Available online at www.sciencedirect.com
SciVerse ScienceDirect
Current Opinion in Microbiology

Viral pathogen discovery

Charles Y Chiu[1,2]

Viral pathogen discovery is of critical importance to clinical microbiology, infectious diseases, and public health. Genomic approaches for pathogen discovery, including consensus polymerase chain reaction (PCR), microarrays, and unbiased next-generation sequencing (NGS), have the capacity to comprehensively identify novel microbes present in clinical samples. Although numerous challenges remain to be addressed, including the bioinformatics analysis and interpretation of large datasets, these technologies have been successful in rapidly identifying emerging outbreak threats, screening vaccines and other biological products for microbial contamination, and discovering novel viruses associated with both acute and chronic illnesses. Downstream studies such as genome assembly, epidemiologic screening, and a culture system or animal model of infection are necessary to establish an association of a candidate pathogen with disease.

Addresses
[1] Department of Laboratory Medicine, University of California San Francisco, San Francisco, CA, USA
[2] UCSF-Abbott Viral Diagnostics and Discovery Center, San Francisco, CA, USA

Corresponding author: Chiu, Charles Y (charles.chiu@ucsf.edu)

Current Opinion in Microbiology 2013, **16**:468–478

This review comes from a themed issue on **Host–microbe interactions: viruses**

Edited by **Carlos F Arias**

For a complete overview see the Issue and the Editorial

Available online 29th May 2013

1369-5274

http://dx.doi.org/10.1016/j.mib.2013.05.001

Introduction

The identification of novel pathogens has a tremendous impact on infectious diseases, microbiology, and human health. Nearly all of the outbreaks of clinical and public health importance over the past two decades have been caused by novel emerging viruses, including Severe Acute Respiratory Syndrome (SARS) coronavirus [1], Sin Nombre hantavirus [2], 2009 pandemic influenza H1N1 [3,4], and the recently described coronavirus EMC [5–7] and H7N9 avian influenza viruses [8], with most originating from animal reservoirs. Changes in the environment, globalization, growth of wet (live animal) markets, and the rapid expansion of the human population into wildlife habitats all promote the rapid spread of previously unidentified pathogens that are capable of causing widespread and devastating epidemics of human illness [9]. Arthropods such as mosquitoes and ticks are vectors for emerging pathogens including West Nile virus [10,11], and the Severe Fever and Thrombocytopenia Syndrome (SFTS) [12,13] and Heartland bunyaviruses [14]. Moreover, the link between new viruses and disease is not only restricted to acute illnesses, but also can be seen in chronic disease states, as demonstrated by the strong association between infection by the novel Merkel cell polyomavirus (MCPyV) and a rare, highly aggressive skin tumor in elderly patients [15].

Currently available diagnostic tests for pathogens are generally narrow in scope and fail to detect an agent in a significant fraction of cases. Traditional methods such as culture, serology, or targeted nucleic acid-based testing, such as specific polymerase chain reaction (PCR), have limited utility in investigations where there is no *a priori* knowledge of the identity of potential infectious agents. Notably, in certain infectious diseases such as encephalitis, conventional testing fails to identify a pathogen in up to 70% of cases [16–18]. In contrast, state-of-the-art genomic technologies such as pan-microbial microarrays or unbiased next-generation sequencing (NGS) can be attractive tools for broad-based pathogen discovery. Nearly all infectious agents, with the sole exception of prions [19], contain either RNA or DNA, and are thus amenable to nucleic acid-based detection. In principle, these technologies are capable of comprehensively identifying all potential pathogens in clinical samples from humans and animals. This review will describe the genomic approaches for pathogen discovery currently being employed in the field, and highlight recent examples of their use in the discovery and characterization of novel viral pathogens (Table 1).

Genomic approaches for pathogen discovery

Pathogen discovery entails the use of genomic-based methods to identify novel microbes, followed by further investigation to determine potential associations with disease (Figure 1). As a pathogen discovery tool, consensus PCR uses degenerate primers to detect conserved sequences that are broadly shared between members of a group. This approach was recently used to identify novel paramyxoviruses in samples from large-scale surveys of bats and rodents [20–22] and emerging viruses such as coronavirus EMC, the cause of a new severe and

Table 1

Some recent examples of viral pathogen discovery[d]

Name[a]	Detection platform	NGS bioinformatics approach	Disease assocation	Strength of association[c]
Coronavirus-EMC	Culture and 454 NGS [7]	*De novo* genome assembly	Severe pneumonia (humans)	++
SFTS (severe fever with thrombocytopenia virus) bunyavirus [12,13]	Illumina NGS [12]	Subtraction and BLAST search	Severe fever with thrombocytopenia	++
Heartland bunyavirus	Culture and 454 NGS [14]	BLAST search and *de novo* gene assembly	Severe febrile illness	++
MCPyV (Merkel cell polyomavirus)	454 NGS [15]	Subtraction and BLAST search	Merkel cell carcinoma (MCC)	++
Bat paramyxoviruses	Consensus PCR [20–22]	N/A	–	–
Raccoon polyomavirus	Consensus PCR and RCA [38]	N/A	Brain tumors (racoons)	++
HPyV6 and HPyV7 (human polyomaviruses 6 and 7)	RCA [39]	N/A	N/A	–
TSPyV (trichodysplasia spinulosa-associated polyomavirus)	RCA [40]	N/A	Trichodysplasia spinulosa	++
2009 pandemic influenza A(H1N1)[b]	Microarray and Illumina NGS [51]	Subtraction and BLAST search	Febrile illness	++
	454 NGS [110,111]	BLAST search	Febrile illness	++
	Illumina NGS [112]	BLAST search	Febrile illness	++
TMAdV (titi monkey adenovirus)	Microarray and Illumina NGS [52]	BLAST search	Pneumonia (titi monkeys)	++
BASV (Bas-Congo virus), a rhabdovirus	Illumina NGS [58]	Subtraction, BLAST search, and *de novo* genome assembly	Acute hemorrhagic fever	++
Novel circoviruses and cycloviruses in humans and monkeys	454 NGS [59]	Subtraction and BLAST search	Diarrhea	–
Human klassevirus/salivirus	Illumina NGS [61]	Subtraction and BLAST search	Diarrhea	–
	454 NGS [60,62]	Subtraction and BLAST search	Diarrhea	–
MWPyV/HPy10/MXPyV (MW polyomavirus)	454 NGS [63]	Subtraction and BLAST search	Diarrhea	–
	Illumina NGS [64]	Subtraction and BLAST search	Diarrhea	–
	Illumina NGS [65]	BLAST search	WHIM syndrome	–
HPyV9 (human polyomavirus 9)	Illumina NGS [66]	Subtraction and BLAST search	–	–
	Consensus PCR [24]	N/A	–	–
Human bufavirus	454 NGS [67]	Subtraction and BLAST search	Diarrhea	–
HAstV-PS (human astrovirus Puget Sound)	454 NGS [68]	Subtraction and BLAST search	Encephalitis	++
Human enterovirus 109	Consensus PCR and Illumina NGS [69]	Subtraction and BLAST search	Acute respiratory illness	+
Dandenong arenavirus	454 NGS [70]	Subtraction and BLAST search	Fatal febrile illness in transplant patients	++
Lujo arenavirus	454 NGS [71]	Subtraction and BLAST search	Acute hemorrhagic fever	++
TDAV (Theiler's disease-associated virus), a novel pegivirus	Illumina NGS [73]	BLAST search and *de novo* genome assembly	Hepatitis (horses)	++
Bat, canine, horse, and rodent hepaciviruses and pegiviruses	454 NGS [74–77]	BLAST search	Respiratory infection (dogs)	–
Canine bocavirus 3	Illumina NGS [78]	BLAST search	Hemorrhagic diarrhea and vasculitis (dog)	–
Snake arenaviruses	Illumina NGS [79]	Subtraction and BLAST search	Inclusion body disease (snakes)	++

Table 1 (*Continued*)

Name[a]	Detection platform	NGS bioinformatics approach	Disease assocation	Strength of association[c]
SAdV-C (simian adenovirus C)	Culture and Illumina NGS [105]	BLAST search and *de novo* genome assembly	Pneumonia (baboons) Acute respiratory illness (humans)	++ +

[a] Viruses were detected in clinical samples from humans unless otherwise specified.
[b] Discovered previously in 2009 by conventional testing [3,4].
[c] –, unknown or no association observed with the given disease; +, moderate association; ++, strong association.
[d] Viruses are listed in order in which they are first mentioned. Note that table is not comprehensive and only lists viruses that are specifically highlighted in the text. *Abbreviations*: NGS, next-generation sequencing; 454, Roche 454 pyrosequencing platform; PCR, polymerase chain reaction; RCA, rolling circle amplification; subtraction, computational 'digital' subtraction of host background sequences from NGS data; BLAST, basic local alignment search tool; WHIM syndrome, Warts, Hypogammaglobulinemia, Infections, and Myelokathexis syndrome; N/A, not applicable.

