OPINION

Clinical and biological insights from viral genome sequencing

Charlotte J. Houldcroft, Mathew A. Beale and Judith Breuer

Abstract | Whole-genome sequencing (WGS) of pathogens is becoming increasingly important not only for basic research but also for clinical science and practice. In virology, WGS is important for the development of novel treatments and vaccines, and for increasing the power of molecular epidemiology and evolutionary genomics. In this Opinion article, we suggest that WGS of viruses in a clinical setting will become increasingly important for patient care. We give an overview of different WGS methods that are used in virology and summarize their advantages and disadvantages. Although there are only partially addressed technical, financial and ethical issues in regard to the clinical application of viral WGS, this technique provides important insights into virus transmission, evolution and pathogenesis.

Since the publication of the first shotgun-sequenced genome (cauliflower mosaic virus[1]), the draft human genome[2] and the first bacterial genomes (*Haemophilus influenzae*[3] and *Mycoplasma genitalium*[4]), and enabled by the rapidly decreasing cost of high-throughput sequencing[5], genomics has changed our understanding of human and pathogen biology. Several large projects that aim to systematically analyse microbial genomes have recently been completed or are ongoing (for example, sequencing thousands of microbiomes[6] and fungal genomes[6,7]); these projects are shaping our knowledge of the genetic variation that is present in pathogen populations, the genetic changes that underlie disease and the diversity of microorganisms with which we share our environment.

The methods and data from whole-genome sequencing (WGS), which have been developed through basic scientific research, are increasingly being applied to clinical medicine, involving both humans[8] and pathogens. For example, WGS has been used to identify new routes of transmission of *Mycobacterium abscessus*[9] in healthcare facilities (nosocomial transmission) and to understand *Neisseria meningitidis* epidemics in Africa[10], whereas the sequencing of partial genomes has been used to detect drug resistance in RNA viruses, such as influenza virus[11], and DNA viruses, such as human cytomegalovirus (HCMV)[12]. Viral genome sequencing is becoming ever more important, especially in clinical research and epidemiology. WGS of pathogens has the advantage of detecting all known drug-resistant variants in a single test, whereas deep sequencing (that is, sequencing at high coverage) can identify low levels of drug-resistant variants to enable intervention before resistance becomes clinically apparent[13,14]. Whole genomes also provide good data with which to identify linked infections for public health and infection control purposes[15,16]. However, progress in using viral WGS for clinical practice has been slow. By contrast, WGS of bacteria is now well accepted, particularly for tracking outbreaks and for the management of nosocomial transmission of antimicrobial-resistant bacteria[17,18].

In this Opinion article, we will address the challenges and opportunities for making WGS, using modern next-generation sequencing (NGS) methods, standard practice in clinical virology. We will discuss the strengths, weaknesses and technical challenges of different viral WGS methods (TABLE 1), and the importance of deeply sequencing some viral pathogens. We will also explore two areas in which viral WGS has recently proven its clinical utility: metagenomic sequencing to identify viruses that cause encephalitis (BOX 1); and the role of WGS in molecular epidemiology and public health management of the Pan-American Zika virus outbreak (BOX 2). Finally, we will briefly consider the ethical and data analysis challenges that clinical viral WGS presents.

Why sequence viruses in the clinic?

For small viruses, such as HIV, influenza virus, hepatitis B virus (HBV) and hepatitis C virus (HCV), the sequencing of partial genomes has been widely used for research, but it also has important clinical applications. One of the main applications and reasons for sequencing viruses is the detection of drug resistance. For example, the management of highly active antiretroviral therapy (HAART) for HIV relies on viral sequencing for the detection of drug-resistant variants. HAART has substantially improved the survival of patients who have HIV, but successful therapy requires long-term suppression of viral replication with antiretroviral drugs, which may be prevented by impaired host immunity, suboptimal drug penetration in certain tissue compartments and incomplete adherence to therapy[19]. When viral replication continues despite treatment, the high mutation rate of HIV enables resistant variants to develop. It has become standard practice in many parts of the world to sequence the HIV *pol* gene, which encodes the main viral enzymes, to detect variants that confer resistance to inhibitors of reverse transcriptase, integrase or protease[20], particularly when patients are first diagnosed and when viral loads indicate treatment failure. Sequencing resistant variants has enabled targeted changes in treatment, which has resulted in greater reductions in viral loads than with standard care (undetectable HIV load in 32% versus 14% of patients after six months)[21,22]. Thus, sequencing resistant variants to guide HIV treatment improves disease outcomes. Similar approaches have been used to identify resistant variants of HCV[23], HBV[24] and influenza virus[25].

Table 1 | **Advantages and disadvantages of different viral sequencing methods**

Method	Advantages	Disadvantages
Metagenomic sequencing	• Simple, cost-effective sample preparation • Can sequence novel or poorly characterized genomes • Effective in 'fishing' approaches to identify a potential underlying pathogen • Lower required number of PCR cycles causes few amplification mutations • Preservation of minor variant frequencies reflects *in vivo* variation • No primer or probe design required, which enables a rapid response to novel pathogens or sequence variants	• High sequencing cost to obtain sufficient data • Relatively low sensitivity to target pathogen • Coverage is proportional to viral load • High proportion of non-pathogen reads increases computational challenges • Incidental sequencing of human and off-target pathogens raises ethical and diagnostic issues
PCR amplification sequencing	• Tried and trusted well-established methods and trained staff • Highly specific; most sequencing reads will be pathogen-specific, which decreases sequencing costs • Highly sensitive, with good coverage even at low pathogen load • Relatively straightforward design and application of new primers for novel sequences	• Labour-intensive and difficult to scale for large genomes • Iterating standard PCRs across large genomes requires high sample volume • PCR reactions are subject to primer mismatch, particularly in poorly characterized or highly diverse pathogens, or those with novel variants • Limited ability to sequence novel pathogens • High number of PCR cycles may introduce amplification mutations • Uneven amplification of different PCR amplicons may influence minor variant and haplotype reconstruction
Target enrichment sequencing	• Single tube sample preparation that is suited to high-throughput automation and the sequencing of large genomes • Higher specificity than metagenomics decreases sequencing costs • Overlapping probes increases tolerance for individual primer mismatches • Fewer PCR cycles (than PCR amplification) limits the introduction of amplification mutations • Preservation of minor variant frequencies reflects *in vivo* variation	• High cost and technical expertise for sample preparation • Unable to sequence novel pathogens and requires well-characterized reference genomes for probe design • Sensitivity is comparable to PCR, but coverage is proportional to pathogen load; low pathogen load yields low or incomplete coverage • Cost and time to generate new probe sets limit a rapid response to emerging and novel viruses

Why sequence whole genomes?

