The Genomics of Emerging Pathogens

Cadhla Firth and W. Ian Lipkin

Center for Infection and Immunity, Mailman School of Public Health, Columbia University, New York, NY 10032; email: cbf2118@columbia.edu, wil2001@columbia.edu

Annu. Rev. Genomics Hum. Genet. 2013. 14:281–300

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi: 10.1146/annurev-genom-091212-153446

Keywords

viruses, bacteria, emerging infectious diseases, pathogen discovery

Abstract

Globalization and industrialization have dramatically altered the vulnerability of human and animal populations to emerging and reemerging infectious diseases while shifting both the scale and pace of disease outbreaks. Fortunately, the advent of high-throughput DNA sequencing platforms has also increased the speed with which such pathogens can be detected and characterized as part of an outbreak response effort. It is now possible to sequence the genome of a pathogen rapidly, inexpensively, and with high sensitivity, transforming the fields of diagnostics, surveillance, forensic analysis, and pathogenesis. Here, we review advances in methods for microbial discovery and characterization, as well as strategies for testing the clinical and public health significance of microbe-disease associations. Finally, we discuss how genetic data can inform our understanding of the general process of pathogen emergence.

INTRODUCTION

Infectious disease research has been transformed by the recent renaissance of the One Health approach, which recognizes the importance of the interrelationships among humans, animals, and the environment in health and disease. Industrialization, globalization, and the large-scale commercialization of agriculture have increased the susceptibility of both animal and human populations to infectious diseases (22). Humans are at risk from a wide range of zoonotic diseases: Up to 75% of emerging infectious diseases have been estimated to originate from animals, including such striking examples as HIV/AIDS, rabies, Ebola, and Lyme disease (65, 137, 158). Domestic and wild animals may serve as intermediate or amplifying hosts prior to transmission to humans: Hendra virus, Nipah virus, and SARS coronavirus (SARS-CoV) jumped from bats to humans through horses, pigs, and palm civets, respectively (40, 150). Similarly, the amplification of Japanese encephalitis virus in pigs is often required before spillovers into human populations can occur (143). The rate of contact between a reservoir and a novel host species is a key determinant for successful cross-species transmission events; unfortunately, many anthropogenic changes to the environment (e.g., deforestation, habitat fragmentation, urbanization, and agriculture) act to increase these contact rates (111). Side effects of industrialization and globalization alter the abundance, density, and physical proximity of multiple species, including humans, and these factors have already been implicated in the emergence of malaria, Lyme disease, dengue virus, Nipah virus, SARS-CoV, Rift Valley fever virus, and hantaviruses, to name only a few (1, 65, 159). In fact, many common modern phenomena, such as the illegal movement of animals or their tissues for use as pets, food, or medicine, have the potential to introduce dangerous pathogens into new environments. For example, 35 people were infected with monkeypox in 2003 when the virus was carried from Africa by a Gambian pouched rat brought in through the illegal pet trade (30, 140).

Although One Health surveillance typically focuses on risk to humans, increasing contact rates between humans and wild animal populations have also exposed animals to new pathogens, often with devastating consequences. Multiple human viruses, including respiratory syncytial virus, metapneumovirus, and several anelloviruses, have been identified in diseased primates, threatening the health of sensitive populations and the ecotourism industries that rely on them (72, 106, 110). In addition, the damaging economic impact of livestock diseases on human populations cannot be ignored. Livestock are particularly vulnerable to infectious diseases due to the effects of high-density farming, intense breeding practices, and global trade networks on population structure, genetic diversity, and immune system health (22, 129). Recent outbreaks of African swine fever, foot and mouth disease, bovine spongiform encephalopathy, and Marek's disease have all resulted in massive economic losses for affected countries (19, 55, 90, 138).

Globalization has forever shifted the scale and pace of disease outbreaks such that each new epidemic must be considered a potential global health threat. This was powerfully demonstrated during the 2009 H1N1 influenza pandemic, where laboratory-confirmed cases were documented in more than 214 countries and overseas territories and communities within a year of emergence (47). Fortunately, technologies of all kinds are rapidly progressing in response to the challenges of a growing and globalized world. Mechanisms of surveillance and reporting are becoming increasingly flexible and available to epidemiologists, clinicians, and biologists worldwide as the need to rapidly identify, characterize, and monitor emerging diseases continues to grow. Internet-based surveillance services—such as ProMED-mail (Program for Monitoring Emerging Diseases), GPHIN (Global Public Health Intelligence Network), HealthMap, and others—work to integrate and distribute submissions involving new or recurring epidemics from contributors all over the world, allowing response efforts to be initiated in real time (43, 93, 103).

Table 1 Examples of viruses that acquired specific mutations in association with emergence in a new host species

Virus	Original host	New host	Associated mutation (gene)	Effect of mutation	Reference(s)
West Nile virus	Passerine birds	Humans	T249P (*NS3*)	Enhanced virulence	15
Chikungunya virus	Various vertebrate species	Humans	A226V (*E1*)	Vector specificity	141
H3N8 influenza A	Horses	Dogs	W222L, N483T (*HA*)	Receptor function, host switch	23
H5N1 influenza A	Water birds	Humans	E627K (*PB2*)	Receptor function, host switch	53
HIV-1	Chimpanzees	Humans	M30R (*Gag*)	Viral fitness in new host	147
SARS coronavirus	Bats	Humans, civets, and related carnivores	K479N, S487T (*S*)	Receptor function, host switch	125, 131, 135
Canine parvovirus	Cats	Dogs	Multiple mutations (*VP3*)	Receptor function	3, 133

The widespread adoption of one of the most transformative innovations in biology, DNA sequencing, has opened up a wealth of new information, and the rapidly decreasing cost and increasing speed of DNA sequencing have ramifications for all aspects of biomedical research and clinical medicine, including ecology and evolution, diagnostics, immunology, vaccinology, drug development, and public health. In the field of infectious disease, this is perhaps best illustrated by the results of years of genetic and genomic analyses of influenza viruses from a range of host species, which have influenced every aspect of the field, from surveillance efforts to treatment programs. Phylogenetic analyses have exposed the transmission dynamics of both avian and mammalian strains, revealed that metapopulation structure better characterizes influenza than source-sink dynamics, and uncovered the importance of antigenic shift and genetic drift in the evolution of both seasonal and pandemic influenza (7, 76, 115, 127). Reverse genetics, in combination with in vitro experiments, has identified the mutations necessary for resistance to adamantine and oseltamivir, and has revealed that effective transmission in birds and/or mammals depends on receptor-binding affinity, which is linked to the identity of the amino acid at position 627 of the PB2 protein (14, 53, 112, 146) (**Table 1**). Genetic analyses have revealed more about influenza than about nearly any other pathogen (with the possible exception of HIV)—a direct result of years of concentrated sampling and sequencing effort. With more than 10,000 full genome sequences of influenza now publicly available (see the Influenza Genome Sequencing Project at **http://www.niaid.nih.gov/labsandresources/resources/dmid/gsc/influenza**), the success of this work suggests what is possible for the future of many of our most important infectious diseases.

The first few decades following the advent of the polymerase chain reaction (PCR) and DNA sequencing were largely dedicated to applying these tools to characterizing existing agents or the evolutionary relationships between them. In the current era, applications of DNA sequencing have evolved to occupy a place on the front lines of public health, helping to diagnose, characterize, and even treat some of today's most important diseases. High-throughput pyrosequencing enabled the implication of arenaviruses in a cluster of three women who died of encephalitis in Australia following organ transplantation, and in an outbreak of hemorrhagic fever in southern

Africa (18, 109). In the latter instance, this discovery led to the use of an antiviral drug that may have halted progression to fatal disease. More recently, other high-throughput sequencing (HTS) platforms have been used to resolve epidemics of *Escherichia coli* O104:74–associated hemolytic uremic syndrome and to track the distribution of antibiotic-resistant *Klebsiella pneumoniae* through a hospital (96, 130). The widespread availability of metagenomic techniques and HTS technology has had an enormous impact on infectious disease research, and much of this has understandably been focused on the discovery of novel pathogens. It is now not only possible but almost routine to sequence the metagenome of a particular sample or species, in addition to those associated with a particular syndrome, such as autism with gastrointestinal disturbances (50, 67, 105, 153).

In this review, we first discuss the current state of the technology and techniques used to detect and characterize bacterial and viral pathogens in the context of an emergence event (although the discussion is admittedly skewed toward viruses) and then explore how these recent advances can help further our understanding of emerging infectious diseases, particularly in outbreak scenarios.