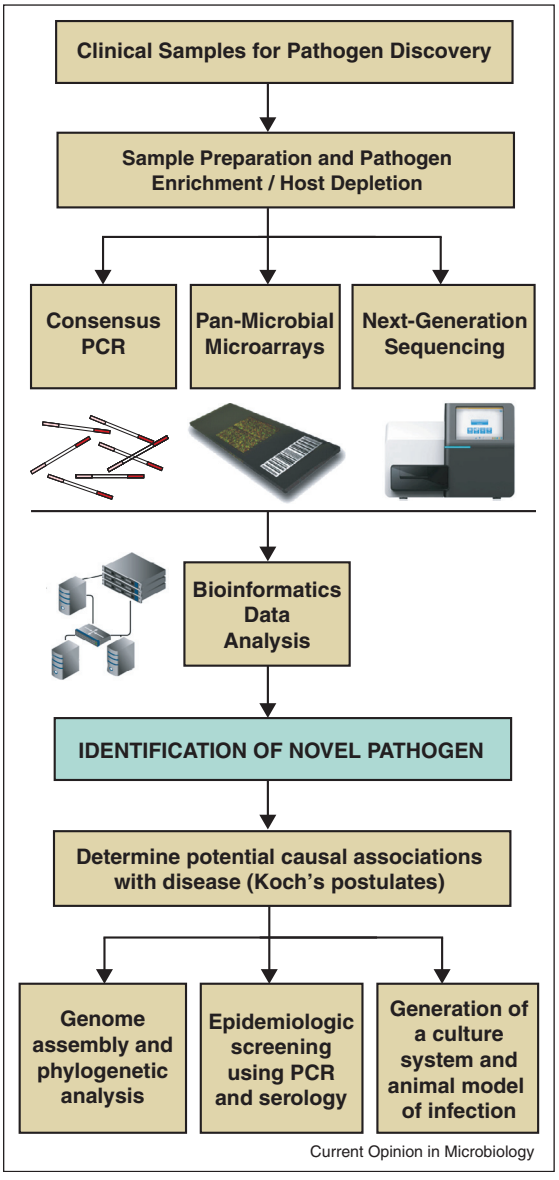
occasionally fatal respiratory disease in the Middle East and Europe [6], although many other examples of this strategy for viral discovery exist [23,24]. However, the identity of an infectious agent is often not known *a priori*, and a random, unbiased, and sequence-independent method for 'universal' amplification becomes necessary for pathogen discovery [25]. In the past, such universal amplification methods have been used in combination with conventional shotgun Sanger sequencing to detect novel human viruses such as human metapneumovirus in respiratory secretions [26], PARV4, a novel parvovirus in blood from patients with acute viral infection syndrome [27], and novel astroviruses, parvoviruses, picornaviruses (cardioviruses and cosaviruses), and polyomaviruses in diarrheal stool [25,28–34]. One caveat with this approach may be the relatively low detection sensitivity of $\sim 10^6$ genome equivalents per milliliter [25]. A related strategy is the use of rolling circle amplification (RCA) [35,36], which has been successful in the unbiased detection and/or characterization of DNA viruses with circular genomes, such as novel papillomaviruses, circoviruses, and polyomaviruses [37–40].

DNA microarrays have been used for multiplexed detection of a defined set of known pathogens using conserved primers [41], or for broad pan-microbial detection by universal amplification [42–44]. Microarrays are miniaturized detection platforms consisting of short (25-mer to 70-mer) single-stranded oligonucleotide probes deposited onto a solid substrate. These probes are typically designed to target conserved sequences at different levels of the taxonomy (family, genus, and species), which allows detection of novel pathogens that share homology with known, previously characterized viruses. Fluorescently labeled clinical samples are hybridized to the microarray, and hybridization patterns are analyzed to identify the specific pathogens that are present (Figure 2a) [43–47].

Pan-microbial DNA microarrays currently in use include the ViroChip (University of California, San Francisco) [42,48], GreeneChip (Columbia University) [43], and the Lawrence Livermore Microbial Detection Array, or LLMDA (Lawrence Livermore National Laboratory) [44]. The ViroChip is a pan-viral DNA microarray and was originally employed to characterize the coronavirus responsible for the 2003 outbreak of SARS [1]. Since then, studies have employed the ViroChip to discover a number of novel viruses including a previously undescribed rhinovirus clade [49], human cardioviruses [50], and 2009 pandemic influenza H1N1 (Figure 2a) [51]. In 2011, the ViroChip was also used to identify a novel adenovirus that caused a fulminant pneumonia outbreak in a New World titi monkey colony, with serologic evidence of concurrent cross-species infection of a human researcher [52]. The GreeneChip is a pan-microbial array that includes ~30k 60-mer probes and is designed to broadly detect all viruses, as well as pathogenic bacteria, fungi, and protozoa on the basis of conserved 16S/18S sequences [43]. The LLMDA is yet another comprehensive pan-microbial detection array that targets all potential pathogens, with probes derived from their full genome sequences [44]. The GreeneChip and LLMDA have been used to detect *Plasmodium falciparum* in a patient with an unknown febrile illness [43] and porcine circovirus as a contaminant in a rotavirus vaccine [53], respectively. Although useful for the detection of a wide spectrum of pathogens, and for the detection of novel strains, microarrays are still limited by the genome sequence information available at the time of design.

NGS, otherwise known as massively parallel or deep sequencing, has emerged as one of the most promising strategies for the detection of novel infectious agents in clinical specimens [54,55]. This 'needle-in-a-haystack' approach involves analysis of millions of sequences derived from nucleic acid present in clinical specimens to detect sequences corresponding to candidate pathogens. Given low amounts of input nucleic acid in clinical samples, an unbiased, random method employing universal amplification is typically performed during NGS library generation [25,56], similar to that used in pan-microbial microarray assays [42]. Because of its unbiased nature, NGS can identify both known but unexpected agents and highly divergent novel agents. NGS is thus particularly attractive for the identification of novel

Figure 1



Genomic approaches to pathogen discovery. Clinical samples are subjected to pathogen enrichment and host depletion methods, followed by genomic analysis using consensus PCR, pan-microbial microarrays, and/or NGS. After a novel agent is identified, downstream studies are needed to establish a causal association between the candidate pathogen and disease.

emerging viruses, which can exhibit high inherent sequence diversity and rapid rates of mutation, recombination, or reassortment [57]. For example, NGS was recently used to identify and recover the genome of a novel, highly divergent rhabdovirus, Bas-Congo virus (BASV), associated with a 2009 hemorrhagic fever outbreak in the Congo, Africa (Figure 3a) [58]. In this study, the genome of BASV was *de novo* assembled from 140 million deep sequencing reads corresponding to an acute serum sample from an affected patient (Figure 3b). The discovery of BASV underscores the potential of NGS in facilitating early identification of pathogens causing unknown outbreaks in remote areas of the world before they gain a foothold in human populations.