Limited sequencing of the small number of genes that encode targets of antiviral agents, such as HIV *pol*, has been the norm in clinical practice. For the detection of a limited number of antiviral-resistant variants, WGS has been too costly and labour-intensive to use compared with sequencing only the specific genes that are targeted by the drugs. However, the increasing number of resistance genes that are located across viral genomes, together with decreasing costs of sequencing and the use of sequence data for transmission studies, are driving a reappraisal of the need for WGS. For example, antiviral treatment for HCV now targets four gene products (NS3, NS4A, NS5A and NS5B), and these genes encompass more than 50% of the viral genome[26]. Individually sequencing each of these genes can be as expensive and time-consuming as WGS[27]. Partial-genome sequencing is particularly problematic for large viral genomes, in particular those of the herpesviruses HCMV[12], varicella zoster virus (VZV)[28], herpes simplex virus 1 (HSV-1)[29] and HSV-2 (REF. 30). These viruses have traditionally been treated with drugs that target the viral thymidine or serine/threonine protein kinases and DNA polymerase. However, the increasing number of drugs in development that interact with different proteins that are encoded by viral genes scattered across the genome, means that targeted sequencing for resistance testing is costly, involves more PCR reactions (which increases the chance of failure), requires more starting material, is more labour intensive and generally less tractable for diagnostic use[31]. Sequencing the whole genome simultaneously captures all resistant variants and removes the need to design and optimize PCR assays for the detection of resistance to new drugs. A good example of this is HCMV, for which WGS can simultaneously capture the genes that encode targets of licensed therapies, such as UL27 (unknown function), UL54 (DNA polymerase) and UL97 (serine/threonine protein kinase), and of newer drugs, such as letermovir, which targets UL56 (terminase complex). This enables comprehensive antiviral-resistance testing in a single test[12]. In addition, WGS can provide information on antigenic epitopes, virus evolution in a patient over time[12], and evidence of recombination between HCMV strains[32]. WGS can also detect putative novel drug-resistant variants and predict changes to epitopes, although phenotypic testing of variants is required to confirm clinical resistance[33] and to map epitope changes[34].

As pre-existing resistance to antiviral drugs increases (for example, HCV that is resistant to protease inhibitors[35] and HBV that is resistant to nucleoside analogue reverse-transcriptase inhibitors[36]), WGS will provide the comprehensive resistance data that are required for selecting appropriate treatment. The complete knowledge of all resistant variants can also support novel decisions in clinical management; for example, the identification of extensive genome-wide HCMV drug resistance in a patient supported the decision to treat the individual with autologous cytomegalovirus-specific T cells instead of antiviral drugs[37].

WGS may also better identify transmission events and outbreaks, which is not always possible with sequences of subgenomic fragments. For example, WGS of respiratory syncytial virus (RSV) identified variation outside of the gene that is traditionally used for genotyping, and such information could be used to track outbreaks in households when the genetic variability in single genes is too low for transmission studies[38]. The numerous phylogenetically informative variant sites that can be obtained from full-length or near full-length genomes removes the need for high-quality sequences, which enabled the robust linking of cases of Ebola virus infection and public health interventions in real time during the 2015 epidemic[39].

Box 1 | **RNA-seq and metagenomic diagnostics**

For cases of encephalitis of unknown origin, metagenomic techniques are promising diagnostic tools. There are various protocols in use, but the main methods that are used are RNA sequencing (RNA-seq) and metagenomics. For RNA-seq, the total RNA, or a subset of RNA, is extracted from a sample (for example, cerebrospinal fluid or a brain biopsy), converted to complementary DNA (cDNA) and sequenced. Metagenomics generally describes the same procedure for DNA, but may also include simultaneous sequencing of DNA and RNA through the incorporation of a cDNA-synthesis step. RNA-seq may improve the detection of pathogenic viruses, as many viruses have RNA genomes and viral mRNAs in the cerebrospinal fluid (CSF) or brain indicate both the presence of the virus and which viral genes are being transcribed. However, DNA viruses, which experience low-level transcription, may be poorly detected using RNA-seq, and read numbers for DNA viruses may be higher in metagenomic datasets[63].

Both methods have successfully identified new or known viral pathogens in cases of encephalitis of unknown origin. Metagenomics has been used to aid the diagnosis and characterization of enterovirus D68 in cases of acute flaccid paralysis[83]. Metagenomics identified herpesviruses in the CSF of four patients who had suspected viral meningoencephalitis[136]. RNA-seq also identified herpes simplex virus 1 (HSV-1) in a patient with encephalitis, although the use of a DNase I digestion (which was intended to decrease the amount of host nucleic acid) decreased the number of HSV-1 reads[63]. Mumps vaccine virus has also been detected in a patient with chronic encephalitis using RNA-seq[137].

RNA-seq has been very successful in the identification of encephalitis caused by astroviruses[138,139] and coronaviruses[65]. The deaths of three squirrel breeders from encephalitis were linked to a novel squirrel bornavirus, which was identified by separate metagenomic sequencing of DNA and RNA[62]. Metagenomics provides more information about the virus in a sample than PCR alone, which may be important for molecular epidemiology, whereas RNA-seq can identify viral sequences and viral gene expression.

Another example of WGS supporting public health efforts is with the recent outbreak of Zika virus (BOX 2).

The routine use of pathogen WGS for diagnostic purposes[40] is likely to have wider clinical and research benefits. For example, Zika virus sequences that were generated for epidemiological purposes inform public health decisions[41]. In addition, HIV genomes that were sequenced to identify antiviral-resistant variants have also been used to study virus evolution[42] and viral genetic association with disease, including genotype–phenotype association studies and genome-to-genome association studies, which look for associations between viral genetic variants, host genetic variants and outcomes of infection, such as viral load set point in HIV infection[43,44].

Why do we need deep sequencing?

Modern methods, which use massively parallel sequencing, enable better examination of diversity and the analysis of virus populations that contain nucleotide variants or haplotypes at low frequencies (less than 50% of the consensus sequence). Minority variant analysis is particularly powerful for RNA viruses, reverse-transcribing DNA viruses and retroviruses, because they typically show high diversity, even in a single host. HIV is the classic example; the reverse transcriptase of HIV is error-prone and introduces mutations at an extremely high rate ($4.1 \pm 1.7 \times 10^{-3}$ per base per cell)[45]. Many closely related, but subtly different, viral variants exist in a single patient. These variants are sometimes described as a quasispecies or a cloud of intra-host virus diversity. The presence of a mixed population of viruses introduces problems for the determination of the true consensus 'majority' sequence, but these minority (non-consensus) variants may also change the clinical phenotype of the virus, and can be used to predict changes in genotype, tropism or drug resistance. For example, a minor variant that confers drug resistance in HIV that is present in only 2.1% of sequencing reads in a patient at baseline can rapidly become the majority (consensus) variant under the selective pressure of drug treatment[46]. Similar changes in the frequency of resistance-associated alleles during treatment have been observed for HBV[47], HCV[48], HCMV[12] and influenza virus[49].

Deep sequencing of viruses is not only required to detect drug resistance, it is also key for genotypically predicting the receptor tropism of HIV, which has treatment implications. HIV can be grouped by its use of cellular co-receptor into R5 (uses CC-chemokine receptor 5 (CCR5)), X4 (uses CXC chemokine receptor 4 (CXCR4)) or R5X4 (dual tropism). Maraviroc is a CCR5 antagonist that blocks infection of R5-tropic HIV, but not of X4-tropic and R5X4-tropic HIV. Just a 2% frequency of X4 or R5X4 genotypes is predictive of maraviroc treatment failure[50]. Sub-consensus frequencies of X4-tropic or R5X4-tropic HIV are also important for the success[51] and failure[52] of bone marrow transplants from CCR5-deficient (CCR5-Δ32) donors, and this information may influence the decision to stop antiviral therapy in these patients[51].