SEQUENCING THE GENOMES OF NOVEL PATHOGENS

Pathogen Identification and Characterization

When responding to an outbreak, the most important and immediate question to address is the identity of the agent responsible for the disease. Identification of the causative agent can provide critical insights into estimates of the basic reproductive number (R_0), probable route(s) of transmission, and possible intervention strategies. The tools and techniques used to determine the causative agents of disease have evolved greatly since the discovery that it was not mysterious "miasmatas" but rather *Vibrio cholerae* that was responsible for cholera (62). Many reviews have covered in excellent detail the history and evolution of the increasingly sophisticated techniques used in pathogen discovery (5, 11, 27, 88, 100, 134). As the aim of this review is to discuss the current state of the field and its future prospects, we touch only briefly on the most significant advances in pathogen discovery on our way to the burgeoning field of metagenomics.

The era of culturing infectious agents was ushered in with the propagation of poliovirus in 1949, yet this remains one of the most significant accomplishments in microbial research to date (35). Despite the utility of cell culture for studying replication and pathogenesis as well as creating purified virus for antibody generation or vaccine development, many viral agents (e.g., all hepatitis viruses and human rhinovirus C) cannot easily be cultured (52, 63, 154). The amount of time and level of difficulty involved in culturing a virus have bolstered the movement toward pathogen discovery through molecular techniques. Consensus PCR, with degenerate primer pairs designed from conserved genomic regions of known microbes, has been one of the most effective approaches to identify novel agents, including coronaviruses, flaviviruses, and herpesviruses (17, 102, 118, 145, 151). Microarrays, which were originally designed to monitor gene expression across multiple targets, have been modified with probes targeting highly conserved gene regions and have been successfully applied to the characterization of new pathogens, including SARS-CoV, avian bornavirus, and a gammaretrovirus in patients with prostate cancer (69, 73, 122, 142, 149). However, each of these approaches is severely limited for the discovery of highly divergent or completely novel viruses and in cases where no conserved targets are available for primer design. Because of these limitations, sequence-independent approaches such as sequence-independent single-primer amplification (SISPA), random PCR, and rolling-circle amplification (RCA) have become popular alternative strategies for pathogen discovery, laying the groundwork for the evolution of metagenomics (44, 116, 117). Whereas SISPA involves the ligation of oligonucleotides to viral nucleic acid followed by PCR using primers complementary to the ligated fragments, random PCR and

RCA do not require a ligation step. Each of these methods has been highly effective at facilitating the identification and characterization of a range of pathogens, including human parvovirus 4, human coronavirus NL, GB virus C, rotaviruses, beta- and gammapapillomaviruses, and human polyomaviruses (41, 66, 71, 77, 87, 144). However, unless these techniques are applied to a purified sample containing only the agent of interest, all of the background material (e.g., host nucleic acid and environmental contaminants) will also be amplified, potentially complicating identification of the target sequences. This is an important limitation and one of the greatest confounding factors still present in all current pathogen discovery approaches. The ultimate aim of nearly all molecular methods used for discovery (whether sequence-dependent or -independent) is the generation of sequence data that can be used to definitively confirm both the presence and identity of the infectious agent.

Metagenomic approaches directly characterize the genetic material of viral and bacterial communities (the virome and bacteriome, respectively), while circumventing the need for agent-specific amplification techniques (86, 100). Modern HTS technology is capable of sequencing the virome or bacteriome in a sample rapidly and with high sensitivity, reducing the quantity of input material that is necessary relative to conventional approaches. In addition, the incorporation of barcodes during sample preparation creates a multiplexing capability that further lowers the cost of sequencing by increasing throughput. Although the concept of metagenomic analysis is often inexorably tied to HTS methods, early attempts at characterizing environmental metagenomes involved traditional molecular approaches—the creation of shotgun cDNA libraries followed by classical Sanger (i.e., first-generation) sequencing. However, there is little doubt that without the development of these new sequencing platforms, we would not have seen the rise of the metagenomics era (16). Outbreak investigation and pathogen discovery have benefited greatly from the ongoing race toward cheaper, faster, and more sensitive sequencing technologies, perhaps more than any other field.

Multiple HTS platforms are available, and the methods of template preparation, sequencing/imaging, and data analysis employed by each affect how they can best be used for pathogen discovery and genetic characterization. The variety of platforms and the novelty of the technology have made this field highly competitive, and as a result, DNA sequencing technologies continue to evolve at an impressive rate. Many outstanding reviews and benchmark studies have discussed in great detail the effects of the chemistry and physics of each platform on the data they generate (e.g., 89, 91, 97, 114). Therefore, we only briefly review the fundamental aspects of the most popular platforms for pathogen discovery, before turning to the applications of HTS in this field. With respect to pathogen discovery, the most important characteristics of an HTS platform are (*a*) the amount of template required for input; (*b*) the time and cost associated with sequencing; (*c*) the number of (pathogen) reads generated, i.e., the depth of coverage attained; and (*d*) the length of each read. The amount of template required is important because pathogen genomes are relatively small compared with those of their eukaryotic hosts, and in some sample types (e.g., tissues, feces, and environmental samples) the ratio of virus to host or background material will also be low. As a result, enrichment techniques are routinely applied to isolate microbial genetic material (discussed in more detail in the next section), further reducing the total amount of input material available. The time and cost required for template preparation and sequencing are obviously factors of vital importance for any scientific endeavor; however, pathogen discovery often takes place within the constraints of an outbreak, and the time frame in which answers are needed is measured in hours rather than in days or weeks. Lastly, the number and length of reads generated by each platform have critical implications for the downstream analytical approaches necessary to generate meaning from the data (discussed in further detail below).

Sample Preparation for Pathogen Discovery Using High-Throughput Sequencing

The sample preparation protocols used for pathogen discovery differ from those used for human genome sequencing because of the added complexity of nontarget material, which may be present in high concentrations in a sample (100) (**Figure 1**). Several approaches can be used to increase the ratio of microbial to background nucleic acid, contingent on the qualities of the starting material. Large sample volumes that are primarily cell free (e.g., water samples, cell culture medium, and in some cases cerebrospinal fluid) can be concentrated by ultracentrifugation or progressive filtration, which also removes cellular organisms (24, 139). Samples such as feces, oral or anal swabs, and serum or plasma should also be filtered to reduce volume and/or remove larger particulate matter. If the detection of viral agents is desired, bacterial, eukaryotic cells, and any free-floating organelles from lysed cells (e.g., mitochondria) should be removed from the sample by a final 0.22- or 0.45-μm filtration step (139). Tissue samples (e.g., diseased organs and lesion biopsies), although perhaps the most difficult starting material for pathogen discovery work due to the large quantities of host material, are often important from a clinical perspective. Initial sample preparation from tissue often includes mechanical homogenization with inert beads [e.g., using the TissueLyser (Qiagen)], freezing with liquid nitrogen followed by pulverization with a pestle, or passage through a small-gauge needle. Subsequent enzymatic digestion with proteinase K can be advantageous for difficult tissue types, as it not only denatures structural proteins, but also inactivates contaminating RNases and DNases in the sample. Following homogenization, cellular debris should be removed by filtration or centrifugation. If the bacteriome is of interest, nucleic acid extraction should proceed directly from this step, using DNA as template and employing 16S rRNA primers designed to amplify regions conserved in bacteria; when sequenced, the amplified region then provides significant taxonomic information (64). However, if the goal is to sequence the virome, a freeze-thaw step is sometimes used to lyse remaining cellular membranes, followed by RNase A and DNase I digestion to remove nucleic acid not protected by a viral nucleocapsid (2). Nucleic acid extraction can then be performed on all processed samples using standard techniques.

Following extraction, tissue samples can be further subjected to an rRNA-removal procedure [e.g., subtractive hybridization or depletion methods such as those used in RiboMinus (Life Technologies) and Ribo-Zero (Epicentre)] to avoid the domination of rRNA sequences in subsequent library preparations (54). Downstream preparations for sequencing (DNA/RNA shearing, cDNA generation, and library preparation) are dependent on both the input material and intended sequencing platform; an in-depth discussion of these protocols is beyond the scope of this review.