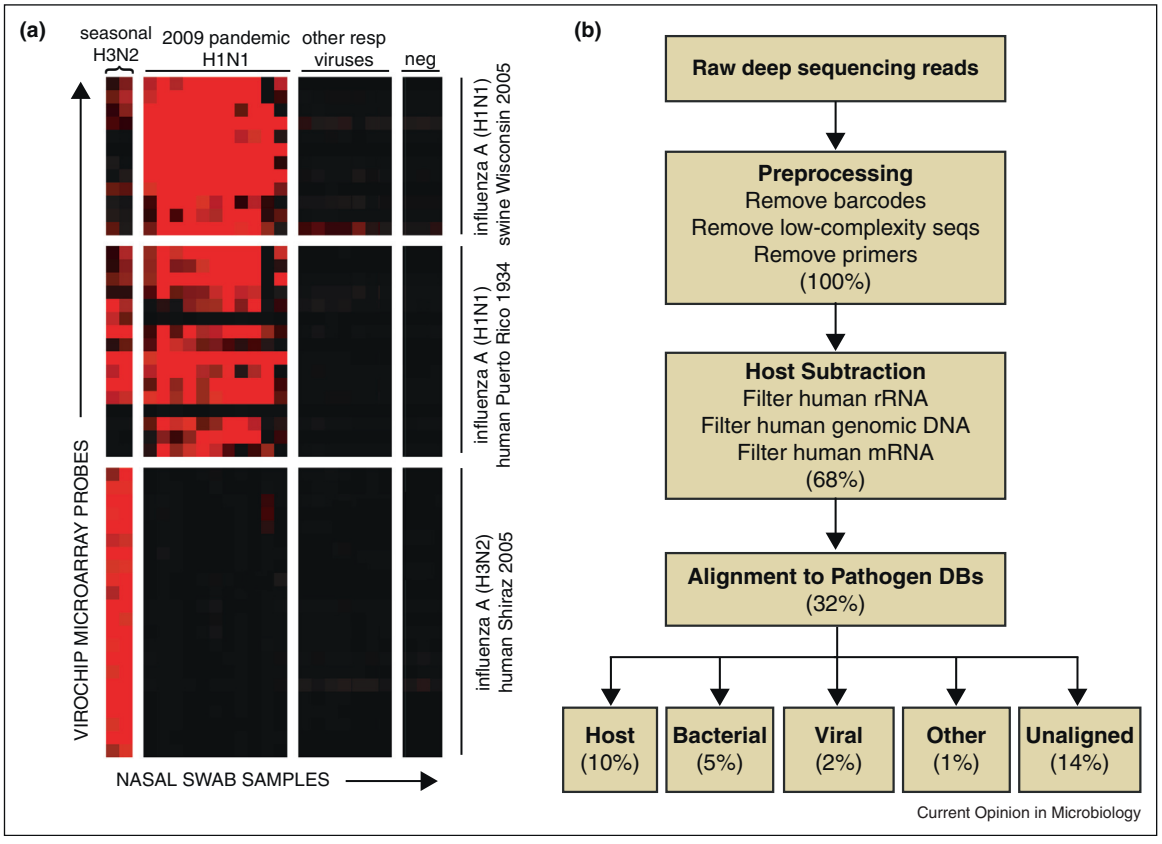
In addition to the identification of BASV, the use of NGS technology has led to the discovery of many novel human viruses over the past decade, including, among others, the aforementioned MCPyV [15]; novel circoviruses/cycloviruses [59], kobuviruses (klassevirus/salivirus) [60–62]; polyomaviruses such as the HPyV9 and MWPyV/HPyV10/MXPyV [63–66]; a novel parvovirus named bufavirus [67]; a novel astrovirus associated with encephalitis [68]; a novel enterovirus species in tropical febrile illness [69]; as well as novel arenaviruses in a fatal outbreak of transplant recipients [70] and a hemorrhagic fever outbreak from South Africa [71]. In 2011, an unknown outbreak of fever and thrombocytopenia involving hundreds of patients occurred in rural China [12,13]. Unbiased NGS of pooled patient serum samples was used by one research group to identify the causal agent as a novel, highly divergent bunyavirus in the *Phlebovirus* genus referred to as Severe Fever and Thrombocytopenia Syndrome (SFTS) virus [12]. Furthermore, NGS has been used to enable whole-genome sequencing and assembly of highly divergent viruses identified from unknown cultures exhibiting cytopathic effect. Heartland virus, a presumed novel tick-borne bunyavirus in the *Phlebovirus* genus associated with two cases of severe febrile illness in hospitalized patients in Missouri [14], and Lone Star virus, another phlebovirus infecting the *Amblyomma americanum* tick [72], were both successfully sequenced from virally infected cell culture supernatants using NGS.

NGS approaches have also been successful in the identification of novel animal viruses, including the discovery of bats, dogs, horses, and rodents as reservoirs for novel flaviviruses (pegiviruses and hepaciviruses distantly related to human hepatitis C) [73–77], a novel bocavirus in canine liver [78], and novel arenaviruses associated with inclusion body disease in snakes [79]. Recently, a novel flavivirus in the *Pegivirus* genus, named Theiler's disease-associated virus (TDAV), was found by NGS to be the likely cause of an mysterious acute hepatitis in horses associated with the administration of equine blood products, a diagnosis that had eluded microbiologists for nearly a century [73]. Finally, infection by non-viral agents, such as *Fusobacterium nucleatum* bacteria in the setting of colon cancer, has also been detected by NGS [80].

Sample preparation methods

Both unbiased NGS, and, to a lesser extent, pan-microbial microarrays are affected by the level of host background, limiting sensitivity for detection of pathogen-derived sequences. In a study using NGS to investigate occult bacterial infection in tissues, microbial sequences were