Minority variants and the identification of haplotypes can also be used to detect mixed infections. Infections with different HCMV genotypes or super-infections[53] are associated with poor clinical outcomes, and the detection of such mixed infections by WGS might justify more aggressive treatment.

Sanger sequencing of a virus population can detect minority variants at frequencies between 10% and 40%[54], whereas NGS can sequence those same PCR amplicons to a much greater depth[55], and, consequently, capture more of the variability that is present. Sensitivity and specificity are specific for the analysed virus and the sequencing method. Many studies of drug resistance in HIV that use deep-sequencing of PCR amplicons require minority variants to be present at >1% to decrease the possibility of false-positive results[56,57]. This may miss drug-resistance mutations at frequencies of 0.1–1% and lead to poor treatment outcome[57]. Although a 1–2% frequency threshold (or lower) may be clinically relevant for the detection of drug resistance in HIV, it is less clear whether the same degree of sensitivity is required for monitoring vaccine escape in HBV or drug resistance in herpesviruses (discussed below). Large cohorts of patients need to be tested before, during and after treatment[46,50] to establish thresholds for minority drug-resistant variants[12] and vaccine-escape variants that are clinically relevant for each virus.

Direct deep sequencing of clinical material, either by shotgun methods or RNA-seq methods (so called metagenomic methods), also enables the unbiased detection and diagnosis of pathogens, and provides an alternative to culture, electron microscopy and quantitative PCR (qPCR; see below).

Practical considerations

Sequencing viral nucleic acids, whether from cultures or directly from clinical specimens, is complicated by the presence of contaminating host DNA[58]. By contrast, most bacterial sequencing is currently carried out on clinical isolates that are cultured; thus, sample preparation is comparatively straightforward (TABLE 2 and

Box 2 | **Whole-genome sequencing of Zika virus**

Whole-genome sequencing (WGS) of Zika virus can help to understand the epidemiology of the recent outbreak in South America, including the origin and spread of the virus, and the connection between the virus and microcephaly. It also informs control measures, such as stopping importation of Zika cases or disrupting transmission from a reservoir, and blood safety measures in hospitals.

For flaviviruses, such as Zika virus, WGS, or at least near whole-genome sequencing, is required to provide molecular epidemiology studies sufficient power[41]. WGS, phylogenetic analysis and molecular clock dating, combined with other epidemiological data, were useful to study the introduction of Zika virus to South America[41]. For example, the most recent common ancestor of strains that are circulating in Brazil pre-dates the 2014 football World Cup, which makes it highly unlikely that this event was responsible for the introduction of the Asian-lineage Zika virus to South America[41].

WGS is also central for understanding the pathogenesis of Zika virus; for example, by trying to identify sequence changes that are associated with microcephaly, as it is currently unclear which genome regions determine pathogenesis. It is likely that numerous whole-genome sequences of Zika virus from around the world and from individuals with microcephaly and asymptomatic infection are required to link particular mutations to birth defects. So far, no changes in the Zika virus genome have been unambiguously associated with microcephaly[41,72,81].

WGS and fragment sequencing were used to identify a case of Zika virus transmission through platelet transfusion[140]. This case suggested that asymptomatic donors can transmit the virus to immunocompromised individuals. PCR-based testing had already established the presence of Zika virus in the blood supply in a previous outbreak, but no infection was detected in recipients of blood products[141]. Based on this new evidence, blood products may need to be screened routinely for Zika virus[140].

Finally, WGS of Zika virus isolates has identified sequence polymorphisms in primer binding sites[142], which may make PCR-based diagnosis and the quantification of viral load more difficult. This highlights the need to characterize population-level diversity, especially in epidemics in which the locally circulating virus may have diverged from viruses from other locations or time periods. Several projects are underway to determine population-level diversity, including the Zika in Brazil real time analysis (ZIBRA) mobile laboratory project[143], which uses portable metagenomic sequencing of Zika virus and real-time reporting of results[107].

reviewed in REF. 59). Currently, genome sequencing of viruses can be achieved by ultra-deep sequencing or through the enrichment for viral nucleic acids before sequencing, either directly or by concentrating virus particles. All of these approaches have their own costs and complexities.

Three main methods are currently used for viral genome sequencing: metagenomic sequencing, PCR amplicon sequencing and target enrichment sequencing (FIG. 1).

Metagenomics. Metagenomic approaches have been used extensively for pathogen discovery and for the characterization of microbial diversity in environmental and clinical samples[60,61]. Total DNA and/or RNA, including from the host, bacteria, viruses, fungi and other pathogens, are extracted from a sample, and a library is prepared and sequenced by shotgun sequencing or RNA sequencing (RNA-seq). BOX 1 explores the diagnostic applications for metagenomics and RNA-seq; for example, in encephalitis of unknown aetiology[62–64], for which conventional methods such as PCR are often not diagnostic, metagenomics and RNA-seq have detected viral infections[65–67] and other causes[68] of encephalitis. In addition, these methods have been used to sequence the whole genome of some viruses, including Epstein–Barr virus (EBV)[69] and HCV[27]. However, in clinical specimens, the presence of contaminating nucleic acids from the host and commensal microorganisms[58] (TABLE 2) decreases sensitivity. The proportion of reads that match the target virus genome from metagenomic WGS is often low; for example, 0.008% for EBV in the blood of a healthy adult[70], 0.0003% for Lassa virus in clinical samples[71] and 0.3% for Zika virus in a sample that was enriched for virus particles through filtration and centrifugation[72]. The read depth is often inadequate to detect resistance[27] and the cost is high. Thus, metagenomic sequencing has typically only been carried out on a small number of samples for research purposes[72,73]. The concentration of virus particles (see the Zika virus example above[72]), depletion of host material and/or sequencing to high read depth can increase the amount of virus sequence, but all of these methods add to the cost. The concentration of virus particles from clinical specimens by antibody-mediated pull-down (for example, virus discovery based on complementary DNA (cDNA)–amplified fragment length polymorphism (AFLP), abbreviated to VIDISCA), filtration, ultracentrifugation and the depletion of free nucleic acids, which mostly come from the host, have all been tried[74–77]; however, these methods may also decrease the total amount of viral nucleic acids so that it is insufficient for preparing a sequencing library. Non-specific amplification methods, such as multiple displacement amplification (MDA), which make use of random primers and Φ29 polymerases, can increase the DNA yield. However, these approaches are time consuming, costly, and may increase the risk of biases, errors and contamination, without necessarily improving sensitivity[78,79]. Moreover, the proportion of host reads often remains high[80].

When metagenomic methods are used for pathogen discovery or diagnosis, it is crucial to use appropriate bioinformatic tools and databases that can evaluate whether detected pathogen sequences are likely to be the cause of infection, incidental findings or contaminants. Bioinformatic analyses of large metagenomic datasets require high-performance computational resources.

