.1371/journal.ppat.1000455]

87 **Palacios G**, Druce J, Du L, Tran T, Birch C, Briese T, Conlan S, Quan PL, Hui J, Marshall J, Simons JF, Egholm M, Paddock CD, Shieh WJ, Goldsmith CS, Zaki SR, Catton M, Lipkin WI. A new arenavirus in a cluster of fatal transplant-associated diseases. *N Engl J Med* 2008; **358**: 991-998 [PMID: 18256387 DOI: 10.1056/NEJMoa073785]

88 **Pyrc K**, Jebbink MF, Berkhout B, van der Hoek L. Detection of new viruses by VIDISCA. Virus discovery based on cDNA-amplified fragment length polymorphism. *Methods Mol Biol* 2008; **454**: 73-89 [PMID: 19057862 DOI: 10.1007/978-1-59745-181-9_7]

89 **van der Hoek L**, Pyrc K, Jebbink MF, Vermeulen-Oost W, Berkhout RJ, Wolthers KC, Wertheim-van Dillen PM, Kaandorp J, Spaargaren J, Berkhout B. Identification of a new human coronavirus. *Nat Med* 2004; **10**: 368-373 [PMID: 15034574 DOI: 10.1038/nm1024]

90 **de Souza Luna LK**, Baumgarte S, Grywna K, Panning M, Drexler JF, Drosten C. Identification of a contemporary human parechovirus type 1 by VIDISCA and characterisation of its full genome. *Virol J* 2008; **5**: 26 [PMID: 18269761 DOI: 10.1186/1743-422X-5-26]

91 **de Vries M**, Pyrc K, Berkhout R, Vermeulen-Oost W, Dijkman R, Jebbink MF, Bruisten S, Berkhout B, van der Hoek L. Human parechovirus type 1, 3, 4, 5, and 6 detection in picornavirus cultures. *J Clin Microbiol* 2008; **46**: 759-762 [PMID: 18077635 DOI: 10.1128/JCM.02009-07]

92 **Shaukat S**, Angez M, Alam MM, Jebbink MF, Deijs M, Canuti M, Sharif S, de Vries M, Khurshid A, Mahmood T, van der Hoek L, Zaidi SS. Identification and characterization of unrecognized viruses in stool samples of non-polio acute flaccid paralysis children by simplified VIDISCA. *Virol J* 2014; **11**: 146 [PMID: 25112200 DOI: 10.1186/1743-422X-11-146]

93 **van der Heijden M**, de Vries M, van Steenbeek FG, Favier RP, Deijs M, Brinkhof B, Rothuizen J, van der Hoek L, Penning LC. Sequence-independent VIDISCA-454 technique to discover new viruses in canine livers. *J Virol Methods* 2012; **185**: 152-155 [PMID: 22664180 DOI: 10.1016/j.jviromet.2012.05.019]

94 **de Vries M**, Deijs M, Canuti M, van Schaik BD, Faria NR, van de Garde MD, Jachimowski LC, Jebbink MF, Jakobs M, Luyf AC, Coenjaerts FE, Claas EC, Molenkamp R, Koekkoek SM, Lammens C, Leus F, Goossens H, Ieven M, Baas F, van der Hoek L. A sensitive assay for virus discovery in respiratory clinical samples. *PLoS One* 2011; **6**: e16118 [PMID: 21283679 DOI: 10.1371/journal.pone.0016118]

95 **Meiring TL**, Salimo AT, Coetzee B, Maree HJ, Moodley J, Hitzeroth II, Freeborough MJ, Rybicki EP, Williamson AL. Next-generation sequencing of cervical DNA detects human papillomavirus types not detected by commercial kits. *Virol J* 2012; **9**: 164 [PMID: 22897914 DOI: 10.1186/1743-422X-9-164]

96 **Sijmons S**, Thys K, Corthout M, Van Damme E, Van Loock M, Bollen S, Baguet S, Aerssens J, Van Ranst M, Maes P. A method enabling high-throughput sequencing of human cytomegalovirus complete genomes from clinical isolates. *PLoS One* 2014; **9**: e95501 [PMID: 24755734 DOI: 10.1371/journal.pone.0095501]

97 **Kohl C**, Brinkmann A, Dabrowski PW, Radonić A, Nitsche A, Kurth A. Protocol for metagenomic virus detection in clinical specimens. *Emerg Infect Dis* 2015; **21**: 48-57 [PMID: 25532973 DOI: 10.3201/eid2101.140766]

98 **Bishop-Lilly KA**, Turell MJ, Willner KM, Butani A, Nolan NM, Lentz SM, Akmal A, Mateczun A, Brahmbhatt TN, Sozhamannan S, Whitehouse CA, Read TD. Arbovirus detection in insect vectors by rapid, high-throughput pyrosequencing. *PLoS Negl Trop Dis* 2010; **4**: e878 [PMID: 21085471 DOI: 10.1371/journal.pntd.0000878]

99 **Quan PL**, Firth C, Street C, Henriquez JA, Petrosov A, Tashmukhamedova A, Hutchison SK, Egholm M, Osinubi MO, Niezgoda M, Ogunkoya AB, Briese T, Rupprecht CE, Lipkin WI. Identification of a severe acute respiratory syndrome coronavirus-like virus in a leaf-nosed bat in Nigeria. *MBio* 2010; **1**: pii: e00208-10 [PMID: 21063474 DOI: 10.1128/mBio.00208-10]

100 **Benson DA**, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res* 2013; **41**: D36-D42 [PMID: 23193287 DOI: 10.1093/nar/gks1195]

101 **Pop M**, Salzberg SL. Bioinformatics challenges of new sequencing technology. *Trends Genet* 2008; **24**: 142-149 [PMID: 18262676 DOI: 10.1016/j.tig.2007.12.006]

102 **Henson J**, Tischler G, Ning Z. Next-generation sequencing and large genome assemblies. *Pharmacogenomics* 2012; **13**: 901-915 [PMID: 22676195 DOI: 10.2217/pgs.12.72]

103 **Sharma D**, Priyadarshini P, Vrati S. Unraveling the web of viroinformatics: computational tools and databases in virus research. *J Virol* 2015; **89**: 1489-1501 [PMID: 25428870 DOI: 10.1128/JVI.02027-14]

104 **Altschul SF**, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215**: 403-410 [PMID: 2231712 DOI: 10.1016/S0022-2836(05)80360-2]

P- Reviewer: Chen YD, Demonacos C, Qiu HJ **S- Editor**: Song XX
L- Editor: A **E- Editor**: Yan JL

Published by **Baishideng Publishing Group Inc**
8226 Regency Drive, Pleasanton, CA 94588, USA
Telephone: +1-925-223-8242
Fax: +1-925-223-8243
E-mail: bpgoffice@wjgnet.com
Help Desk: http://www.wjgnet.com/esps/helpdesk.aspx
http://www.wjgnet.com