Figure 2

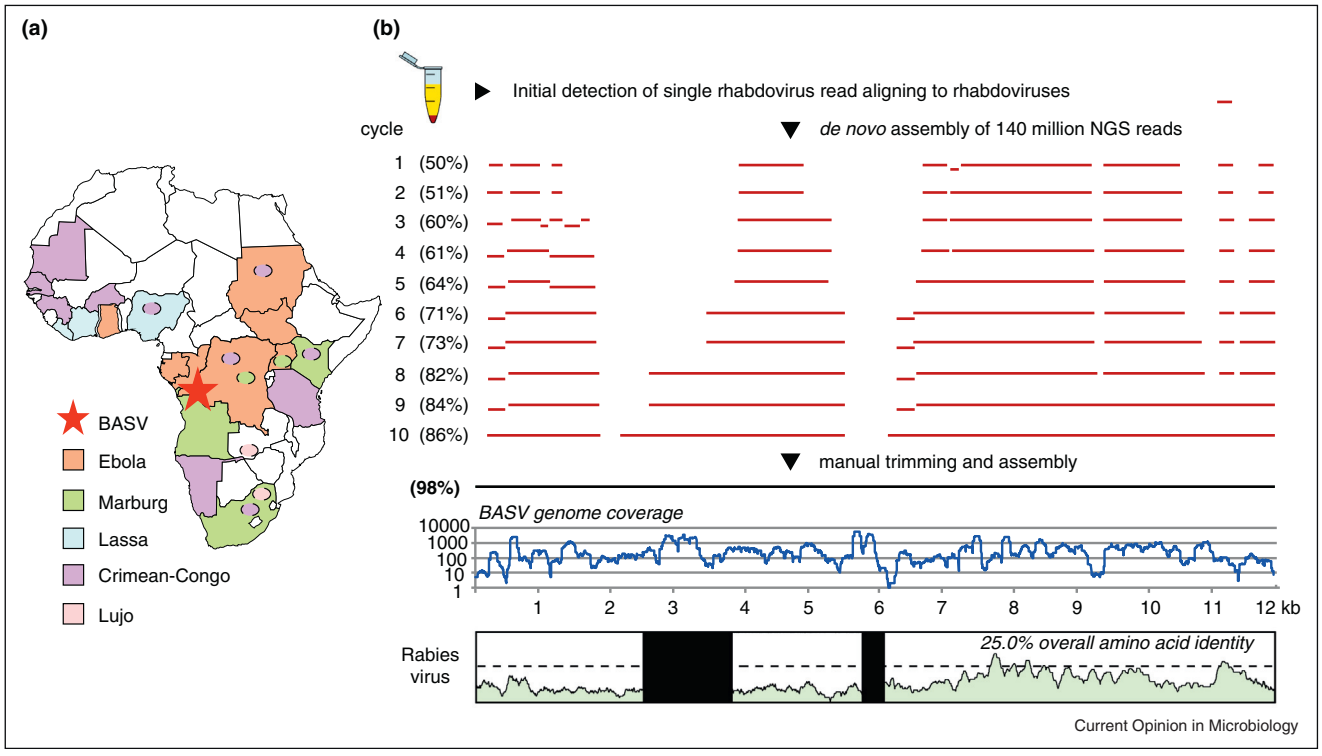


Microarray and NGS analyses of pandemic 2009 influenza A(H1N1) infection in humans. **(a)** Heat map of ViroChip microarray hybridization patterns obtained from nasal swab samples from patients with influenza-like illness and asymptomatic negative controls ('neg'). The samples (*x*-axis) and microarray probes (*y*-axis) are clustered using a hierarchical clustering algorithm [45]. High-intensity probes derived from swine influenza A(H1N1) and human influenza A(H1N1) sequences are observed in samples from patients infected by pandemic 2009 influenza A(H1N1), with higher relative signal intensity in the swine influenza A(H1N1) probes. In contrast, the ViroChip signature in nasal swabs from patients infected with seasonal H3N2 influenza consists primarily of influenza A(H3N2) probes. No microarray cross-hybridization is observed in patients infected with other respiratory viruses or negative controls. Note that the influenza probes on the ViroChip microarray shown here were designed before onset of the pandemic 2009 influenza A(H1N1) outbreak. **(b)** Computational pipeline for analysis of NGS data. Preprocessing and computational subtraction of host (human) sequences are then followed by alignment to pathogen reference databases. The percentages show the remaining proportion of reads after each step, beginning with 100% of the preprocessed reads. *Abbreviations*: DBs, databases; rRNA, ribosomal RNA; mRNA, messenger RNA.
Modified from [51] with permission.

only detected in 0.00067% of NGS reads, corresponding to fewer than 10 per million [80]. Pathogen enrichment or host depletion before microarray and deep sequencing analyses hence becomes critical to maximize sensitivity for identification of novel agents in clinical samples (Figure 1). For viruses, capsid purification procedures involving repeated freeze/thaw cycles, filtration, ultracentrifugation, and prenuclease digestion have been developed to enrich host tissues or body fluids for infectious particles [78,81]. Strategies to deplete the sample of background host DNA can also be implemented, including the use of methylation-specific DNAse to selectively degrade host genomes [82], removal of host ribosomal RNA [83], and/or removal of the most abundant host sequences by duplex-specific nuclease (DSN) normalization [84]. Another complementary approach is to perform target enrichment using biotinylated probes to enrich NGS libraries for sequences corresponding to pathogens, akin to now well-established techniques that have been developed in the cancer field [85]. This strategy can also potentially harness prior experience with microarrays for pathogen discovery by the use of previously validated microarray probes to enrich NGS libraries for microbial sequences.

The choice of NGS platforms on the market today for pathogen discovery is driven by two main parameters: read length and read depth. NGS reads must be long enough (typically at least 100–300 nt) to unambiguously identify the presence of a novel pathogen, and to discriminate reads from host or background flora. There must also be sufficient read depth, or number of sequence

Figure 3



Discovery of Bas-Congo virus (BASV), a novel rhabdovirus associated with acute hemorrhagic fever in humans. **(a)** Map of Africa showing viral hemorrhagic fever outbreak regions. Hemorrhagic fever due to flaviviruses, such as dengue and yellow fever, is widespread throughout the continent. The location of the BASV hemorrhagic fever outbreak is designated by a red star. **(b)** Deep sequencing and *de novo* genome assembly of BASV. The BASV genome is highly divergent, sharing only 25% amino acid identity with rabies and <42% amino acid identity with any other rhabdovirus. Modified from [58] with permission.

reads generated per run, to detect novel agents with a high degree of sensitivity. For pathogen discovery, the Roche 454 GS-FLX+pyrosequencing^{TM} platform has been widely applied given the long reads (currently up to 1 million single or paired-end reads with average read lengths of 400–500 nt with the GS-FLX+ Titanium^{TM} platform) and high accuracy. More recently, Illumina NGS sequencing platforms (GAIIx^{TM}, HiSeq^{TM}, and MiSeq^{TM}) have been used for pathogen discovery given the ~10–1000× improved read depth relative to 454, resulting in much greater sensitivity for the detection of viruses [86], and gradually improving read lengths (currently up to 150 nt paired-end reads for the HiSeq and 250 nt paired-end reads for the MiSeq). In fact, previous studies suggest that the limits of detection of viruses in clinical samples by NGS with Illumina sequencing are comparable to specific PCR [51,86]. The use of paired-end sequencing, or sequencing from each end of the DNA fragment in NGS libraries, can be particularly useful for pathogen discovery given that the forward and reverse reads can facilitate the design of PCR primers to confirm potential sequence 'hits' to novel microbes and *de novo* genome assembly [87]. Other NGS technologies, such as platforms by Ion Torrent (very fast run times of under three hours) and Pacific Biosciences (very long reads of up to 7 kb; average read lengths 3–4 kb) [88], have yet to be used widely for pathogen discovery, although one application may be rapid genome sequencing of emerging pathogens such as *Escherichia coli* O104:H4, associated with a recent foodborne outbreak of hemolytic-uremic syndrome in Germany [89,90]. One particular concern for all unbiased NGS technologies is the high potential for reagent and laboratory contamination, especially with the use of universal amplification methods [51,86,91].

Bioinformatics analysis challenges

Whereas for microarrays, specialized bioinformatics algorithms for pathogen detection are in routine use [43–47], analysis of NGS data for pathogen discovery poses enormous computational challenges. The most widely used strategy is computational subtraction, in which reads are first sequentially aligned to reference databases to filter out sequences corresponding to host background [92]. Sequences derived from microbes are then typically identified by nucleotide or translated amino acid alignments

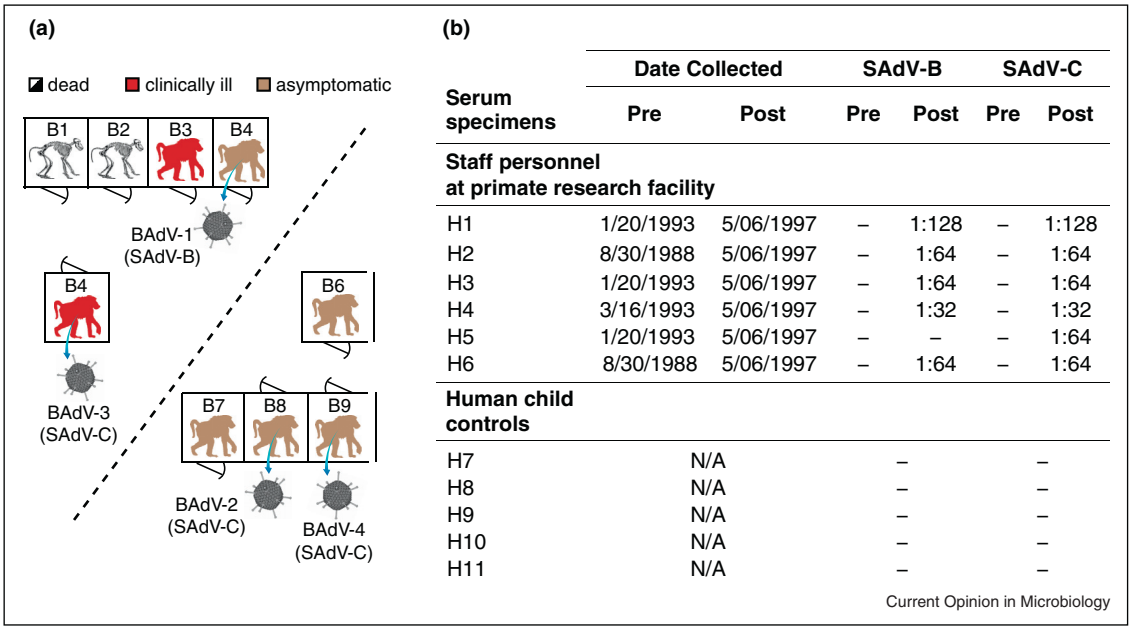
using BLAST [93]. This approach was previously used, for example, to detect pandemic 2009 influenza A(H1N1) in nasal swabs from affected patients with respiratory illness (Figure 2b) [51]. For highly divergent viruses, successful identification can sometimes only be made by searching for remote homologs of protein sequences using methods such as HMMER [94,95]. Dedicated bioinformatics analysis pipelines, such as PathSeq, used to detect *Fusobacterium* bacteria in colon cancer tissues [80], RINS, CaPSID, and READSCAN are now available for automated pathogen identification from NGS data [96–99], although their performance has yet to be rigorously tested on a large number of clinical samples. Ongoing limitations of available bioinformatics software for pathogen discovery include the data-intensive computing workloads that are not amenable to real-time analysis in the absence of ultra-rapid processing algorithms, the lack of a graphical user interface, the requirement for a minimum level of computer hardware and bioinformatics expertise, and the lack of a validated scoring system to permit confident identification of microbes from NGS data. In addition, existing reference sequence databases, such as NIH GenBank, can be heavily baised and fraught with annotation errors. Notably, over 40% of the GenBank viral database consists of overrepresented HIV or influenza sequences. Comprehensive, well-annotated reference databases for pathogens are thus needed in support of NGS-based pathogen discovery efforts.