The fact that metagenomics requires no prior knowledge of the viral genome, can be considered an advantage[27] as it enables novel viruses to be sequenced without the need for primer or probe design and synthesis. This is particularly relevant for rapid responses to emerging threats, such as Zika virus[81]. For virus-associated cancers, metagenomics can inform clinical care, provide information on cancer evolution and generate high-coverage data of integrated virus genomes[69]. However, incidental findings, both in host and microbial sequences, may also present ethical and even diagnostic dilemmas for clinical metagenomics[82] (see below). A recent example involved a cluster of cases of acute flaccid myelitis that were associated with enterovirus D68 (REF. 83). The metagenomic data from samples taken from patients showed the presence of alternative pathogens, some of which are treatable, and was debated in formal[84] and informal scientific channels (see Omicsomics blogspot article). Regulation and reporting frameworks will be important to resolve future issues of this kind.

PCR amplicon enrichment. An alternative to metagenomic approaches is to enrich the specific viral genome before sequencing. PCR amplification of viral genetic material using primers that are complementary to

Table 2 | **Limitations of viral sequencing compared with bacterial sequencing**

Feature	Bacteria	Viruses	Challenges
Genome	dsDNA	dsDNA, ssDNA, partially dsDNA, ssRNA or dsRNA	Different extraction protocols for different viruses. RNA viruses require cDNA synthesis and ssDNA second strand synthesis
Gene conservation	Highly conserved, essential genes (for example, 16s rRNA) enabling broad microbiome studies and surveys of taxa	No homologous genes between viruses of different phyla	Lack of conserved homology between viral phyla prevents universal primer-based surveys of viromes
Culture	Often straightforward to culture and obtain pure, highly enriched bacterial DNA and RNA	Challenging to culture, and require a host cell for replication	Cultured viruses are heavily contaminated with host cell nucleic acids, which decreases viral sequencing output
Clinical specimens	Hardy bacterial cells with cell walls can often be separated from human cells in clinical specimens using differential lysis methods or flow cytometry[144] prior to extraction	Viruses are intracellular pathogens, and although separation from the host is possible (for example, by filtration or antibody pull-down), viruses cannot easily be separated from clinical samples prior to extraction	Clinical specimens are heavily contaminated with host nucleic acids, which decreases viral sequencing output
Methylation patterns	Bacteria use different methylation patterns from eukaryotes; host DNA can be depleted post-extraction using restriction endonucleases that are directed against CpG methylation[145]	DNA viruses are often methylated by the host intracellular machinery, and may have similar methylation patterns	DNA digestion according to methylation patterns is less effective as a means of host depletion for viral sequencing

cDNA, complementary DNA; dsDNA, double-stranded DNA; dsRNA, double-stranded RNA; rRNA, ribosomal RNA; ssDNA, single-stranded DNA; ssRNA, single-stranded RNA.

a known nucleotide sequence has been the most common approach for enriching small viral genomes, such as HIV and influenza virus. Recent examples of PCR amplicon enrichment followed by WGS include phylogenetic analysis of a measles virus outbreak at the 2010 Winter Olympics[85] and tracking the recent Ebola virus[39] and Zika virus (BOX 2) epidemics. PCR amplicon WGS of norovirus, which has a genome size of 7.5 kb, has been used to understand virus transmission in community[86] and hospital[87] settings, which revealed both independent introductions of the pathogen to the hospital and nosocomial transmission despite measures to control infection[87]. Other PCR-based deep-sequencing studies have generated several whole genomes of influenza virus[88] (~13.5 kb), dengue virus[89] (~11 kb) and HCV[90] (9.6 kb). This was feasible because these viruses all have relatively small genomes that require only a few PCR amplicons to assemble whole-genome sequences. However, the heterogeneity of RNA viruses, such as HCV[27], norovirus[86], rabies virus[91] and RSV[38], may necessitate the use of multiple overlapping sets of primers to ensure the amplification of all genotypes. PCR amplicon sequencing is more successful for WGS from samples that have low virus concentrations than metagenomic methods[27], although other methods such as target enrichment of viral sequences may work equally well in such samples, as shown for norovirus samples[92].

Overlapping PCRs combined with NGS have been used to sequence the whole genomes of larger viruses, such as HCMV[93], but this method has limited scalability, as many primers and a relatively large amount of starting DNA are required[93]. This limits the number of suitable samples that are available and also the genomes that can be studied using this method. For example, 8–19 PCR products were required to amplify the genome of Ebola virus[39], and two studies of norovirus needed 14 and 22 PCR products, respectively[86,87]. For clinical applications this is problematic because of the high laboratory workload that is associated with numerous discrete PCR reactions, the necessity for individually normalizing concentrations of different PCR amplicons before pooling, the increasing probability of reaction failure due to primer mismatch (particularly for highly variable viruses), and the high costs of labour and consumables[94]. Therefore, although PCR-based sequencing of viruses as large as 250 kb is technically possible, the proportional relationship between genome size and technical complexity make PCR-based sequencing of viral genomes that are more than 20–50 kb impractical with current technologies, particularly for large multi-sample studies or routine diagnostics. Another consideration is that increasing numbers of PCR reactions require a corresponding increase in sample amount, and this is not always possible as clinical specimens are limited. Improvements in microfluidic technologies may help to overcome some of these barriers; for example, Fluidigm, RainDance and other 'droplet' sequencing technologies. Microfluidics-based PCR and the pooling of multiple amplicons have been used successfully to sequence several antimicrobial-resistance loci (for example, from the microbiome of pigs)[95] and can also be used for viral genomes, potentially down to the single-genome level[96].

Highly variable pathogens, particularly those that have widely divergent genetic lineages or genotypes, such as HCV[97] and norovirus, cause problems for PCR amplification, such as primer amplification[27,92] and primer mismatches[86]. Careful primer design may help to mitigate these problems, but novel variants remain problematic.

Target enrichment. Methods of target enrichment (also known as pull-down, capture or specific enrichment methods) can be used to sequence whole viral genomes directly from clinical samples without the need for prior culture or PCR[98–100]. These methods typically involve small RNA or DNA probes that are complementary to the pathogen reference sequence (or a panel of reference sequences). Unlike specific PCR amplicon-based methods, the reaction can be carried out in a single tube that contains overlapping probes that cover the whole genome. In a hybridization reaction, the probes, which are bound to a solid phase (for example, streptavidin-labelled