High-Throughput Sequencing Platforms in Pathogen Discovery

The first HTS platform to be widely used for pathogen discovery was also the first publicly available system—the 454 Life Sciences pyrosequencer, a platform that is still in use today as the Roche GS FLX Titanium and GS Junior systems (94). Sheared DNA (or cDNA) from extracted material is ligated to biotinylated linkers, which are then bound to streptavidin-coated beads inside a droplet of water along with PCR reagents. Individual strands of DNA are clonally amplified inside an oil emulsion using primers complementary to the ligated linker, and each individual DNA-bound bead is transferred to a PicoTiterPlate, where the sequencing reaction takes place (78). As nucleotides are added during the reaction, a pyrophosphate is released, emitting light that is detected by a charge-coupled device (CCD) camera within the GS system. Although the outputs from the GS systems are lower than those of many newer platforms (~0.7 Gb for the GS FLX Titanium, 14 Gb for the GS Junior), they also offer the longest available read lengths (a maximum of 700 bases),

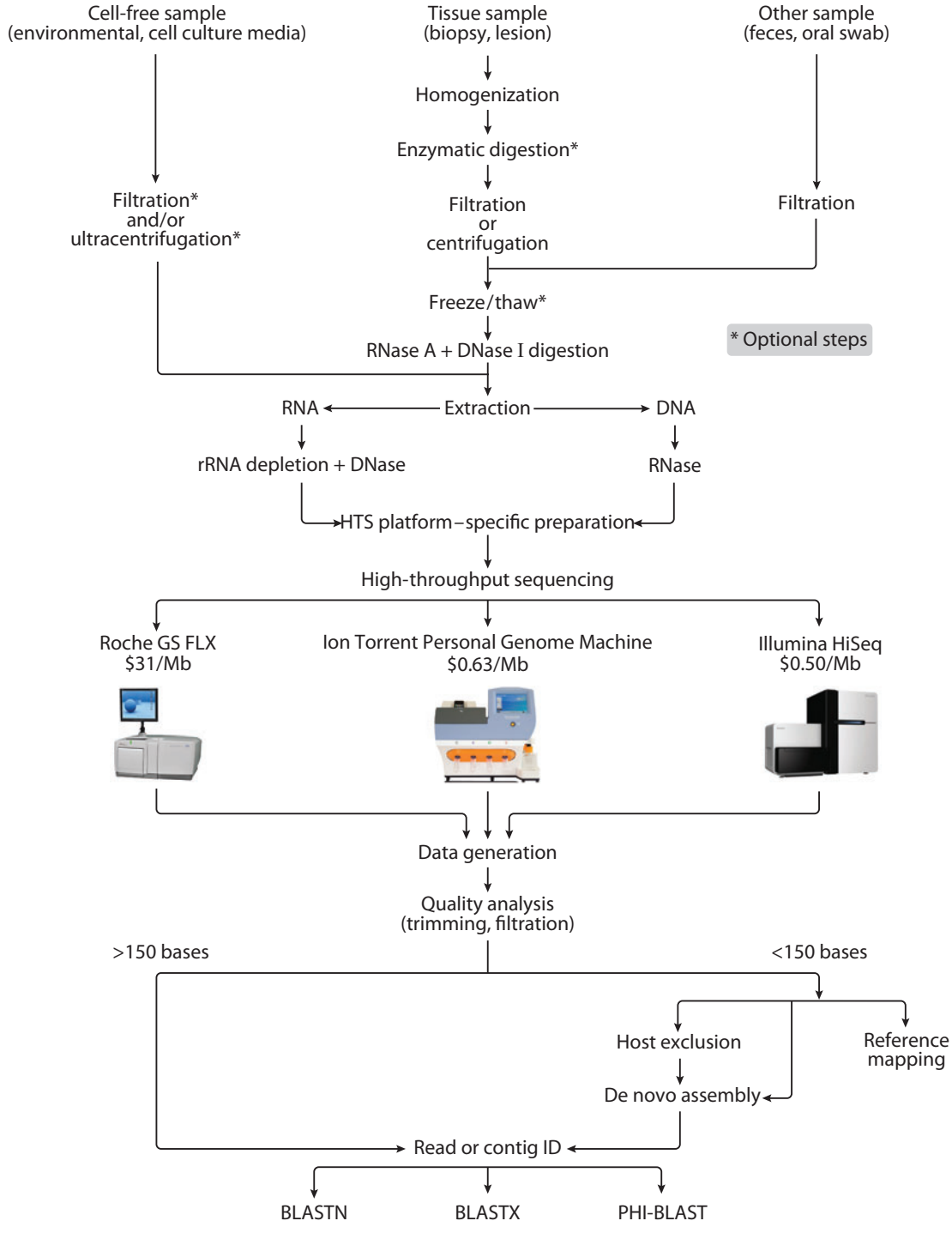

Figure 1

Standard sample preparation and data analysis workflows for viral discovery using high-throughput sequencing (HTS).

which is an advantage for pathogen discovery (89, 91). Having a small number of reads per sample makes the assembly of a viral or bacterial genome difficult, particularly in cases where most of the reads can be attributed to host or background material. However, this can be offset somewhat by having individual reads that are long enough to use directly for homology searches with nucleotide BLAST (BLASTN) or translated BLAST (BLASTX) (160). Dandenong virus, a novel arenavirus associated with a fatal outbreak in organ transplant recipients, was identified using only 14 reads from the GS FLX in the first use of HTS technology to identify a novel virus in an outbreak setting (109). Similarly, Merkel cell polyomavirus was identified by only 2 sequences from a pool of more than 400,000 generated from tumor biopsies, even after a polyadenylated RNA selection procedure was used (39).

Illumina currently produces two DNA sequencing platforms: the massively parallel HiSeq 2000, which can handle thousands of multiplexed samples simultaneously, and the MiSeq, a faster, lower-throughput benchtop sequencer launched in 2011. These platforms use sequencing-by-synthesis technology, where sheared DNA fragments are ligated to fixed adapters that act as primers (10). These are grafted to an acrylamide-coated glass flow cell where bridge amplification takes place to form clusters of clonal DNA fragments. Fluorescently labeled reversible-terminator nucleotides are incorporated by DNA polymerase one base at a time, and the signal is captured by a CCD camera. Illumina platforms are currently the most widely used HTS systems, and generate 600 Gb (HiSeq) or 3 Gb (MiSeq) of data per run with a read length of up to 150 bases (97, 114). Although the runtime for these machines is much longer than that of the GS FLX (approximately 27 h for the MiSeq and 11 days for the HiSeq, compared with 8 h for the GS FLX), the cost per megabase is an order of magnitude lower, making it a more accessible sequencing platform (91). However, the short read length is a significant challenge when using an Illumina platform for pathogen discovery (156). Because (*a*) the individual reads are short, (*b*) the read quality tends to decrease dramatically near the end of the read, and (*c*) tens of millions of reads are regularly generated per sample, it is not possible to BLAST each read to determine its origin. Instead, the reads must be assembled into longer contiguous sequences (contigs) that can be used in downstream applications. Unfortunately, assembly requires significant coverage of the genome in question (measured by the average number of reads that represent a given nucleotide), which can be problematic due to the relatively small amount of viral/bacterial template that is often present, relative to that contributed by background nucleic acid. Despite this constraint, Illumina sequencing technology was successfully used to identify a novel arterivirus associated with wobbly possum disease in the Australian brushtail possum, based on the de novo assembly of a 4.8-kb viral contig (33).

The Ion Torrent Personal Genome Machine (PGM) is a relatively new platform that, like the GS FLX, exploits emulsion PCR for amplification. However, rather then measuring the light emitted by nucleotide incorporation, the PGM system uses a modified silicon chip to detect the pH change that occurs when hydrogen ions are released during base incorporation. By not incorporating camera scanning, the PGM is faster (runtime of ~3 h) and less expensive than many other platforms. Although the PGM is not currently a high-throughput system—depending on the size of the chip used, it can generate between 20 Mb and 1 Gb of data and a maximum read length of 200 bases—the technology is new, and performance improvements are ongoing (114). Benchmark studies have indicated that de novo assemblies from PGM data are significantly more fragmented than those created from the Roche or Illumina systems, perhaps owing to the high number of miscalls that occur following homopolymeric sequences as short as two or three residues (91, 114). For pathogen discovery work, these miscalls and fragmented assemblies are likely to create difficulties primarily with assembling appropriate open reading frames, thereby interfering with BLASTX analyses, and in some cases additional conventional PCR-based Sanger sequencing may be required to confirm the sequence.

Despite the low throughput and high error rate of the PGM, the rapid speed with which it can generate data makes it an ideal platform for use in outbreak response. This was perfectly illustrated during the German outbreak of Shiga-toxin-producing *E. coli* O104:H4, which infected nearly 4,000 people in the summer of 2011, killing nearly 50 of these (74). The Beijing Genomics Institute (BGI) used the PGM to rapidly generate the first draft genome sequence of the outbreak strain, and took a generous approach to data release in one of the first examples of crowd-sourced analysis in an outbreak response effort (75, 120). BGI released the PGM-derived sequence data only three days after receiving the sample, and despite issues with homopolymeric errors, the assembly and annotations were completed by the community only 24 h later. These data were used to determine that the progenitor of the outbreak strain had an enteroaggregative phenotype that had acquired a Shiga-toxin-encoding phage and antibiotic resistance genes. BGI continued to improve on the draft genome by resequencing the same sample using the Illumina HiSeq, generating a high-quality data set within two weeks of sample receipt (120). The magnitude of this achievement is especially striking when compared with the decade of work that resulted in publication of the first full genome sequence of *E. coli* in 1997 (13, 75).