Linking a novel pathogen to disease

The mere discovery of a candidate pathogen is only the first step in determining whether or not it is associated with disease. Clinical samples are colonized with a variety of commensal organisms (the 'microbiome') [100], and it is often difficult, if not impossible, to unambiguously identify a single causal infectious agent. Highly divergent, novel agents such as torque teno virus (TTV) [101,102] may be nonpathogenic and part of the normal microbial flora. Follow-up studies to establish causality are thus needed to establish a link between a candidate infectious agent and disease (Figure 1).

To assign causality, attempts should be made to address Koch's postulates, which require that the agent be isolated in culture, or River's modifications, which recognize the added significance of the generation of specific antibodies in response to infection [103]. For novel viruses, this begins with assembly of the entire genome, either *de novo* directly from NGS data [58,72,87] or by standard methods such as primer walking, probe enrichment [104], and/or specific PCR to fill in gaps [52]. Full or partial genomic sequence permits a detailed phylogenetic

Figure 4



Serum specimens	Date Collected		SAdV-B		SAdV-C	
	Pre	Post	Pre	Post	Pre	Post
Staff personnel at primate research facility						
H1	1/20/1993	5/06/1997	–	1:128	–	1:128
H2	8/30/1988	5/06/1997	–	1:64	–	1:64
H3	1/20/1993	5/06/1997	–	1:64	–	1:64
H4	3/16/1993	5/06/1997	–	1:32	–	1:32
H5	1/20/1993	5/06/1997	–	–	–	1:64
H6	8/30/1988	5/06/1997	–	1:64	–	1:64
Human child controls						
H7	N/A		–		–	
H8	N/A		–		–	
H9	N/A		–		–	
H10	N/A		–		–	
H11	N/A		–		–	

Baboon and human infections from a novel adenovirus species. **(a)** A 1997 acute respiratory outbreak in a baboon colony. A novel adenovirus, named simian adenovirus C (SAdV-C), was discovered in association with an outbreak at a primate research facility that sickened 4 of 9 baboons and resulted in two cases of fatal pneumonia. **(b)** Serological testing of staff personnel at the facility and controls (five epidemiologically unrelated young children) for exposure to simian adenoviruses SAdV-B and SAdV-C. Neutralizing antibodies to SAdV-C are absent before the outbreak but detected in 6 of 6 staff personnel after the outbreak, indicating recent or prior exposure to the virus. *Abbreviations*: BAdV, baboon adenovirus; SAdV, simian adenovirus; Pre, pre-outbreak; Post, post-outbreak; N/A, not applicable.Modified from [105] with permission.

analysis of the novel agent, which can provide clues as to its potential host range and pathogenicity [58]. The availability of sequence information also facilitates the development of specific PCR-based or serological assays for detection. Epidemiological screening of the distribution of the candidate pathogen in diseased patients and asymptomatic controls by PCR, as well as assessment of the geographic and temporal distribution of infections, can help in establishing a link to disease. Serology can also play a critical role in determining pathogenicity, as increases in titer support the association of a given pathogen with infection. For example, serologic analyses of a novel adenovirus species named 'simian adenovirus C (SAdV-C)' associated with a pneumonia outbreak in a baboon colony (Figure 4a) were recently used to establish that staff personnel at the facility had also been exposed to this newly discovered virus (Figure 4b) [105]. Finally, development of a culture system and animal model for infection can directly confirm that a candidate novel agent plays a causal role in disease.

One advantage of using microarrays and NGS for pathogen discovery is that these same technologies can also be applied to evaluate the potential pathogenicity of newly identified novel agents. Host transcriptome analysis using gene expression microarrays [106] or RNA-Seq [107] can enable the characterization of associated host biomarkers in response to infection. Detailed NGS-based quasispecies analysis of novel pathogens that exhibit high mutation rates, such as RNA viruses [108,109], can also provide insights into how these agents infect and invade the host.

Conclusions

Although sometimes derided as a 'fishing expedition', pathogen discovery is, in actuality, a highly worthwhile scientific endeavor. Without a cause identified for many presumed infectious diseases, it is not possible to conduct downstream investigations in pathogenesis and host–microbial interactions, nor is it possible to design effective vaccines or antimicrobial drugs to combat the associated illness. Potential applications of pathogen discovery range from outbreak investigation of emerging pathogens, to screening of blood products, vaccines, and other biologics for viral contaminants, to clinical diagnosis of unknown acute or chronic infectious diseases. The current availability of state-of-the-art genomic technologies such as pan-microbial microarrays and NGS provides an unprecedented opportunity to 'cast a wide net' and survey the full breadth of as-yet undiscovered pathogens in nature that pose significant threats to human health.

Competing interests statement

The author's research on viral pathogen discovery is partially supported by an award by Abbott Laboratories, Inc. The author has also filed provisional patent applications related to Lone Star virus, a novel bunyavirus in the *Amblyomma americanum* tick, and the novel baboon SAdV-C adenoviruses referred to in this article.

Acknowledgements

The author thanks Drs. Eric Delwart and Jerome Bouquet for thoughtful comments and the U.S. National Institutes of Health (grants R56-AI08953 and R01-HL105704), UC MEXUS-CONACYT Collaborative Grants Program, and Abbott Laboratories, Inc. for research funding and support.