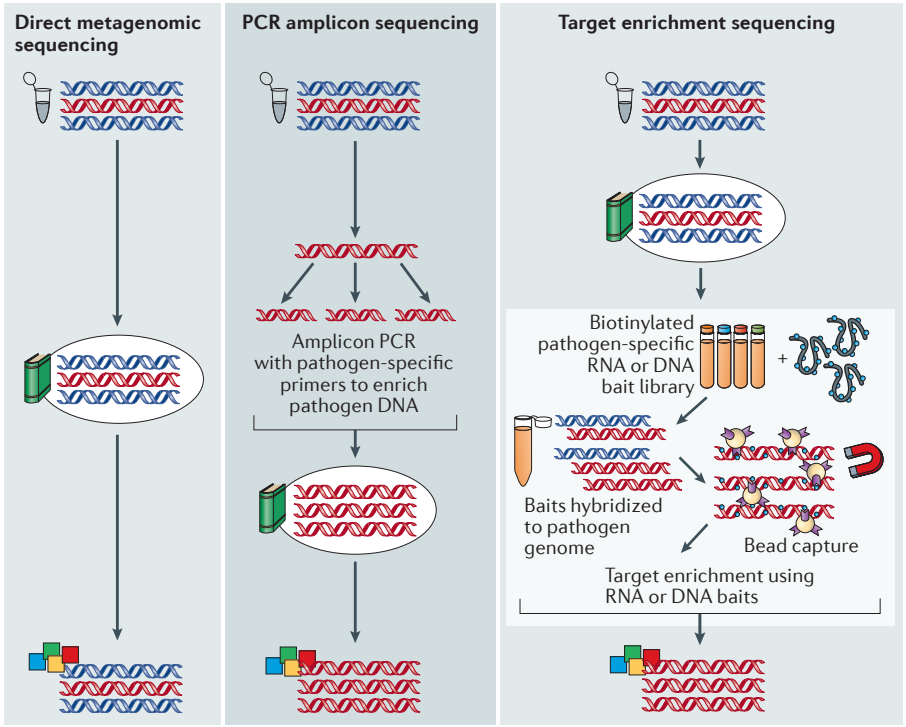

Figure 1 | **Methods for sequencing viral genomes from clinical specimens.** All specimens originally comprise a mix of host (in blue) and pathogen (in red) DNA sequences. For pathogens that have RNA genomes, RNA in the sample is converted into complementary DNA (cDNA) before PCR and library preparation. Direct metagenomic sequencing provides an accurate representation of the sequences in the sample, although at high sequencing and data analysis and storage costs. PCR amplicon sequencing uses many discrete PCR reactions to enrich the viral genome, which increases the workload for large genomes substantially but decreases the costs. Target enrichment sequencing uses virus-specific nucleotide probes that are bound to a solid phase, such as beads, to enrich the viral genome in a single reaction, which reduces workload but increases the cost of library preparation compared with PCR.

magnetic beads), capture or 'pull down' complementary DNA sequences from the total nucleic acids that are present in a sample. Capture is followed by sequencer-specific adaptor ligation and a small number of PCR cycles to enrich for successfully ligated fragments. This has been used successfully to characterize small and large, clinically relevant viruses, such as HCV[27], HSV-1 (REF. 101), VZV[100], EBV[102], HCMV[69], human herpesvirus 6 (HHV6)[103] and HHV7 (REF. 104). The reaction is carried out in a single well and, similarly to microfluidics-based PCR, is amenable to high-throughput automation[102]. The lack of a culture step means that the sequences that are obtained are more representative of original virus rather than cultured virus isolates, and there are fewer mutations than in PCR-amplified templates[69,100]. The success of this method depends on the available reference sequences for the virus of interest; specificity increases when probes are designed against a larger panel of reference sequences, as this leads to better capture of the diversity in and between samples. Target enrichment is possible despite small mismatches between template and probe; however, whereas PCR amplification requires only knowledge of flanking regions of a target region, target enrichment requires knowledge of the internal sequence to design probes. However, if one probe fails, internal and overlapping regions may still be captured by other probes[69,100]. Target enrichment is not suitable for the characterization of novel viruses that have low homology to known viruses for which metagenomics, and, in some cases, PCR using degenerate primers, which are a mix of similar but variable primers, may be more appropriate.

As with all methods, the technique is constrained by the starting virus concentration. Although viruses could be sequenced from samples with viral loads as low as 2,000 International Units (IU) ml^{-1} (for HCV) or 2,500 IU ml^{-1} (for HCMV), there was a reduced depth of coverage in sequencing data at lower virus concentrations[27,69]. With metagenomics, the proportion of sequencing data that map to the pathogen from unenriched clinical samples is small. Target enrichment can increase the percentage of on-target viral reads from 0.01% to 80% or more[69]. The improvement in quality and depth of sequence that results enables more samples to be sequenced per run than unenriched metagenomic libraries for equivalent on-target sequencing performance. This improvement also decreases the price of sequencing, although the cost of library preparation is increased. There are alternative approaches for the enrichment of viral reads, including pulsed-field gel electrophoresis (PFGE)[105], which separates large viral genomes from smaller fragments of host DNA.

Enrichment techniques that use degenerate RNA or DNA probes to capture hundreds of viral species have also been developed; for example, virome capture sequencing (VirCapSeq)[106]. This method is designed for the detection of both known and novel viruses, although its performance remains to be evaluated.

Comparison of methods. To date, there has been very little direct comparison between the three methods for viral genome sequencing in clinical practice, with only one paper evaluating relative performance for the sequencing of HCV[27]. Results from this study, in which three different enrichment protocols, two metagenomic methods and one overlapping PCR method were evaluated, showed that metagenomic methods were the least sensitive, yielded the lowest genome coverage for comparable sequencing effort and were more prone to result in incomplete genome assemblies. The PCR method required repeated amplification and was the most likely to miss mixed infections, but when reactions were successful it resulted in the most consistent read depth, whereas read depth was proportional to virus copy number in metagenomics and target enrichment. PCR generated more incomplete sequences for some HCV genotypes (particularly genotype 2) than metagenomics and target enrichment. Target enrichment was the most consistent method to result in full genomes and identical consensus sequences. The ease of library preparation for metagenomic and target enrichment sequencing of HCV was considered a major advantage for clinical applications, but PCR may still be appropriate for samples that have very low viral loads.

Similar results were achieved in a study that compared PCR amplicon sequencing and target enrichment sequencing of norovirus[92]. With target enrichment sequencing, the whole viral genome could be sequenced in all 164 samples, whereas PCR-based capsid sequencing was only possible in 158 out of the 164 samples, owing to low virus titres and PCR primer mismatches, which suggests that target enrichment is more sensitive than PCR for sequencing norovirus and better accommodates between-strain sequence heterogeneity[92]. Target enrichment has also been used for samples that have low viral loads and incomplete genome coverage in metagenomic sequencing[107]. Both metagenomic and target enrichment sequencing can be used for pathogen genomes of all sizes, whereas PCR-based methods are less suitable for large viral genomes or for non-viral (that is, bacterial, fungal and parasite) genomes.

Direct comparisons of different methods[27,92] will be important for determining when each method should be used, based on sensitivity and specificity, as well as factors such as cost, scalability and turn-around time, which are particularly important in clinical applications (TABLE 1).

Analysis and interpretation challenges

Beyond the technical challenges of viral WGS that are mentioned above, there are several other roadblocks that may slow the advance of WGS in the clinic. They may be considered in three groups: ethical issues, including incidental host and microbiological findings; regulatory issues, such as the establishment of standards, good laboratory practice, and sensitivity and specificity thresholds for sequencing; and analytical issues regarding data interpretation and the numerous choices of analysis options.