Analyzing Metagenomic Data for Pathogen Discovery

The initial stages of HTS data analysis are highly similar regardless of the type of sample (environmental sample, serum, or biopsy) or the goal of the analysis (full genome assembly or pathogen discovery). Although the platform type does alter the mechanics of the initial processing steps due to differences in the data generated (e.g., Illumina sequencers produce data of a fixed length, whereas the GS FLX and PGM instruments produce variable-length reads), the strategy does not change. Initial quality analyses are performed to assess potential problems with sample preparation or sequencing, as well as to prepare the data for downstream analyses. This step generally consists of assessing the quality scores assigned to each nucleotide across the length of the read to remove lower-quality bases, identifying and removing data artifacts (e.g., low-complexity reads and homopolymers), and determining the size of the prefix to be trimmed (i.e., primer/adapter sequences) (48, 95, 124).

The next step in the HTS data analysis pipeline involves efficiently assembling individual reads into contigs, which is one of the biggest challenges in the use of HTS data for pathogen discovery. Unlike genome sequencing projects, where the majority of sequence data derive from the target organism, samples that are analyzed for pathogen discovery not only are often dominated by host material but may also contain sequences from multiple viruses and/or bacteria. As a result, the ratio of reads from each pathogen to "other" reads can be low, leading to a fragmented assembly of the target genome sequence and an increased probability of assembling chimeric contigs (contigs generated from reads of multiple origin) (104, 123, 157). High-quality de novo assembly is typically dependent on high coverage; therefore, many more reads may be required from a clinical sample used for pathogen discovery than from one used for a host genome sequencing project. Several approaches have been developed to reduce the number of contaminating host reads from a mixed sample, which may be applied pre- or postassembly (11). The most straightforward approach requires the mapping of reads or contigs to the host genome, if available, followed by the subtraction of the mapped sequences from the data set. Alternatively, contigs can be aligned against a database containing only host sequences using BLAST, and those that match with 99% similarity removed. Although this approach is common, it is also computationally time consuming with large data sets. Fortunately, algorithms that remove nontarget sequences with higher accuracy and greater computational ease than a BLAST-based approach are now freely available (e.g., DeconSeq) (123). Further complications with assembly can arise when the pathogen of interest

exists as a highly variable population within an individual, as with RNA viruses (31). In general, high-quality de novo assemblies are created when strict parameters for matching and extension are used (i.e., zero mismatches between the "seed" sequence and the extending sequence, with most of the read lengths completely overlapping). However, the use of these stringent parameters on mixed populations of RNA viruses may result in poor assemblies and a failure to generate contigs. This is especially problematic when contigs are required for downstream homology searches due to short read lengths, as with the Illumina platforms. Therefore, it can be important to use an assembler with flexible parameters that can be adjusted by the user to suit each data set, with the caveat that using relaxed parameters for assembly can increase the number of chimeric contigs that then must be separated (104).

The final (and most important) step in the pipeline for pathogen discovery involves determining the identity of the agent(s) present in the sample. The success of this process is determined by three factors: (*a*) how well the sample preparation and sequencing methods were able to amplify the target material and modulate the host or environmental noise, (*b*) the efficacy of assembly, and (*c*) the similarity between the agent(s) and known microbes. The most common method used to identify unknown pathogens is a homology search against all known agents using BLAST, an approach that relies on the idea that new pathogens will share homology with those that are known. Unfortunately, a BLASTN will reveal matches only between sequences with reasonably high similarity (even with relaxed parameters); therefore, more distantly related or completely new pathogens will remain undetected. Fortunately, evolutionary relationships remain evident at the protein level much longer than at the nucleotide level. Therefore, translated contigs or long reads (>150 bases) should be searched against a translated database (BLASTX) to identify more distant relationships. Finally, a pattern-based homology detection program (e.g., PHI-BLAST) has been used to identify more divergent relationships, although this technique has yet to be successfully employed in a pathogen discovery setting (134).

APPLYING SEQUENCE DATA TO PATHOGEN DISCOVERY

Proof of Causation

Until recently, the central problem in pathogen discovery was the aspect of discovery—research was directed primarily toward identifying an infectious agent that could be the cause of a particular disease syndrome. Guided by what are now known as Koch's postulates, this process can take many years, and is perfectly illustrated by the decades-long campaign to identify and characterize the etiological agent of the 1918 influenza pandemic. Painstaking sequence analysis of samples obtained from archived materials and the remains of a victim frozen in the arctic tundra was required, followed by reconstitution of the virus using reverse genetics methods (136). The criteria of causation defined by Koch's postulates include (*a*) the presence of the agent in every case of the disease, (*b*) the specificity of the agent for that disease, (*c*) the successful propagation of the agent in pure culture, and (*d*) the ability of this inoculum to reproduce disease in a naive host (70) (**Table 2**). The central role of cell culture in these early tenets for proof of causation both emphasized the importance of cell culture in early pathogen work and cemented its place as the gold standard of discovery.

As the field of pathogen discovery has progressed, technological advances have likewise influenced the development and execution of the milestones toward proof of causation. In 1937, Rivers (119) modified Koch's postulates to reflect the knowledge that viruses cannot always be cultured, asymptomatic carriers exist, and antibodies can be used to aid in determining the timing, specificity, and degree of immune response to an agent. Subsequent advances in immunology, including the ability to purify viral antigen and detect specific antibodies, led Evans (36) to apply population-based serological characteristics toward proof of causation. The discovery of nucleic

Table 2 Strategies for proof of causation

Key criteria for proof of causation	Technological milestone(s)	Author(s)
Approaches motivated by specific technological advances		
The agent is present in every case of the disease The agent is specific for the disease After being isolated in pure culture, the agent can induce the disease The culture can reproduce the disease in a naive host	Cell culture	Koch, 1891 (70)
The virus is regularly (but not always) associated with the disease The agent must be associated with an immune response The association must be causative	Antibody detection	Rivers, 1937 (119)
Disease prevalence and incidence should be higher in individuals exposed to the agent than in the unexposed A measurable host immune response should occur following exposure to the agent Experimental reproduction of the disease should occur with higher incidence in exposed individuals Modification of host immune response should decrease or eliminate the disease	Specific antibody detection, viral antigen purification	Evans, 1976 (36)
A candidate gene should be associated with a pathogenic bacterium Inactivation or deletion of the candidate gene should lead to a loss in pathogenicity or virulence Reversion or allelic replacement of the gene should restore pathogenicity	Bacterial genetics, molecular cloning	Falkow, 1988 (37)
Candidate sequences should be present in most cases of the disease and pathology Sequences should be present prior to symptoms Few or no sequences should be present in disease-free tissues or individuals Sequences should diminish with recovery from the disease	Polymerase chain reaction, in situ hybridization	Fredericks & Relman, 1996 (42)
Metagenomic traits (e.g., sequence reads or assembled contigs, genes, or genomes) should be present and more abundant in diseased subjects compared with matched controls Inoculating a healthy individual with a sample containing the metagenomic trait should re-create the disease The candidate metagenomic traits should be recovered in newly diseased individuals	Metagenomics	Mokili et al., 2012 (100)
An integrative approach to causation		
Possible causal relationship: There is evidence of exposure to a microbe in diseased individuals *Probable causal relationship*: The microbial burden is high and localized to diseased tissues, the antibody response is consistent with recent exposure, and the microbe is present in multiple diseased individuals *Confirmed causal relationship*: Koch's postulates have been fulfilled, and/or disease prevention can be carried out through targeted therapies	Serology, polymerase chain reaction, in situ hybridization, immunohistochemistry, high-throughput sequencing	Lipkin, 2010 (88)

acids and subsequent advances in DNA sequencing and genetic characterization have led to multiple amendments of Koch's postulates, including a version focused on the association between bacterial genes and pathogenicity (37), and an adaptation that emphasized PCR-based identification of agents and the use of in situ hybridization for localization to the site of pathogenesis (42) (**Table 2**).

