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

1369-5274/8 – see front matter, © 2013 The Author. Published by Elsevier Ltd. All rights reserved.

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>
<thead>
<tr>
<th>Name</th>
<th>Detection platform</th>
<th>NGS bioinformatics approach</th>
<th>Disease association</th>
<th>Strength of association</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronavirus-EMC</td>
<td>Culture and 454 NGS [7]</td>
<td>De novo genome assembly</td>
<td>Severe pneumonia (humans)</td>
<td>++</td>
</tr>
<tr>
<td>SFTS (severe fever with thrombocytopenia virus)</td>
<td>Illumina NGS [12]</td>
<td>Subtraction and BLAST search</td>
<td>Severe fever with thrombocytopenia</td>
<td>++</td>
</tr>
<tr>
<td>bunyavirus [12,13]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heartland bunyavirus</td>
<td>Culture and 454 NGS [14]</td>
<td>BLAST search and de novo gene assembly</td>
<td>Severe febrile illness</td>
<td>++</td>
</tr>
<tr>
<td>MCPyV (Merkel cell polyomavirus)</td>
<td>454 NGS [15]</td>
<td>Subtraction and BLAST search</td>
<td>Merkel cell carcinoma (MCC)</td>
<td>++</td>
</tr>
<tr>
<td>Bat paramyxoviruses</td>
<td>Consensus PCR [20–22]</td>
<td>N/A</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>Raccoon polyomavirus</td>
<td>Consensus PCR and RCA [38]</td>
<td>N/A</td>
<td>Brain tumors (racoons)</td>
<td>++</td>
</tr>
<tr>
<td>HPyV6 and HPyV7 (human polyomaviruses 6 and 7)</td>
<td>RCA [39]</td>
<td>N/A</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>TSPyV (trichodysplasia spinulosa-associated polyomavirus)</td>
<td>RCA [40]</td>
<td>N/A</td>
<td>Trichodysplasia spinulosa</td>
<td>++</td>
</tr>
<tr>
<td>2009 pandemic influenza A(H1N1) [3]</td>
<td>Microarray and Illumina NGS [51]</td>
<td>Subtraction and BLAST search</td>
<td>Febrile illness</td>
<td>++</td>
</tr>
<tr>
<td></td>
<td>454 NGS [110,111]</td>
<td>BLAST search</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Illumina NGS [112]</td>
<td>BLAST search</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TMAAdv (titi monkey adenovirus) [52]</td>
<td>Micorarray and illumina NGS</td>
<td>BLAST search</td>
<td>Pneumonia (titi monkeys)</td>
<td>++</td>
</tr>
<tr>
<td>BASV (Bas-Congo virus), a rhabdovirus</td>
<td>Illumina NGS [58]</td>
<td>Subtraction, BLAST search, and de novo genome assembly</td>
<td>Acute hemorrhagic fever</td>
<td>++</td>
</tr>
<tr>
<td>Novel circoviruses and cycloviruses in humans and monkeys</td>
<td>454 NGS [59]</td>
<td>Subtraction and BLAST search</td>
<td>Diarrhea</td>
<td>–</td>
</tr>
<tr>
<td>Human klassevirus/salivirus</td>
<td>Illumina NGS [61]</td>
<td>Subtraction and BLAST search</td>
<td>Diarrhea</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>454 NGS [60,62]</td>
<td>Subtraction and BLAST search</td>
<td>Diarrhea</td>
<td>–</td>
</tr>
<tr>
<td>MWPyV/HPy10/MXPyV (MW polyomavirus)</td>
<td>454 NGS [63]</td>
<td>Subtraction and BLAST search</td>
<td>Diarrhea</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Illumina NGS [64]</td>
<td>Subtraction and BLAST search</td>
<td>Diarrhea</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Illumina NGS [65]</td>
<td>BLAST search</td>
<td>WHIM syndrome</td>
<td>–</td>
</tr>
<tr>
<td>HPyV9 (human polyomavirus 9)</td>
<td>Illumina NGS [66]</td>
<td>Subtraction and BLAST search</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Consensus PCR [24]</td>
<td>N/A</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>Human bufavirus</td>
<td>454 NGS [67]</td>
<td>Subtraction and BLAST search</td>
<td>Diarrhea</td>
<td>–</td>
</tr>
<tr>
<td>HAsTV-PS (human astrovirus Puget Sound)</td>
<td>454 NGS [68]</td>
<td>Subtraction and BLAST search</td>
<td>Encephalitis</td>
<td>++</td>
</tr>
<tr>
<td>Human enterovirus 109</td>
<td>Consensus PCR and Illumina NGS [69]</td>
<td>Subtraction and BLAST search</td>
<td>Acute respiratory illness</td>
<td>+</td>
</tr>
<tr>
<td>Dandenong arenavirus</td>
<td>454 NGS [70]</td>
<td>Subtraction and BLAST search</td>
<td>Fatal febrile illness in transplant patients</td>
<td>++</td>
</tr>
<tr>
<td>Lujo arenavirus</td>
<td>454 NGS [71]</td>
<td>Subtraction and BLAST search</td>
<td>Acute hemorrhagic fever</td>
<td>++</td>
</tr>
<tr>
<td>TDAV (Theiler’s disease-associated virus), a novel pegivirus</td>
<td>Illumina NGS [73]</td>
<td>BLAST search and de novo genome assembly</td>
<td>Hepatitis (horses)</td>
<td>++</td>
</tr>
<tr>
<td>Bat, canine, horse, and rodent hepaviruses and pegiviruses</td>
<td>454 NGS [74–77]</td>
<td>BLAST search</td>
<td>Respiratory infection (dogs)</td>
<td>–</td>
</tr>
<tr>
<td>Canine bocavirus 3</td>
<td>Illumina NGS [78]</td>
<td>BLAST search</td>
<td>Hemorrhagic diarrhea and vasculitis (dog)</td>
<td>–</td>
</tr>
<tr>
<td>Snake arenaviruses</td>
<td>Illumina NGS [79]</td>
<td>Subtraction and BLAST search</td>
<td>Inclusion body disease (snakes)</td>
<td>++</td>
</tr>
</tbody>
</table>
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], PARV, 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 ~10^5 genome equivalents per milliliter [25]. A related strategy is the use of rolling circle amplification (RCA) [35, 36], which has been succesful 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 cardiomyoviruses [50], and 2009 pandemic influenza H1N1 (Figure 2a) [51]. In 2011, the ViroChip was also used to identify a novel adenovirus that caused a fulminating 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
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 bUMP virus [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 Theliler’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
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 pnucluse 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 DNAsce 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) normalizazion [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
reads generated per run, to detect novel agents with a high degree of sensitivity. For pathogen discovery, the Roche 454 GS-FLX+pyrosequencing™ 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™ platform) and high accuracy. More recently, Illumina NGS sequencing platforms (GAIIx™, HiSeq™, and MiSeq™) 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 biased 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**

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 Bouvet 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


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


74. 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 hepatic viruses and pegiviruses. Proc Nat Acad Sci U S A 2013. [Epub ahead of print].


www.sciencedirect.com

Current Opinion in Microbiology 2013, 16:468-478