References

1. Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, Penaranda S, Bankamp B, Maher K, Chen MH *et al.*: **Characterization of a novel coronavirus associated with severe acute respiratory syndrome**. *Science* 2003, **300**:1394-1399.
2. Nichol ST, Spiropoulou CF, Morzunov S, Rollin PE, Ksiazek TG, Feldmann H, Sanchez A, Childs J, Zaki S, Peters CJ: **Genetic identification of a hantavirus associated with an outbreak of acute respiratory illness**. *Science* 1993, **262**:914-917.
3. Dawood FS, Jain S, Finelli L, Shaw MW, Lindstrom S, Garten RJ, Gubareva LV, Xu X, Bridges CB, Uyeki TM: **Emergence of a novel swine-origin influenza A (H1N1) virus in humans**. *N Engl J Med* 2009, **360**:2605-2615.
4. Shinde V, Bridges CB, Uyeki TM, Shu B, Balish A, Xu X, Lindstrom S, Gubareva LV, Deyde V, Garten RJ *et al.*: **Triple-reassortant swine influenza A (H1) in humans in the United States, 2005–2009**. *N Engl J Med* 2009, **360**:2616-2625.
5. Bermingham A, Chand MA, Brown CS, Aarons E, Tong C, Langrish C, Hoschler K, Brown K, Galiano M, Myers R *et al.*: **Severe respiratory illness caused by a novel coronavirus, in a patient transferred to the United Kingdom from the Middle East, September 2012**. *Euro Surveill* 2012, **17**:20290.
6. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA: **Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia**. *N Engl J Med* 2012, **367**:1814-1820.
7. van Boheemen S, de Graaf M, Lauber C, Bestebroer TM, Raj VS, Zaki AM, Osterhaus AD, Haagmans BL, Gorbalenya AE, Snijder EJ *et al.*: **Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans**. *MBio* 2012, **3**.
8. Gao R, Cao B, Hu Y, Feng Z, Wang D, Hu W, Chen J, Jie Z, Qiu H, Xu K *et al.*: **Human infection with a novel avian-origin influenza A (H7N9) virus**. *N Engl J Med* 2013. [Epub ahead of print].
9. Chomel BB, Belotto A, Meslin FX: **Wildlife exotic pets, and emerging zoonoses**. *Emerg Infect Dis* 2007, **13**:6-11.
10. Briese T, Jia XY, Huang C, Grady LJ, Lipkin WI: **Identification of a Kunjin/West Nile-like flavivirus in brains of patients with New York encephalitis**. *Lancet* 1999, **354**:1261-1262.
11. Lanciotti RS, Roehrig JT, Deubel V, Smith J, Parker M, Steele K, Crise B, Volpe KE, Crabtree MB, Scherret JH *et al.*: **Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States**. *Science* 1999, **286**:2333-2337.
12. Xu B, Liu L, Huang X, Ma H, Zhang Y, Du Y, Wang P, Tang X, Wang H, Kang K *et al.*: **Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus**. *PLoS Pathog* 2011, **7**:e1002369.
13. Yu XJ, Liang MF, Zhang SY, Liu Y, Li JD, Sun YL, Zhang L, Zhang QF, Popov VL, Li C *et al.*: **Fever with thrombocytopenia associated with a novel bunyavirus in China**. *N Engl J Med* 2011, **364**:1523-1532.
14. McMullan LK, Folk SM, Kelly AJ, MacNeil A, Goldsmith CS, Metcalfe MG, Batten BC, Albarino CG, Zaki SR, Rollin PE *et al.*: **A new phlebovirus associated with severe febrile illness in Missouri**. *N Engl J Med* 2012, **367**:834-841.

15. Feng H, Shuda M, Chang Y, Moore PS: **Clonal integration of a polyomavirus in human Merkel cell carcinoma**. *Science* 2008, **319**:1096-1100.

16. Bloch KC, Glaser C: **Diagnostic approaches for patients with suspected encephalitis**. *Curr Infect Dis Rep* 2007, **9**:315-322.

17. Glaser CA, Gilliam S, Schnurr D, Forghani B, Honarmand S, Khetsuriani N, Fischer M, Cossen CK, Anderson LJ: **In search of encephalitis etiologies: diagnostic challenges in the California Encephalitis Project, 1998–2000**. *Clin Infect Dis* 2003, **36**:731-742.

18. Glaser CA, Honarmand S, Anderson LJ, Schnurr DP, Forghani B, Cossen CK, Schuster FL, Christie LJ, Tureen JH: **Beyond viruses: clinical profiles and etiologies associated with encephalitis**. *Clin Infect Dis* 2006, **43**:1565-1577.

19. Prusiner SB: **Prions**. *Sci Am* 1984, **251**:50-59.

20. Wilkinson DA, Temmam S, Lebarbenchon C, Lagadec E, Chotte J, Guillebaud J, Ramasindrazana B, Heraud JM, de Lamballerie X, Goodman SM *et al.*: **Identification of novel paramyxoviruses in insectivorous bats of the Southwest Indian Ocean**. *Virus Res* 2012, **170**:159-163.

21. Drexler JF, Corman VM, Muller MA, Maganga GD, Vallo P, Binger T, Gloza-Rausch F, Rasche A, Yordanov S, Seebens A *et al.*: **Bats host major mammalian paramyxoviruses**. *Nat Commun* 2012, **3**:796.

22. Kurth A, Kohl C, Brinkmann A, Ebinger A, Harper JA, Wang LF, Muhldorfer K, Wibbelt G: **Novel paramyxoviruses in free-ranging European bats**. *PLoS ONE* 2012, **7**:e38688.

23. Kapoor A, Li L, Victoria J, Oderinde B, Mason C, Pandey P, Zaidi SZ, Delwart E: **Multiple novel astrovirus species in human stool**. *J Gen Virol* 2009, **90**:2965-2972.

24. Scuda N, Hofmann J, Calvignac-Spencer S, Ruprecht K, Liman P, Kuhn J, Hengel H, Ehlers B: **A novel human polyomavirus closely related to the African green monkey-derived lymphotropic polyomavirus**. *J Virol* 2011, **85**:4586-4590.

25. 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 U S A* 2001, **98**:11609-11614.

26. van den Hoogen BG, de Jong JC, Groen J, Kuiken T, de Groot R, Fouchier RA, Osterhaus AD: **A newly discovered human pneumovirus isolated from young children with respiratory tract disease**. *Nat Med* 2001, **7**:719-724.

27. Jones MS, Kapoor A, Lukashov VV, Simmonds P, Hecht F, Delwart E: **New DNA viruses identified in patients with acute viral infection syndrome**. *J Virol* 2005, **79**:8230-8236.

28. Allander T, Andreasson K, Gupta S, Bjerkner A, Bogdanovic G, Persson MA, Dalianis T, Ramqvist T, Andersson B: **Identification of a third human polyomavirus**. *J Virol* 2007, **81**:4130-4136.

29. Jones MS, Lukashov VV, Ganac RD, Schnurr DP: **Discovery of a novel human picornavirus in a stool sample from a pediatric patient presenting with fever of unknown origin**. *J Clin Microbiol* 2007, **45**:2144-2150.

30. Kapoor A, Victoria J, Simmonds P, Slikas E, Chieochansin T, Naeem A, Shaukat S, Sharif S, Alam MM, Angez M *et al.*: **A highly prevalent and genetically diversified Picornaviridae genus in South Asian children**. *Proc Natl Acad Sci U S A* 2008, **105**:20482-20487.

31. Gaynor AM, Nissen MD, Whiley DM, Mackay IM, Lambert SB, Wu G, Brennan DC, Storch GA, Sloots TP, Wang D: **Identification of a novel polyomavirus from patients with acute respiratory tract infections**. *PLoS Pathog* 2007, **3**:e64.

32. Arthur JL, Higgins GD, Davidson GP, Givney RC, Ratcliff RM: **A novel bocavirus associated with acute gastroenteritis in Australian children**. *PLoS Pathog* 2009, **5**:e1000391.

33. Finkbeiner SR, Allred AF, Tarr PI, Klein EJ, Kirkwood CD, Wang D: **Metagenomic analysis of human diarrhea: viral detection and discovery**. *PLoS Pathog* 2008, **4**:e1000011.

34. Kapoor A, Slikas E, Simmonds P, Chieochansin T, Naeem A, Shaukat S, Alam MM, Sharif S, Angez M, Zaidi S *et al.*: **A newly identified bocavirus species in human stool**. *J Infect Dis* 2009, **199**:196-200.

35. Rector A, Tachezy R, Van Ranst M: **A sequence-independent strategy for detection and cloning of circular DNA virus genomes by using multiply primed rolling-circle amplification**. *J Virol* 2004, **78**:4993-4998.

36. Dean FB, Nelson JR, Giesler TL, Lasken RS: **Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification**. *Genome Res* 2001, **11**:1095-1099.

37. Niel C, Diniz-Mendes L, Devalle S: **Rolling-circle amplification of Torque teno virus (TTV) complete genomes from human and swine sera and identification of a novel swine TTV genogroup**. *J Gen Virol* 2005, **86**:1343-1347.

38. Dela Cruz FN Jr, Giannitti F, Li L, Woods LW, Del Valle L, Delwart E, Pesavento PA: **Novel polyomavirus associated with brain tumors in free-ranging raccoons, western United States**. *Emerg Infect Dis* 2013, **19**:77-84.