The molecular era provides a new set of challenges for proof of causation. A plethora of genetic information from known and novel infectious agents is continually being revealed from the analysis

of samples from both diseased and healthy individuals and animals. New viruses are often described without observed pathology, reversing the traditional workflow of searching for the causative agent of disease (28, 45, 99). An unintended outcome of this is that viruses and bacteria are occasionally associated with disease without rigorous attention to the fulfillment of a modern version of Koch's postulates. In some cases, these associations do not stand up to more careful scrutiny, resulting in the phenomenon of "de-discovery"—as with echoviruses and motor neuron disease (148), Borna disease virus and neuropsychiatric disease (61), human coronavirus NL63 and Kawasaki syndrome (32), and TT viruses and a range of syndromes (107). One of the most well known examples of de-discovery involved xenotropic murine leukemia virus–related virus (XMRV), which was initially associated with both prostate cancer and chronic fatigue syndrome before rigorous examination revealed a lack of association in both cases (4, 79, 92, 142). Although the discovery of XMRV using DNA microarrays is not in dispute, suggestions of any disease association attributed to this or related viruses have since been retracted (58).

Because of the vast amounts of data now being generated by advanced molecular techniques, strict adherence to guidelines regarding proof of causation may be more critical than ever. As in the past, modern modifications of Koch's postulates have focused on applying the data generated using the most contemporary (molecular) techniques. The "metagenomic Koch's postulates" focus on the use of molecular markers (individual sequence reads, assembled contigs, genes, or genomes) that uniquely discriminate diseased metagenomes from those of matched healthy controls (100). The case-control format of this approach uniquely allows for nonpathogenic microbes to be distinguished from those that may be causative, based on the premise that unassociated organisms should be equally present in diseased and healthy individuals (**Table 2**). Alternatively, one can apply a more holistic approach to proof of causation that focuses less on the types of data available, and more on the strength of the associations that can be inferred from the data (88) (**Table 2**). This method of assessing the certainty surrounding a proposed association is particularly useful in an outbreak scenario, where it may not be possible to perform the rigorous experiments necessary to fulfill Koch's postulates (or similar guidelines) before a response effort is needed. For a detailed description of the methods for testing confidence in the strength of an association between a candidate pathogen and a disease, including quantitation of microbial burden, localization of microbial footprints in affected tissues, antibody status, and biological plausibility, readers are referred to Reference 88.

Inferring the Origin and Dynamics of Emerging Pathogens from Genetic Data

Despite a growing understanding of the ecological, environmental, and genetic factors that influence the probability of emergence, considerable effort is still required to determine the origin of novel pathogens. When investigating an outbreak in real time, the rapid acquisition of sequence data can be critical in generating an appropriate response effort, determining reservoirs (if applicable), and identifying routes of transmission. Early genomic analysis of the 2009 pandemic H1N1 influenza virus revealed that initial swine-to-human transmission events had occurred months before recognition of the outbreak, confirmed Mexico as the likely geographic origin, and inferred a history of natural (rather than laboratory-created) reassortment that had remained undetected in swine for nearly a decade (25, 81, 128). Phylogenetic analysis of the genes or genome of a pathogen and related agents can also be used to (*a*) reconstruct contact networks, (*b*) identify cross-species transmission events (e.g., the movement of a pandemic parvovirus between cats, dogs, and raccoons, or multiple host-switching events between humans and bovids in *Staphylococcus aureus*), (*c*) estimate the rate and pattern of spatial spread (e.g., reconstruction of the circulation of human H1 influenza A in swine populations or the rapid dispersal of West Nile virus in North America), and (*d*) identify critical mutations associated with increased transmission

(e.g., the stepwise mutations in the receptor binding domain of the S protein needed for human transmission of SARS-CoV) or pathogenicity (e.g., the T249P NS3 mutation in West Nile virus) in a new host (3, 15, 83, 85, 113, 152) (**Table 1**).

HIV is one of the best examples of how the analysis of genetic data can be used to understand the history and dynamics of emergence over multiple timescales. Phylogenetic analyses of HIV-1, HIV-2, and multiple SIVs have (*a*) revealed a western African origin for HIV, (*b*) identified chimpanzee subspecies *Pan troglodytes troglodytes* as the reservoir of HIV-1 and the sooty mangabey (*Cercocebus atys*) as the reservoir of HIV-2, and (*c*) inferred that the emergence of the prominent global variant of HIV-1, group M subtype B, originated from a single migration event out of Haiti around 1970 (26, 46, 51, 68, 126). The reconstruction of contact networks using genetic data has also proved successful, revealing multiple webs of transmission linked by risk factor (e.g., hemophiliacs, heterosexuals, intravenous drug users, men who have sex with men) and migration pattern (9, 60, 80, 84). In addition, multiple instances of HIV transmission between health-care workers and their patients have been demonstrated using sequence data, resulting in interventions to prevent future transmission as well as evidentiary support for legal compensation (12, 49, 57, 108). Phylogenetic analyses have been used to support both convictions and acquittals in accusations of deliberate or accidental HIV transmission, indicating the power and relevance of using genetic analyses to reconstruct outbreaks and emergence events (29, 82, 98, 121).

Aside from reconstructing epidemiological history, genetic data can be used to add insight into the more general process of emergence. Examination of the genetic characteristics of a range of pathogens has revealed that specific types of viruses tend to emerge more frequently than would be expected in a purely stochastic process. For example, the large amount of sequence data generated from viral pathogens in particular has been critical for the understanding that RNA viruses may be much more likely to emerge through cross-species transmission than other pathogens (59, 65, 111). Analysis of the genetic diversity present in populations of RNA viruses through time has revealed that they are characterized by high mutation rates, high replication rates, and large population sizes (31, 59, 101). These features allow external selective pressures, such as vaccination, drug therapy, or cross-species transmission, to shape their diversity in real time. Therefore, RNA viruses may be able to quickly adapt to using novel cell receptors or transmission routes in a new host species, as observed with SARS-CoV, influenza, Venezuelan equine encephalitis (which required only a single glycoprotein mutation to move from rodents into horses), and Chikungunya virus (which was able to colonize a new vector species as the result of a single mutation in the envelope protein) (6, 141) (**Table 1**). Similarly, sequence data from RNA viruses that infect multiple host species have revealed a capacity to utilize cell receptors that are phylogenetically conserved across species (e.g., rabies virus uses the conserved nicotinic acetylcholine receptor, and foot and mouth disease virus can enter a cell using a variety of receptors, even within the same cell type), making host transitions more accessible (8, 21, 111).

CONCLUSION AND FUTURE PROSPECTS

DNA sequencing platforms will continue to evolve, becoming less expensive, less complex, and more portable. Several independent groups are developing single-molecule DNA sequencers that use nanopore technology, promise longer read lengths at lower cost, and may not require extensive sample preparation (20, 34). Such advances could enable microbial diagnostics as well as surveillance and discovery in clinics and remote field sites, further narrowing the time between the recognition of an outbreak and the generation of an appropriate response effort. Advances that facilitate human genome sequencing will also allow us to examine host factors that contribute to susceptibility and resistance to infectious diseases, and may lead to new personalized strategies

for targeting drugs and vaccines. As diagnostic sensitivity improves and samples continue to be collected from prospective population cohorts, we may also be able to detect evidence of exposure to microbes in prenatal and early life that ultimately results in mental illness, degenerative and autoimmune disorders, or neoplasia (132).

The era of infectious disease research characterized by the challenge of data acquisition has now been eclipsed by the challenge associated with effective data analysis. Whereas previous work has focused on analyzing tens of sequences, or at most several hundred, from a few genes or genomes, the availability of HTS data now requires the analysis of hundreds or thousands of genome sequences. Thus, developing the necessary analytical tools is a substantial obstacle. Many investigators are outsourcing both sequencing and sequence analysis to commercial centers, preferring to focus their efforts on subsequent steps in the process, including the verification and extension of these results by using PCR, serology, anatomical techniques, and animal models. In a parallel yet complementary discipline, synthetic genomics is allowing investigators to exploit sequence data to build known or novel microorganisms with features that may lead to increased virulence or transmission to new hosts (56, 155). The rapid pace with which DNA sequencing and synthetic biology are advancing has increased the need to establish policies that address the challenges of dual-use research, i.e., research that is scientifically valuable yet yields information or technologies that have the potential for misuse (38). The future of genomics and synthetic biology will depend as much on our ability to ensure open access, transparency, and ethical conduct of research as it will on material advances in platform technologies.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank Meera Bhat for her assistance in the preparation of this manuscript. Work in the Center for Infection and Immunity is supported by grants from the National Institutes of Health, the Department of Defense, and USAID-PREDICT.