Ethical issues and incidental findings. In many clinical tests (for example, magnetic resonance imaging (MRI) scans and sequencing of the genomes of patients), there is a risk of detecting a disease association that is not part of the original investigation but might be of clinical importance for the individual or their family. These incidental findings remain a topic of intense medical ethical debate[108]. The risk of incidental findings in pathogen sequencing (for example, the discovery of HIV infection during metagenomic sequencing for other pathogens) is not novel and the solution in clinical virology laboratories that use multiplex PCRs is to suppress results that have not been requested (J.B., unpublished observations). In the United Kingdom, the clinical virologist who interprets the test results is part of the team that manages the patient, and, as such, may decide to discuss an unexpected result with the physician in charge. Incidental host genetic findings (for example, the detection of variants that predispose to cancer development) in a pathogen metagenomic analysis are not reported to the individual in the United Kingdom, because this is only permissible with the consent of the patient. In regard to both host and virus incidental findings, target enrichment and PCR have the advantage of only providing results about the pathogen of interest. The ethical and privacy concerns that are associated with the presence of host genetic data in publically available metagenomic datasets have been well reviewed[82] and represent a separate challenge.

Regulatory challenges. Regulation, as well as helping to address some of the ethical concerns, is also important in standardizing WGS of viruses. The framework that is required to make viral WGS sufficiently robust and reproducible in clinical practice will come from several areas.

The framework of laboratory accreditation and benchmark testing that are already available (for example, Clinical Laboratory Improvement Amendments of 1988 (CLIA) regulations in the USA, or accreditation according to medical laboratory quality and competence standardization criteria for ISO 15189) will support the development of viral WGS standards, provided that there is sufficient need and pressure to implement clinical viral WGS.

Lessons learned from the use of PCR in diagnostics may be useful here, starting with ensuring good clinical laboratory and molecular practices[109,110]. This will mean including negative samples in every sequencing run to assess contamination thresholds, spiking samples with a known virus to provide a sensitivity threshold, and including positive controls and controls for batch-to-batch variation[111], all of which will increase sequencing costs and are likely to deter the adoption of pathogen genome sequencing by laboratories that are sequencing only small batches of samples. The centralization of virus WGS can help to ensure the maintenance of adequate standards, the processing of large batches of samples and reducing costs.

The issues of sensitivity and contamination are especially important in WGS, because of the risk of both false-negative and false-positive detection of pathogens. Highly sensitive sequencing (whether metagenomic, PCR-based or target enrichment-based) may detect low-level contaminating viral nucleic acids[112,113]. For example, murine leukaemia virus[114,115] and parvovirus-like sequences[116,117] are just two of many contaminants that can come from common laboratory reagents, such as nucleic acid extraction columns[118]. As with other highly sensitive technologies, robust laboratory practices and protocols are required to minimize contamination. It is also important to remember that the detection of viral nucleic acid does not necessarily identify the cause of illness, and it is good practice when using NGS methods for the diagnosis of viral infections to confirm the findings with alternative independent methods that do not rely on testing for nucleic acids. For example, in cases of encephalitis of unknown origin, positive NGS findings can be confirmed through immunohistochemical analysis of the affected tissue[65,119], or identification of the virus by electron microscopy or tissue culture[82].

The standardization of methods, including bioinformatics, will be key to the success of NGS and WGS in clinical virology. Software packages that use a graphical user interface (GUI) are preferable to tools that require command-line expertise. Strict version control of software and analysis pipelines is required to ensure that results are reproducible, to make best practices easily shareable, and to enable the accreditation of analysis software. However, best-practice analysis methods are continually evolving and the premature standardization of best practices in an overly rigid manner may inhibit innovation. Commercialization and regulation may help, as they provide financial and regulatory incentives to ensure that analysis tools and technologies meet clinical needs. Finally, the development of well-curated databases that show which variants are truly indicative of drug resistance will be crucial for accurate clinical interpretation. Such databases have already been created for HIV[120], HBV[121,122] and HCV[123], but without recognition of their value by funding agencies and corresponding centralized funding to ensure their continued maintenance and upkeep, these databases and associated tools may become swiftly outdated or unusable.

Financial barriers to the clinical use of viral WGS. Although there are good reasons for sequencing whole genomes and, in general, for using NGS, if diagnostic or hospital-based laboratories are to be persuaded to transition away from sequencing subgenomic fragments, they need to see the benefit of the additional information for patient care and the practical feasibility of WGS. This includes WGS workflows that are as scalable and automatable as subgenomic fragment sequencing, a suitable regulatory framework and a price for sequencing whole genomes that is competitive with sequencing fragments.

Currently, the cost of sequencing viral genomes, despite their small size, remains higher than the cost of sequencing subgenomic resistance genes. The cost difference between sequencing a target region and the whole virus genome is largely governed by the size of the genome versus the size and number of target loci. In addition, whole-genome information may provide important additional knowledge, as discussed above.

What does the future hold?

Current NGS technologies that are based around Illumina, 454, Ion Torrent or Sanger methodologies all generate short-read data, which presents challenges for haplotype phasing; that is, determining whether genetic variants (whether inter-host or intra-host) occur on the same genetic background (single viral genome or clonal) or on related, highly similar but different genetic backgrounds in the same population (sometimes called a viral swarm or cloud). Furthermore, repetitive regions and recombination are more difficult to resolve using short reads owing to problems such as mapping ambiguities. The clinical implications of understanding whether, for example, multidrug-resistant variants occur together on a single viral genome or are distributed between a mixed population of viruses, each with different drug-resistance profiles, are currently unclear.

Although there are computational tools[124] to help resolve these issues, new technologies can generate longer reads. Newer, single-molecule sequencers, such as PacBio (Pacific Biosciences) and MinION (Oxford Nanopore), are capable of extremely long-read sequencing, and whole viral genomes (for example, viruses that have genomes less than 20 kb in size, such as Ebola virus, norovirus and influenza A virus) could theoretically be obtained from single reads. MinION also has the advantage of being very fast, taking as little as four hours to go from sample receipt to reporting of analysed data[125]. So far, viral read lengths achieved by MinION sequencing have been relatively modest; examples of mean read lengths include 751 bp for modified vaccinia Ankara virus, 758 bp for cowpox virus[126], 455 bp (range 126–1477 bp) for chikungunya virus, 358 bp (range 220–672 bp) for Ebola virus, 1,576 bp[127] (M.A.B., unpublished observations: 6,895bp) for HCMV and 572bp (range 318–792 bp) for HCV[128]. Results from the better-established PacBio technology are more promising, including a recent report of a mean read length of 12,777 bp for pseudorabies virus[129], which has a double-stranded DNA genome that is around 142 kb in length. 9.2 kb reads have been achieved using PacBio for HCV, although 9.2 kb of the 9.6 kb genome had been pre-amplified by PCR[130].

A drawback of both NGS and single-molecule sequencing is the need for high coverage to minimize the effect of sequencing errors. This is particularly problematic for studies of drug resistance, as drug resistance most frequently results from single-nucleotide mutations or small deletions (1–3 bases), especially in lower-fidelity RNA viruses[131]. Achieving the high coverage that is necessary to ensure accurate variant typing is challenging when there is a lot of host DNA compared with viral sequences, and when the error profile of a technology makes point mutations particularly hard to detect[125]. At the time of writing, MinION sequencing (R9 pore chemistry) has raw high quality (so called '2D reads') read error rates of around 5% (J. Quick, personal communication), which compares unfavourably with the error rates of other technologies (Illumina (<0.1%), Ion Torrent (~1%), but not PacBio (13% single pass))[132], although accuracy can been improved using circular consensus read sequencing[133,134].