39. Schowalter RM, Pastrana DV, Pumphrey KA, Moyer AL, Buck CB: **Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin**. *Cell Host Microbe* 2010, **7**:509-515.

40. van der Meijden E, Janssens RW, Lauber C, Bouwes Bavinck JN, Gorbalenya AE, Feltkamp MC: **Discovery of a new human polyomavirus associated with trichodysplasia spinulosa in an immunocompromized patient**. *PLoS Pathog* 2010, **6**:e1001024.

41. Mikhailovich V, Gryadunov D, Kolchinsky A, Makarov AA, Zasedatelev A: **DNA microarrays in the clinic: infectious diseases**. *Bioessays* 2008, **30**:673-682.

42. 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 U S A* 2002, **99**:15687-15692.

43. Palacios G, Quan PL, Jabado OJ, Conlan S, Hirschberg DL, Liu Y, Zhai J, Renwick N, Hui J, Hegyi H *et al.*: **Panmicrobial oligonucleotide array for diagnosis of infectious diseases**. *Emerg Infect Dis* 2007, **13**:73-81.

44. Gardner SN, Jaing CJ, McLoughlin KS, Slezak TR: **A microbial detection array (MDA) for viral and bacterial detection**. *BMC Genomics* 2010, **11**:668.

45. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**:14863-14868.

46. Allred AF, Wu G, Wulan T, Fischer KF, Holbrook MR, Tesh RB, Wang D: **VIPR: a probabilistic algorithm for analysis of microbial detection microarrays**. *BMC Bioinformatics* 2010, **11**:384.

47. Urisman A, Fischer KF, Chiu CY, Kistler AL, Beck S, Wang D, DeRisi JL: **E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns**. *Genome Biol* 2005, **6**:R78.

48. Chen EC, Miller SA, DeRisi JL, Chiu CY: **Using a pan-viral microarray assay (Virochip) to screen clinical samples for viral pathogens**. *J Vis Exp* 2011. 50, [Epub ahead of print].

49. Kistler A, Avila PC, Rouskin S, Wang D, Ward T, Yagi S, Schnurr D, Ganem D, DeRisi JL, Boushey HA: **Pan-viral screening of respiratory tract infections in adults with and without asthma reveals unexpected human coronavirus and human rhinovirus diversity**. *J Infect Dis* 2007, **196**:817-825.

50. Chiu CY, Greninger AL, Kanada K, Kwok T, Fischer KF, Runckel C, Louie JK, Glaser CA, Yagi S, Schnurr DP *et al.*: **Identification of cardioviruses related to Theiler's murine encephalomyelitis virus in human infections**. *Proc Natl Acad Sci U S A* 2008, **105**:14124-14129.

51. Greninger AL, Chen EC, Sittler T, Scheinerman A, Roubinian N, Yu G, Kim E, Pillai DR, Guyard C, Mazzulli T *et al.*: **A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America**. *PLoS ONE* 2010, **5**:e13381.

52. Chen EC, Yagi S, Kelly KR, Mendoza SP, Tarara RP, Canfield DR, Maninger N, Rosenthal A, Spinner A, Bales KL *et al.*: **Cross-species**

transmission of a novel adenovirus associated with a fulminant pneumonia outbreak in a new world monkey colony. *PLoS Pathog* 2011, **7**:e1002155.

53. Victoria JG, Wang C, Jones MS, Jaing C, McLoughlin K, Gardner S, Delwart EL: **Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus**. *J Virol* 2010, **84**:6033-6040.

54. 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.

55. Tang P, Chiu C: **Metagenomics for the discovery of novel human viruses**. *Future Microbiol* 2010, **5**:177-189.

56. 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.

57. Holland J, Spindler K, Horodyski F, Grabau E, Nichol S, VandePol S: **Rapid evolution of RNA genomes**. *Science* 1982, **215**:1577-1585.

58. Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe J-J, Sittler T, Veeraraghavan N, Ruby G, Wang C *et al.*: **A novel rhabdovirus associated with acute hemorrhagic fever in Central Africa**. *PloS Pathogens* 2012, **8**:e1002924.

59. Li L, Kapoor A, Slikas B, Bamidele OS, Wang C, Shaukat S, Masroor MA, Wilson ML, Ndjango JB, Peeters M *et al.*: **Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces**. *J Virol* 2010, **84**:1674-1682.

60. Holtz LR, Finkbeiner SR, Zhao G, Kirkwood CD, Girones R, Pipas JM, Wang D: **Klassevirus 1, a previously undescribed member of the family Picornaviridae, is globally widespread**. *Virol J* 2009, **6**:86.

61. Greninger AL, Runckel C, Chiu CY, Haggerty T, Parsonnet J, Ganem D, DeRisi JL: **The complete genome of klassevirus — a novel picornavirus in pediatric stool**. *Virol J* 2009, **6**:82.

62. Li L, Victoria J, Kapoor A, Blinkova O, Wang C, Babrzadeh F, Mason CJ, Pandey P, Triki H, Bahri O *et al.*: **A novel picornavirus associated with gastroenteritis**. *J Virol* 2009, **83**:12002-12006.

63. Siebrasse EA, Reyes A, Lim ES, Zhao G, Mkakosya RS, Manary MJ, Gordon JI, Wang D: **Identification of MW polyomavirus, a novel polyomavirus in human stool**. *J Virol* 2012, **86**:10321-10326.

64. Yu G, Greninger AL, Isa P, Phan TG, Martinez MA, de la Luz Sanchez M, Contreras JF, Santos-Preciado JI, Parsonnet J, Miller S *et al.*: **Discovery of a novel polyomavirus in acute diarrheal samples from children**. *PLoS ONE* 2012, **7**:e49449.

65. Buck CB, Phan GQ, Raiji MT, Murphy PM, McDermott DH, McBride AA: **Complete genome sequence of a tenth human polyomavirus**. *J Virol* 2012, **86**:10887.

66. Sauvage V, Foulongne V, Cheval J, Ar Gouilh M, Pariente K, Dereure O, Manuguerra JC, Richardson J, Lecuit M, Burguiere A *et al.*: **Human polyomavirus related to African green monkey lymphotropic polyomavirus**. *Emerg Infect Dis* 2011, **17**:1364-1370.

67. Phan TG, Vo NP, Bonkoungou IJ, Kapoor A, Barro N, O'Ryan M, Kapusinszky B, Wang C, Delwart E: **Acute diarrhea in West African children: diverse enteric viruses and a novel parvovirus genus**. *J Virol* 2012, **86**:11024-11030.

68. Quan PL, Wagner TA, Briese T, Torgerson TR, Hornig M, Tashmukhamedova A, Firth C, Palacios G, Baisre-De-Leon A, Paddock CD *et al.*: **Astrovirus encephalitis in boy with X-linked agammaglobulinemia**. *Emerg Infect Dis* 2010, **16**:918-925.

69. Yozwiak NL, Skewes-Cox P, Gordon A, Saborio S, Kuan G, Balmaseda A, Ganem D, Harris E, DeRisi JL: **Human enterovirus 109: a novel interspecies recombinant enterovirus isolated from a case of acute pediatric respiratory illness in Nicaragua**. *J Virol* 2010, **84**:9047-9058.

70. Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, Conlan S, Quan PL, Hui J, Marshall J *et al.*: **A new arenavirus in a cluster of fatal transplant-associated diseases**. *N Engl J Med* 2008, **358**:991-998.

71. Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, Palacios G, Khristova ML, Weyer J, Swanepoel R, Egholm M *et al.*: **Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa**. *PLoS Pathog* 2009, **5**:e1000455.

72. Swei A, Russell BJ, Naccache SN, Kabre B, Veeraraghavan N, Pilgard MA, Johnson BJ, Chiu CY: **The genome sequence of Lone Star virus, a highly divergent bunyavirus found in the Amblyomma americanum tick**. *PLoS ONE* 2013, **8**:e62083.