LITERATURE CITED

1. Aguirre AA, Tabor GM. 2008. Global factors driving emerging infectious diseases. *Ann. N. Y. Acad. Sci.* 1149:1–3
2. Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J. 2001. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc. Natl. Acad. Sci. USA* 98:11609–14
3. Allison AB, Harbison CE, Pagan I, Stucker KM, Kaelber JT, et al. 2012. Role of multiple hosts in the cross-species transmission and emergence of a pandemic parvovirus. *J. Virol.* 86:865–72
4. Alter HJ, Mikovits JA, Switzer WM, Ruscetti FW, Lo SC, et al. 2012. A multicenter blinded analysis indicates no association between chronic fatigue syndrome/myalgic encephalomyelitis and either xenotropic murine leukemia virus-related virus or polytropic murine leukemia virus. *mBio* 3:e00266–12
5. Ambrose HE, Clewley JP. 2006. Virus discovery by sequence-independent genome amplification. *Rev. Med. Virol.* 16:365–83
6. Anishchenko M, Bowen RA, Paessler S, Austgen L, Greene IP, Weaver SC. 2006. Venezuelan encephalitis emergence mediated by a phylogenetically predicted viral mutation. *Proc. Natl. Acad. Sci. USA* 103:4994–99

7. Bahl J, Nelson MI, Chan KH, Chen R, Vijaykrishna D, et al. 2011. Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. *Proc. Natl. Acad. Sci. USA* 108:19359–64
8. Baranowski E, Ruiz-Jarabo CM, Sevilla N, Andreu D, Beck E, Domingo E. 2000. Cell recognition by foot-and-mouth disease virus that lacks the RGD integrin-binding motif: flexibility in aphthovirus receptor usage. *J. Virol.* 74:1641–47
9. Bello G, Zanotto PMDA, Iamarino A, Graf T, Pinto AR, et al. 2012. Phylogeographic analysis of HIV-1 subtype C dissemination in southern Brazil. *PLoS ONE* 7:e35649
10. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59
11. Bexfield N, Kellam P. 2011. Metagenomics and the molecular identification of novel viruses. *Vet. J.* 190:191–98
12. Blanchard A, Ferris S, Chamaret S, Guetard D, Montagnier L. 1998. Molecular evidence for nosocomial transmission of human immunodeficiency virus from a surgeon to one of his patients. *J. Virol.* 72:4537–40
13. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453–62
14. Bloom JD, Gong LI, Baltimore D. 2010. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* 328:1272–75
15. Brault AC, Huang CYH, Langevin SA, Kinney RM, Bowen RA, et al. 2007. A single positively selected West Nile viral mutation confers increased virogenesis in American crows. *Nat. Genet.* 39:1162–66
16. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* 18:630–34
17. Briese T, Jia XY, Huang C, Grady LJ, Lipkin WI. 1999. Identification of a Kunjin/West Nile-like flavivirus in brains of patients with New York encephalitis. *Lancet* 354:1261–62
18. Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, et al. 2009. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. *PLoS Pathog.* 5:e1000455
19. Callaway E. 2012. Pig fever sweeps across Russia. *Nature* 488:565–66
20. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* 4:265–70
21. Cleaveland S, Haydon DT, Taylor L. 2007. Overviews of pathogen emergence: Which pathogens emerge, when and why? *Curr. Top. Microbiol. Immunol.* 315:85–111
22. Cleaveland S, Laurenson MK, Taylor LH. 2001. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philos. Trans. R. Soc. Lond. B* 356:991–99
23. Crawford PC, Dubovi EJ, Castleman WL, Stephenson I, Gibbs EP, et al. 2005. Transmission of equine influenza virus to dogs. *Science* 310:482–85
24. Culley AI, Lang AS, Suttle CA. 2006. Metagenomic analysis of coastal RNA virus communities. *Science* 312:1795–98
25. Dawood FS, Jain S, Finelli L, Shaw MW, Lindstrom S, et al. 2009. Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N. Engl. J. Med.* 360:2605–15
26. De Leys R, Vanderborght B, Vanden Haesevelde M, Heyndrickx L, van Geel A, et al. 1990. Isolation and partial characterization of an unusual human immunodeficiency retrovirus from two persons of west-central African origin. *J. Virol.* 64:1207–16
27. Delwart EL. 2007. Viral metagenomics. *Rev. Med. Virol.* 17:115–31
28. Delwart EL, Li L. 2012. Rapidly expanding genetic diversity and host range of the Circoviridae viral family and other Rep encoding small circular ssDNA genomes. *Virus Res.* 164:114–21
29. de Oliveira TPO, Rambaut A, Salemi M, Cassol S, Ciccozzi M, et al. 2006. Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak. *Nature* 444:836–37
30. Di Giulio DB, Eckburg PB. 2004. Human monkeypox: an emerging zoonosis. *Lancet Infect. Dis.* 4:15–25
31. Domingo E, Holland JJ. 1997. RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.* 51:151–78

32. Dominguez SR, Anderson MS, Glode MP, Robinson CC, Holmes KV. 2006. Blinded case-control study of the relationship between human coronavirus NL63 and Kawasaki syndrome. *J. Infect. Dis.* 194:1697–701
33. Dunowska M, Biggs PJ, Zheng T, Perrott MR. 2012. Identification of a novel nidovirus associated with a neurological disease of the Australian brushtail possum (*Trichosurus vulpecula*). *Vet. Microbiol.* 156:418–24
34. Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–38
35. Enders JF, Weller TH, Robbins FC. 1949. Cultivation of the Lansing strain of poliomyelitis virus in cultures of various human embryonic tissues. *Science* 109:85–87
36. Evans AS. 1976. Causation and disease: the Henle-Koch postulates revisited. *Yale J. Biol. Med.* 49:175–95
37. Falkow S. 1988. Molecular Koch's postulates applied to microbial pathogenicity. *Rev. Infect. Dis.* 10(Suppl. 2):S274–76
38. Fauci AS, Collins FS. 2012. Benefits and risks of influenza research: lessons learned. *Science* 336:1522–23
39. Feng H, Shuda M, Chang Y, Moore PS. 2008. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* 319:1096–100
40. Field HE, Mackenzie JS, Daszak P. 2007. Henipaviruses: emerging paramyxoviruses associated with fruit bats. *Curr. Top. Microbiol. Immunol.* 315:133–59
41. Fouchier RAM, Hartwig NG, Bestebroer TM, Niemeyer B, de Jong JC, et al. 2004. A previously undescribed coronavirus associated with respiratory disease in humans. *Proc. Natl. Acad. Sci. USA* 101:6212–16
42. Fredericks DN, Relman DA. 1996. Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin. Microbiol. Rev.* 9:18–33
43. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. 2008. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *J. Am. Med. Inform. Assoc.* 15:150–57
44. Froussard P. 1992. A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. *Nucleic Acids Res.* 20:2900
45. Ge X, Li Y, Yang X, Zhang H, Zhou P, et al. 2012. Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in China. *J. Virol.* 86:4620–30
46. Gilbert MTP, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. 2007. The emergence of HIV/AIDS in the Americas and beyond. *Proc. Natl. Acad. Sci. USA* 104:18566–70
47. Girard MP, Tam JS, Assossou OM, Kieny MP. 2010. The 2009 A (H1N1) influenza virus pandemic: a review. *Vaccine* 28:4895–902
48. Golovko G, Khanipov K, Rojas M, Martínez-Alcántara A, Howard JJ, et al. 2012. Slim-Filter: an interactive Windows-based application for Illumina Genome Analyzer data assessment and manipulation. *BMC Bioinforma.* 13:166
49. Goujon CP, Schneider VM, Grofti J, Montigny J, Jeantils V, et al. 2000. Phylogenetic analyses indicate an atypical nurse-to-patient transmission of human immunodeficiency virus type 1. *J. Virol.* 74:2525–32
50. Grard G, Fair JN, Lee D, Slikas E, Steffen I, et al. 2012. A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. *PLoS Pathog.* 8:e1002924
51. Hahn BH, Shaw GM, De Cock KM, Sharp PM. 2000. AIDS as a zoonosis: scientific and public health implications. *Science* 287:607–14
52. Hao W, Bernard K, Patel N, Ulbrandt N, Feng H, et al. 2012. Infection and propagation of human rhinovirus C in human airway epithelial cells. *J. Virol.* 86:13524–32
53. Hatta M, Gao P, Halfmann P, Kawaoka Y. 2001. Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. *Science* 293:1840–42
54. He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, et al. 2010. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat. Methods* 7:807–12
55. Henson S, Mazzocchi M. 2002. Impact of bovine spongiform encephalopathy on agribusiness in the United Kingdom: results of an event study of equity prices. *Am. J. Agric. Econ.* 84:370–86
56. Herfst S, Schrauwen EJ, Linster M, Chutinimitkul S, de Wit E, et al. 2012. Airborne transmission of influenza A/H5N1 virus between ferrets. *Science* 336:1534–41
57. Hillis DM, Huelsenbeck JP. 1994. Support for dental HIV transmission. *Nature* 369:24–25

58. Holmes D. 2012. XMRV controversy laid to rest. *Lancet Infect. Dis.* 12:834
59. Holmes EC. 2009. *The Evolution and Emergence of RNA Viruses*. Oxford, UK: Oxford Univ. Press. 254 pp.
60. Holmes EC, Zhang LQ, Robertson P, Cleland A, Harvey E, et al. 1995. The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. *J. Infect. Dis.* 171:45–53
61. Hornig M, Briese T, Licinio J, Khabbaz RF, Altshuler LL, et al. 2012. Absence of evidence for bornavirus infection in schizophrenia, bipolar disorder and major depressive disorder. *Mol. Psychiatry* 17:486–93
62. Howard-Jones N. 1984. Robert Koch and the cholera vibrio: a centenary. *Br. Med. J. (Clin. Res. Ed.)* 288:379–81
63. Hsiung GD. 1984. Diagnostic virology: from animals to automation. *Yale J. Biol. Med.* 57:727–33
64. Hum. Microbiome Proj. Consort. 2012. A framework for human microbiome research. *Nature* 486:215–21
65. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, et al. 2008. Global trends in emerging infectious diseases. *Nature* 451:990–93
66. Jones MS, Kapoor A, Lukashov VV, Simmonds P, Hecht F, Delwart E. 2005. New DNA viruses identified in patients with acute viral infection syndrome. *J. Virol.* 79:8230–36
67. Kapoor A, Simmonds P, Gerold G, Qaisar N, Jain K, et al. 2011. Characterization of a canine homolog of hepatitis C virus. *Proc. Natl. Acad. Sci. USA* 108:11608–13
68. Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, et al. 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* 313:523–26
69. Kistler AL, Gancz A, Clubb S, Skewes-Cox P, Fischer K, et al. 2008. Recovery of divergent avian bornaviruses from cases of proventricular dilatation disease: identification of a candidate etiologic agent. *Virol. J.* 5:88
70. Koch R. 1891. Ueber bakteriologische Forschung. In *Verhandlungen des X Internationalen Medizinische Kongresses, Berlin 1890*, pp. 35–47. Berlin: August Hirschwald
71. Kohler A, Gottschling M, Manning K, Lehmann MD, Schulz E, et al. 2011. Genomic characterization of ten novel cutaneous human papillomaviruses from keratotic lesions of immunosuppressed patients. *J. Gen. Virol.* 92:1585–94
72. Köndgen SKH, N'Goran PK, Walsh PD, Schenk S, Ernst N, et al. 2008. Pandemic human viruses cause decline of endangered great apes. *Curr. Biol.* 18:260–64
73. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, et al. 2003. A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* 348:1953–66
74. Kupferschmidt K. 2011. As *E. coli* outbreak recedes, new questions come to the fore. *Science* 333:27
75. Kupferschmidt K. 2011. Scientists rush to study genome of lethal *E. coli. Science* 332:1249–50
76. Lam TT-Y, Ip HS, Ghedin E, Wentworth DE, Halpin RA, et al. 2012. Migratory flyway and geographical distance are barriers to the gene flow of influenza virus among North American birds. *Ecol. Lett.* 15:24–33
77. Lambden PR, Cooke SJ, Caul EO, Clarke IN. 1992. Cloning of noncultivatable human rotavirus by single primer amplification. *J. Virol.* 66:1817–22
78. Leamon JH, Lee WL, Tartaro KR, Lanza JR, Sarkis GJ, et al. 2003. A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* 24:3769–77
79. Lee D, Das Gupta J, Gaughan C, Steffen I, Tang N, et al. 2012. In-depth investigation of archival and prospectively collected samples reveals no evidence for XMRV infection in prostate cancer. *PLoS ONE* 7:e44954
80. Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J. 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl. Acad. Sci. USA* 93:10864–69
81. Lemey P, Suchard M, Rambaut A. 2009. Reconstructing the initial global spread of a human influenza pandemic. *PLoS Curr. Influ.* 1:RRN1031
82. Lemey P, Van Dooren S, Van Laethem K, Schrooten Y, Derdelinckx I, et al. 2005. Molecular testing of multiple HIV-1 transmissions in a criminal case. *AIDS* 19:1649–58
83. Leventhal GE, Kouyos R, Stadler T, von Wyl V, Yerly S, et al. 2012. Inferring epidemic contact structure from phylogenetic trees. *PLoS Comput. Biol.* 8:e1002413
84. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. 2008. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med.* 5:e50

85. Li F. 2008. Structural analysis of major species barriers between humans and palm civets for severe acute respiratory syndrome coronavirus infections. *J. Virol.* 82:6984–91
86. Li L, Delwart E. 2011. From orphan virus to pathogen: the path to the clinical lab. *Curr. Opin. Virol.* 1:282–88
87. Linnen J, Wages J, Zhang-Keck ZY, Fry KE, Krawczynski KZ, et al. 1996. Molecular cloning and disease association of hepatitis G virus: a transfusion-transmissible agent. *Science* 271:505–8
88. Lipkin WI. 2010. Microbe hunting. *Microbiol. Mol. Biol. Rev.* 74:363–77
89. Liu L, Li Y, Li S, Hu N, He Y, et al. 2012. Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* 2012:251364
90. Lobago F, Woldemeskel M. 2004. An outbreak of Marek's disease in chickens in central Ethiopia. *Trop. Anim. Health Prod.* 36:397–406
91. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al. 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30:434–39
92. Lombardi VC, Ruscetti FW, Das Gupta J, Pfost MA, Hagen KS, et al. 2009. Detection of an infectious retrovirus, XMRV, in blood cells of patients with chronic fatigue syndrome. *Science* 326:585–89
93. Madoff LC. 2004. ProMED-mail: an early warning system for emerging diseases. *Clin. Infect. Dis.* 39:227–32
94. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–80
95. Martínez-Alcántara A, Ballesteros E, Feng C, Rojas M, Koshinsky H, et al. 2009. PIQA: pipeline for Illumina G1 Genome Analyzer data quality assessment. *Bioinformatics* 25:2438–39
96. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, et al. 2011. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS ONE* 6:e22751
97. Metzker ML. 2010. Sequencing technologies—the next generation. *Nat. Rev. Genet.* 11:31–46
98. Metzker ML, Mindell DP, Liu XM, Ptak RG, Gibbs RA, Hillis DM. 2002. Molecular evidence of HIV-1 transmission in a criminal case. *Proc. Natl. Acad. Sci. USA* 99:14292–97
99. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, et al. 2011. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* 21:1616–25
100. Mokili JL, Rohwer F, Dutilh BE. 2012. Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2:63–77
101. Morimoto K, Hooper DC, Carbaugh H, Fu ZF, Koprowski H, Dietzschold B. 1998. Rabies virus quasispecies: implications for pathogenesis. *Proc. Natl. Acad. Sci. USA* 95:3152–56
102. Moureau G, Temmam S, Gonzalez JP, Charrel RN, Grard G, de Lamballerie X. 2007. A real-time RT-PCR method for the universal detection and identification of flaviviruses. *Vector-Borne Zoonotic Dis.* 7:467–77
103. Mykhalovskiy E, Weir L. 2006. The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. *Can. J. Public Health* 97:42–44
104. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 40:e155
105. Ng TFF, Marine R, Wang C, Simmonds P, Kapusinszky B, et al. 2012. High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *J. Virol.* 86:12161–75
106. Ninomiya M, Hoshino Y, Ichiyama K, Simmonds P, Okamoto H. 2009. Analysis of the entire genomes of torque teno midi virus variants in chimpanzees: infrequent cross-species infection between humans and chimpanzees. *J. Gen. Virol.* 90:347–58
107. Okamoto H. 2009. History of discoveries and pathogenicity of TT viruses. *Curr. Top. Microbiol. Immunol.* 331:1–20
108. Ou CY, Ciesielski CA, Myers G, Bandea CI, Luo CC, et al. 1992. Molecular epidemiology of HIV transmission in a dental practice. *Science* 256:1165–71
109. Palacios G, Druce J, Du L, Tran T, Birch C, et al. 2008. A new arenavirus in a cluster of fatal transplant-associated diseases. *N. Engl. J. Med.* 358:991–98
110. Palacios G, Lowenstine LJ, Cranfield MR, Gilardi KV, Spelman L, et al. 2011. Human metapneumovirus infection in wild mountain gorillas, Rwanda. *Emerg. Infect. Dis.* 17:711–13