73. Chandriani S, Skewes-Cox P, Zhong W, Ganem DE, Divers TJ, Van Blaricum AJ, Tennant BC, Kistler AL: **Identification of a previously undescribed divergent virus from the Flaviviridae family in an outbreak of equine serum hepatitis**. *Proc Natl Acad Sci U S A* 2013, **110**:E1407-E1415.

74. Kapoor A, Simmonds P, Cullen JM, Scheel T, Medina JL, Giannitti F, Nishiuchi E, Brock KV, Burbelo PD, Rice CM *et al.*: **Identification of a pegivirus (GBV-like virus) that infects horses**. *J Virol* 2013. [Epub ahead of print].

75. Kapoor A, Simmonds P, Scheel TK, Hjelle B, Cullen JM, Burbelo PD, Chauhan LV, Duraisamy R, Sanchez Leon M, Jain K *et al.*: **Identification of rodent homologs of hepatitis C virus and pegiviruses**. *MBio* 2013, **4**.

76. Kapoor A, Simmonds P, Gerold G, Qaisar N, Jain K, Henriquez JA, Firth C, Hirschberg DL, Rice CM, Shields S *et al.*: **Characterization of a canine homolog of hepatitis C virus**. *Proc Natl Acad Sci U S A* 2011, **108**:11608-11613.

77. Quan PL, Firth C, Conte JM, Williams SH, Zambrana-Torrelio CM, Anthony SJ, Ellison JA, Gilbert AT, Kuzmin IV, Niezgoda M *et al.*: **Bats are a major natural reservoir for hepaciviruses and pegiviruses**. *Proc Natl Acad Sci U S A* 2013. [Epub ahead of print].

78. Li L, Pesavento PA, Leutenegger CM, Estrada M, Coffey LL, Naccache SN, Samayoa E, Chiu C, Qiu J, Wang C *et al.*: **A novel bocavirus in canine liver**. *Virol J* 2013, **10**:54.

79. Stenglein MD, Sanders C, Kistler AL, Ruby JG, Franco JY, Reavill DR, Dunker F, Derisi JL: **Identification, characterization, and in vitro culture of highly divergent arenaviruses from boa constrictors and annulated tree boas: candidate etiological agents for snake inclusion body disease**. *MBio* 2012, **3**:e00180-00112.

80. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J *et al.*: **Genomic analysis identifies association of Fusobacterium with colorectal carcinoma**. *Genome Res* 2012, **22**:292-298.

81. Delwart EL: **Viral metagenomics**. *Rev Med Virol* 2007, **17**:115-131.

82. Oyola SO, Gu Y, Manske M, Otto TD, O'Brien J, Alcock D, Macinnis B, Berriman M, Newbold CI, Kwiatkowski DP *et al.*: **Efficient depletion of host DNA contamination in malaria clinical sequencing**. *J Clin Microbiol* 2013, **51**:745-751.

83. He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG, Wang Z, Chen F, Lindquist EA, Sorek R *et al.*: **Validation of two ribosomal RNA removal methods for microbial metatranscriptomics**. *Nat Methods* 2010, **7**:807-812.

84. 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.

85. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C *et al.*: **Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing**. *Nat Biotechnol* 2009, **27**:182-189.

86. Cheval J, Sauvage V, Frangeul L, Dacheux L, Guigon G, Dumey N, Pariente K, Rousseaux C, Dorange F, Berthet N *et al.*: **Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples**. *J Clin Microbiol* 2011, **49**:3268-3275.

87. Ruby JG, Bellare P, Derisi JL: **PRICE: software for the targeted assembly of components of (meta)genomic sequence data**. *G3 (Bethesda)* 2013. [Epub ahead of print].

88. 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.

89. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D *et al.*: **Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany**. *N Engl J Med* 2011, **365**:709-717.

90. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W *et al.*: **Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology**. *PLoS ONE* 2011, **6**:e22751.

91. Lysholm F, Wetterbom A, Lindau C, Darban H, Bjerkner A, Fahlander K, Lindberg AM, Persson B, Allander T, Andersson B: **Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing**. *PLoS ONE* 2012, **7**:e30875.

92. Weber G, Shendure J, Tanenbaum DM, Church GM, Meyerson M: **Identification of foreign gene sequences by transcript filtering against the human genome**. *Nat Genet* 2002, **30**:141-142.

93. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403-410.

94. Eddy SR: **Accelerated Profile HMM Searches**. *PLoS Comput Biol* 2011, **7**:e1002195.

95. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching**. *Nucleic Acids Res* 2011, **39**:W29-W37.

96. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M: **PathSeq: software to identify or discover microbes by deep sequencing of human tissue**. *Nat Biotechnol* 2011, **29**:393-396.

97. Borozan I, Wilson S, Blanchette P, Laflamme P, Watt SN, Krzyzanowski PM, Sircoulomb F, Rottapel R, Branton PE, Ferretti V: **CaPSID: a bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes**. *BMC Bioinformatics* 2012, **13**:206.

98. Naeem R, Rashid M, Pain A: **READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation**. *Bioinformatics* 2013, **29**:391-392.

99. Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA: **Rapid identification of non-human sequences in high-throughput sequencing datasets**. *Bioinformatics* 2012, **28**:1174-1175.

100. **Structure function and diversity of the healthy human microbiome**. *Nature* 2012, **486**:207-214.

101. 1Gimenez-Barcons M, Forns X, Ampurdanes S, Guilera M, Soler M, Soguero C, Sanchez-Fueyo A, Mas A, Bruix J, Sanchez-Tapias JM *et al.*: **Infection with a novel human DNA virus (TTV) has no pathogenic significance in patients with liver diseases**. *J Hepatol* 1999, **30**:1028-1034.

102. Okamoto H, Takahashi M, Nishizawa T, Ukita M, Fukuda M, Tsuda F, Miyakawa Y, Mayumi M: **Marked genomic heterogeneity and frequent mixed infection of TT virus demonstrated by PCR with primers from coding and noncoding regions**. *Virology* 1999, **259**:428-436.

103. Fredericks DN, Relman DA: **Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates**. *Clin Microbiol Rev* 1996, **9**:18-33.

104. Wang D, Urisman A, Liu YT, Springer M, Ksiazek TG, Erdman DD, Mardis ER, Hickenbotham M, Magrini V, Eldred J *et al.*: **Viral discovery and sequence recovery using DNA microarrays**. *PLoS Biol* 2003, **1**:E2.

105. Chiu CY, Yagi S, Lu X, Yu G, Chen EC, Liu M, Dick EJ Jr, Carey KD, Erdman DD, Leland MM *et al.*: **A novel adenovirus species associated with an acute respiratory outbreak in a baboon colony and evidence of coincident human infection**. *mBio* 2013, **4** e00084-00013.

106. Ekins R, Chu FW: **Microarrays: their origins and applications**. *Trends Biotechnol* 1999, **17**:217-218.

107. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**:57-63.

108. Beerenwinkel N, Gunthard HF, Roth V, Metzner KJ: **Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data**. *Front Microbiol* 2012, **3**:329.

109. Yin L, Liu L, Sun Y, Hou W, Lowe AC, Gardner BP, Salemi M, Williams WB, Farmerie WG, Sleasman JW *et al.*: **High-resolution deep sequencing reveals biodiversity, population structure, and persistence of HIV-1 quasispecies within host ecosystems**. *Retrovirology* 2012, **9**:108.

110. Kuroda M, Katano H, Nakajima N, Tobiume M, Ainai A, Sekizuka T, Hasegawa H, Tashiro M, Sasaki Y, Arakawa Y *et al.*: **Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by de novo sequencing using a next-generation DNA sequencer**. *PLoS ONE* 2010, **5**:e10256.

111. Nakamura S, Yang CS, Sakon N, Ueda M, Tougan T, Yamashita A, Goto N, Takahashi K, Yasunaga T, Ikuta K *et al.*: **Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach**. *PLoS ONE* 2009, **4**:e4219.

112. Yongfeng H, Fan Y, Jie D, Jian Y, Ting Z, Lilian S, Jin Q: **Direct pathogen detection from swab samples using a new high-throughput sequencing technology**. *Clin Microbiol Infect* 2011, **17**:241-244.