111. Parrish CR, Holmes EC, Morens DM, Park EC, Burke DS, et al. 2008. Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol. Mol. Biol. Rev.* 72:457–70
112. Pizzorno A, Bouhy X, Abed Y, Boivin G. 2011. Generation and characterization of recombinant pandemic influenza A(H1N1) viruses resistant to neuraminidase inhibitors. *J. Infect. Dis.* 203:25–31
113. Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, et al. 2012. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci. USA* 109:15066–71
114. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, et al. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13:341
115. Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. 2008. The genomic and epidemiological dynamics of human influenza A virus. *Nature* 453:615–19
116. Rector A, Tachezy R, Van Ranst M. 2004. A sequence-independent strategy for detection and cloning of circular DNA virus genomes by using multiply primed rolling-circle amplification. *J. Virol.* 78:4993–98
117. Reyes GR, Kim JP. 1991. Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Mol. Cell. Probes* 5:473–81
118. Reyes GR, Purdy MA, Kim JP, Luk KC, Young LM, et al. 1990. Isolation of a cDNA from the virus responsible for enterically transmitted non-A, non-B hepatitis. *Science* 247:1335–39
119. Rivers TM. 1937. Viruses and Koch's postulates. *J. Bacteriol.* 33:1–12
120. Rohde H, Qin J, Cui Y, Li D, Loman NJ, et al. 2011. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N. Engl. J. Med.* 365:718–24
121. Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM, Metzker ML. 2010. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proc. Natl. Acad. Sci. USA* 107:21242–47
122. Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–70
123. Schmieder R, Edwards R. 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* 6:e17288
124. Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27:863–64
125. Sheahan T, Rockx B, Donaldson E, Sims A, Pickles R, et al. 2008. Mechanisms of zoonotic severe acute respiratory syndrome coronavirus host range expansion in human airway epithelium. *J. Virol.* 82:2274–85
126. Simon F, Mauclère P, Roques P, Loussert-Ajaka I, Müller-Trutwin MC, et al. 1998. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nat. Med.* 4:1032–37
127. Simonsen L, Viboud C, Grenfell BT, Dushoff J, Jennings L, et al. 2007. The genesis and spread of reassortment human influenza A/H3N2 viruses conferring adamantane resistance. *Mol. Biol. Evol.* 24:1811–20
128. Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, et al. 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 459:1122–25
129. Smith TC, Harper AL, Nair R, Wardyn SE, Hanson BM, et al. 2011. Emerging swine zoonoses. *Vector-Borne Zoonotic Dis.* 11:1225–34
130. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, et al. 2012. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci. Transl. Med.* 4:148ra116
131. Song HD, Tu CC, Zhang GW, Wang SY, Zheng K, et al. 2005. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *Proc. Natl. Acad. Sci. USA* 102:2430–35
132. Stoltenberg C, Schjolberg S, Bresnahan M, Hornig M, Hirtz D, et al. 2010. The Autism Birth Cohort: a paradigm for gene-environment-timing research. *Mol. Psychiatry* 15:676–80
133. Stucker KM, Pagan I, Cifuente JO, Kaelber JT, Lillie TD, et al. 2012. The role of evolutionary intermediates in the host adaptation of canine parvovirus. *J. Virol.* 86:1514–21
134. Tang P, Chiu C. 2010. Metagenomics for the discovery of novel human viruses. *Future Microbiol.* 5:177–89
135. Tang XC, Zhang JX, Zhang SY, Wang P, Fan XH, et al. 2006. Prevalence and genetic diversity of coronaviruses in bats from China. *J. Virol.* 80:7481–90
136. Taubenberger JK, Reid AH, Krafft AE, Bijwaard KE, Fanning TG. 1997. Initial genetic characterization of the 1918 "Spanish" influenza virus. *Science* 275:1793–96

137. Taylor LH, Latham SM, Woolhouse ME. 2001. Risk factors for human disease emergence. *Philos. Trans. R. Soc. Lond. B* 356:983–89
138. Thompson D, Muriel P, Russell D, Osborne P, Bromley A, et al. 2002. Economic costs of the foot and mouth disease outbreak in the United Kingdom in 2001. *Rev. Sci. Technol.* 21:675–87
139. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. 2009. Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* 4:470–83
140. Torres-Vélez F, Brown C. 2004. Emerging infections in animals—potential new zoonoses? *Clin. Lab. Med.* 24:825–38
141. Tsetsarkin KA, Vanlandingham DL, McGee CE, Higgs S. 2007. A single mutation in Chikungunya virus affects vector specificity and epidemic potential. *PLoS Pathog.* 3:e201
142. Urisman A, Molinaro RJ, Fischer N, Plummer SJ, Casey G, et al. 2006. Identification of a novel gammaretrovirus in prostate tumors of patients homozygous for R462Q *RNASEL* variant. *PLoS Pathog.* 2:e25
143. van den Hurk AF, Ritchie SA, Mackenzie JS. 2009. Ecology and geographical expansion of Japanese encephalitis virus. *Annu. Rev. Entomol.* 54:17–35
144. van der Meijden E, Janssens RWA, Lauber C, Bouwes Bavinck JN, Gorbalenya AE, Feltkamp MCW. 2010. Discovery of a new human polyomavirus associated with *Trichodysplasia spinulosa* in an immunocompromized patient. *PLoS Pathog.* 6:e1001024
145. VanDevanter DR, Warrener P, Bennett L, Schultz ER, Coulter S, et al. 1996. Detection and analysis of diverse herpesviral species by consensus primer PCR. *J. Clin. Microbiol.* 34:1666–71
146. van Riel D, Munster VJ, de Wit E, Rimmelzwaan GF, Fouchier RAM, et al. 2006. H5N1 virus attachment to lower respiratory tract. *Science* 312:399
147. Wain LV, Bailes E, Bibollet-Ruche F, Decker JM, Keele BF, et al. 2007. Adaptation of HIV-1 to its human host. *Mol. Biol. Evol.* 24:1853–60
148. Walker MP, Schlaberg R, Hays AP, Bowser R, Lipkin WI. 2001. Absence of echovirus sequences in brain and spinal cord of amyotrophic lateral sclerosis patients. *Ann. Neurol.* 49:249–53
149. Wang D, Urisman A, Liu Y-T, Springer M, Ksiazek TG, et al. 2003. Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol.* 1:e2
150. Wang LF, Eaton BT. 2007. Bats, civets and the emergence of SARS. *Curr. Top. Microbiol. Immunol.* 315:325–44
151. Watanabe S, Masangkay JS, Nagata N, Morikawa S, Mizutani T, et al. 2010. Bat coronaviruses and experimental infection of bats, the Philippines. *Emerg. Infect. Dis.* 16:1217–23
152. Weinert LA, Welch JJ, Suchard MA, Lemey P, Rambaut A, Fitzgerald JR. 2012. Molecular dating of human-to-bovid host jumps by *Staphylococcus aureus* reveals an association with the spread of domestication. *Biol. Lett.* 8:829–32
153. Williams BL, Hornig M, Buie T, Bauman ML, Cho Paik M, et al. 2011. Impaired carbohydrate digestion and transport and mucosal dysbiosis in the intestines of children with autism and gastrointestinal disturbances. *PLoS ONE* 6:e24585
154. Wilson GK, Stamataki Z. 2012. In vitro systems for the study of hepatitis C virus infection. *Int. J. Hepatol.* 2012:292591
155. Wimmer E, Mueller S, Tumpey TM, Taubenberger JK. 2009. Synthetic viruses: a new opportunity to understand and prevent viral disease. *Nat. Biotechnol.* 27:1163–72
156. Wommack KE, Bhavsar J, Ravel J. 2008. Metagenomics: read length matters. *Appl. Environ. Microbiol.* 74:1453–63
157. Wooley JC, Godzik A, Friedberg I. 2010. A primer on metagenomics. *PLoS Comput. Biol.* 6:e1000667
158. Woolhouse MEJ, Gowtage-Sequeria S. 2005. Host range and emerging and reemerging pathogens. *Emerg. Infect. Dis.* 11:1842–47
159. Woolhouse MEJ, Haydon DT, Antia R. 2005. Emerging pathogens: the epidemiology and evolution of species jumps. *Trends Ecol. Evol.* 20:238–44
160. Ye J, McGinnis S, Madden TL. 2006. BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* 34:6–9

Annual Review of Genomics and Human Genetics

Volume 14, 2013

Contents

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* articles may be found at http://genom.annualreviews.org