However, combining these long-read technologies with target enrichment provides a potential way forward[127,135], as ambiguities can be resolved if sufficient depth of sequence is achieved for the target pathogen, and error rates for all methodologies may be decreased by further technological and analytical improvements. Depleting the nucleic acids of the host is an alternative solution, as a higher proportion of virus reads would be recovered from each sequencing run. Although there are already solutions in place to achieve this for bacterial sequencing (for example, depletion of human ribosomal RNA or mitochondria, and selective depletion of DNA with a certain methylation pattern), no similar methods exist, so far, for viral sequencing.

Conclusions

Viral WGS is of increasing clinical importance for diagnosis, disease management, molecular epidemiology and infection control. There are several methods that are available to achieve WGS of viruses from clinical samples; amplicon sequencing, target enrichment or metagenomics. Currently, the choice of method is specific to both the virus and the clinical question. Metagenomic sequencing is most appropriate for diagnostic sequencing of unknown or poorly characterized viruses, PCR amplicon sequencing works well for short viral genomes and low diversity in primer binding sites, and target enrichment works for all pathogen sizes but is particularly advantageous for large viruses and for viruses that have diverse but well-characterized genomes. Two obvious areas of innovation currently exist: methods that can effectively deplete host DNA without affecting viral DNA, and the further development of long-read technologies to achieve the flexibility and competitive pricing of short-read technologies. New technologies are required to unite the strengths of these different methods and enable healthcare providers to invest in a single technology that is suitable for all viral WGS applications.

Charlotte J. Houldcroft is at the Department of Infection, Immunity and Inflammation, Great Ormond Street Institute of Child Health, University College London, London WC1N 1EH, UK; and the Division of Biological Anthropology, University of Cambridge, Cambridge CB2 3QG, UK.

Matthew A. Beale is at the Division of Infection and Immunity, University College London, London WC1E 6BT, UK; and The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

Judith Breuer is at the Division of Infection and Immunity, University College London, London WC1E 6BT, UK; and at Great Ormond Street Hospital for Children NHS Foundation Trust, London WC1N 3JH, UK.

Correspondence to J.B.
j.breuer@ucl.ac.uk

doi:10.1038/nrmicro.2016.182
Published online 16 Jan 2017

1. Gardner, R. C. *et al.* The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Res.* **9**, 2871–2888 (1981).
2. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).

4. Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403 (1995).
5. Hayden, E. C. Technology: the $1,000 genome. *Nature* **507**, 294–295 (2014).
6. Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449**, 804–810 (2007).
7. Grigoriev, I. V. *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42**, D699–D704 (2014).
8. Worthey, E. A. *et al.* Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.* **13**, 255–262 (2011).
9. Bryant, J. M. *et al.* Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. *Lancet* **381**, 1551–1560 (2013).
10. Lamelas, A. *et al.* Emergence of a new epidemic *Neisseria meningitidis* serogroup A clone in the African meningitis belt: high-resolution picture of genomic changes that mediate immune evasion. *mBio* **5**, e01974-14 (2014).
11. Zaraket, H. *et al.* Genetic makeup of amantadine-resistant and oseltamivir-resistant human influenza A/H1N1 viruses. *J. Clin. Microbiol.* **48**, 1085–1092 (2010).
12. Houldcroft, C. J. *et al.* Detection of low frequency multi-drug resistance and novel putative maribavir resistance in immunocompromised pediatric patients with cytomegalovirus. *Front. Microbiol.* **7,** 1317 (2016).
13. Witney, A. A. *et al.* Clinical application of whole-genome sequencing to inform treatment for multidrug-resistant tuberculosis cases. *J. Clin. Microbiol.* **53**, 1473–1483 (2015).
14. Simen, B. B. *et al.* Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. *J. Infect. Dis.* **199**, 693–701 (2009).
15. Smith, G. J. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–1125 (2009).
16. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372 (2014).
17. Koser, C. U. *et al.* Routine use of microbial whole-genome sequencing in diagnostic and public health microbiology. *PLoS Pathog.* **8**, e1002824 (2012).
18. Cartwright, E. J., Koser, C. U. & Peacock, S. J. Microbial sequences benefit health now. *Nature* **471**, 578 (2011).
19. Paredes, R. & Clotet, B. Clinical management of HIV-1 resistance. *Antiviral Res.* **85**, 245–265 (2010).
20. Van Laethem, K., Theys, K. & Vandamme, A. M. HIV-1 genotypic drug resistance testing: digging deep, reaching wide? *Curr. Opin. Virol.* **14**, 16–23 (2015).
21. Durant, J. *et al.* Drug-resistance genotyping in HIV-1 therapy: the VIRADAPT randomised controlled trial. *Lancet* **353**, 2195–2199 (1999).
22. Clevenbergh, P. *et al.* Persisting long-term benefit of genotype-guided treatment for HIV-infected patients failing HAART. The Viradapt Study: week 48 follow-up. *Antivir. Ther.* **5**, 65–70 (2000).
23. Khudyakov, Y. Molecular surveillance of hepatitis C. *Antivir. Ther.* **17**, 1465–1470 (2012).
24. Kim, J. H., Park, Y. K., Park, E. S. & Kim, K. H. Molecular diagnosis and treatment of drug-resistant hepatitis B virus. *World J. Gastroenterol.* **20**, 5708–5720 (2014).
25. McGinnis, J., Laplante, J., Shudt, M. & George, K. S. Next generation sequencing for whole genome analysis and surveillance of influenza A viruses. *J. Clin. Virol.* **79**, 44–50 (2016).
26. Pawlotsky, J. M. Hepatitis C virus resistance to direct-acting antiviral drugs in interferon-free regimens. *Gastroenterology* **151**, 70–86 (2016).
27. Thomson, E. *et al.* Comparison of next generation sequencing technologies for the comprehensive assessment of full-length hepatitis C viral genomes. *J. Clin. Microbiol.* **54**, 2470–2484 (2016).
28. Brunnemann, A. K. *et al.* Drug resistance of clinical varicella-zoster virus strains confirmed by recombinant thymidine kinase expression and by targeted resistance mutagenesis of a cloned wild-type isolate. *Antimicrob. Agents Chemother.* **59**, 2726–2734 (2015).
29. Karamitros, T. *et al. De novo* assembly of human herpes virus type 1 (HHV-1) genome, mining of non-canonical structures and detection of novel drug-resistance mutations using short- and long-read next generation sequencing technologies. *PLoS ONE* **11**, e0157600 (2016).
30. Piret, J. & Boivin, G. Antiviral drug resistance in herpesviruses other than cytomegalovirus. *Rev. Med. Virol.* **24**, 186–218 (2014).
31. Melendez, D. P. & Razonable, R. R. Letermovir and inhibitors of the terminase complex: a promising new class of investigational antiviral drugs against human cytomegalovirus. *Infect. Drug Resist.* **8**, 269–277 (2015).
32. Lassalle, F. *et al.* Islands of linkage in an ocean of pervasive recombination reveals two-speed evolution of human cytomegalovirus genomes. *Virus Evol.* **2,** vew017 (2016).
33. Lanier, E. R. *et al.* Analysis of mutations in the gene encoding cytomegalovirus DNA polymerase in a phase 2 clinical trial of brincidofovir prophylaxis. *J. Infect. Dis.* **214**, 32–35 (2016).
34. Kaverin, N. V. *et al.* Epitope mapping of the hemagglutinin molecule of a highly pathogenic H5N1 influenza virus by using monoclonal antibodies. *J. Virol.* **81**, 12911–12917 (2007).
35. Franco, S. *et al.* Detection of a sexually transmitted hepatitis C virus protease inhibitor-resistance variant in a human immunodeficiency virus-infected homosexual man. *Gastroenterology* **147**, 599–601.e1 (2014).
36. Fujisaki, S. *et al.* Outbreak of infections by hepatitis B virus genotype A and transmission of genetic drug resistance in patients coinfected with HIV-1 in Japan. *J. Clin. Microbiol.* **49**, 1017–1024 (2011).
37. Pierucci, P. *et al.* Novel autologous T-cell therapy for drug-resistant cytomegalovirus disease after lung transplantation. *J. Heart Lung Transplant.* **35**, 685–687 (2016).
38. Agoti, C. N. *et al.* Local evolutionary patterns of human respiratory syncytial virus derived from whole-genome sequencing. *J. Virol.* **89**, 3444–3454 (2015).
39. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).
40. Aanensen, D. M. *et al.* Whole-genome sequencing for routine pathogen surveillance in public health: a population snapshot of invasive *Staphylococcus aureus* in Europe. *mBio* **7,** e00444-16 (2016).
41. Faria, N. R. *et al.* Zika virus in the Americas: early epidemiological and genetic findings. *Science* **352**, 345–349 (2016).
42. Mbisa, J. L. *et al.* Evidence of self-sustaining drug resistant HIV-1 lineages among untreated patients in the United Kingdom. *Clin. Infect. Dis.* **61**, 829–836 (2015).
43. Bartha, I. *et al.* A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *eLife* **2**, e01123 (2013).
44. Power, R. A. *et al.* Genome-wide association study of HIV whole genome sequences validated using drug resistance. *PLoS ONE* **11**, e0163746 (2016).
45. Cuevas, J. M., Geller, R., Garijo, R., Lopez-Aldeguer, J. & Sanjuan, R. Extremely high mutation rate of HIV-1 *in vivo*. *PLoS Biol.* **13**, e1002251 (2015).
46. Vandenhende, M. A. *et al.* Prevalence and evolution of low frequency HIV drug resistance mutations detected by ultra deep sequencing in patients experiencing first line antiretroviral therapy failure. *PLoS ONE* **9**, e86771 (2014).
47. Zhou, B. *et al.* Composition and interactions of hepatitis B virus quasispecies defined the virological response during telbivudine therapy. *Sci. Rep.* **5**, 17123 (2015).
48. Itakura, J. *et al.* Resistance-associated NS5A variants of hepatitis C virus are susceptible to interferon-based therapy. *PLoS ONE* **10**, e0138060 (2015).
49. Rogers, M. B. *et al.* Intrahost dynamics of antiviral resistance in influenza A virus reflect complex patterns of segment linkage, reassortment, and natural selection. *mBio* **6,** e02464-14 (2015).
50. Swenson, L. C., Daumer, M. & Paredes, R. Next-generation sequencing to assess HIV tropism. *Curr. Opin. HIV AIDS* **7**, 478–485 (2012).
51. Hutter, G. *et al.* Long-term control of HIV by *CCR5* delta32/delta32 stem-cell transplantation. *N. Engl. J. Med.* **360**, 692–698 (2009).
52. Kordelas, L. *et al.* Shift of HIV tropism in stem-cell transplantation with *CCR5* delta32 mutation. *N. Engl. J. Med.* **371**, 880–882 (2014).
53. Coaquette, A. *et al.* Mixed cytomegalovirus glycoprotein B genotypes in immunocompromised patients. *Clin. Infect. Dis.* **39**, 155–161 (2004).
54. Solmone, M. *et al.* Use of massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis B virus in drug-resistant and drug-naive patients and to detect minor variants in reverse transcriptase and hepatitis B S antigen. *J. Virol.* **83**, 1718–1726 (2009).
55. Chou, S. *et al.* Improved detection of emerging drug-resistant mutant cytomegalovirus subpopulations by deep sequencing. *Antimicrob. Agents Chemother.* **58**, 4697–4702 (2014).
56. Fonager, J. *et al.* Identification of minority resistance mutations in the HIV-1 integrase coding region using next generation sequencing. *J. Clin. Virol.* **73**, 95–100 (2015).
57. Kyeyune, F. *et al.* Low-frequency drug resistance in HIV-infected Ugandans on antiretroviral treatment is associated with regimen failure. *Antimicrob. Agents Chemother.* **60**, 3380–3397 (2016).
58. Liu, P. *et al.* Direct sequencing and characterization of a clinical isolate of Epstein–Barr virus from nasopharyngeal carcinoma tissue by using next-generation sequencing technology. *J. Virol.* **85**, 11291–11299 (2011).
59. Loman, N. J. & Pallen, M. J. Twenty years of bacterial genome sequencing. *Nat. Rev. Microbiol.* **13**, 787–794 (2015).
60. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
61. Mulcahy-O'Grady, H. & Workentine, M. L. The challenge and potential of metagenomics in the clinic. *Front. Immunol.* **7**, 29 (2016).
62. Hoffmann, B. *et al.* A variegated squirrel bornavirus associated with fatal human encephalitis. *N. Engl. J. Med.* **373**, 154–162 (2015).
63. Perlejewski, K. *et al.* Next-generation sequencing (NGS) in the identification of encephalitis-causing viruses: unexpected detection of human herpesvirus 1 while searching for RNA pathogens. *J. Virol. Methods* **226**, 1–6 (2015).
64. Duncan, C. J. *et al.* Human IFNAR2 deficiency: lessons for antiviral immunity. *Sci. Transl Med.* **7**, 307ra154 (2015).
65. Morfopoulou, S. *et al.* Human coronavirus OC43 associated with fatal encephalitis. *N. Engl. J. Med.* **375**, 497–498 (2016).
66. Naccache, S. N. *et al.* Diagnosis of neuroinvasive astrovirus infection in an immunocompromised adult with encephalitis by unbiased next-generation sequencing. *Clin. Infect. Dis.* **60**, 919–923 (2015).
67. Huang, W. *et al.* Whole-genome sequence analysis reveals the enterovirus D68 isolates during the United States 2014 outbreak mainly belong to a novel clade. *Sci. Rep.* **5**, 15223 (2015).
68. Wilson, M. R. *et al.* Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N. Engl. J. Med.* **370**, 2408–2417 (2014).
69. Depledge, D. P. *et al.* Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS ONE* **6**, e27805 (2011).
70. Allen, U. D. *et al.* The genetic diversity of Epstein–Barr virus in the setting of transplantation relative to non-transplant settings: a feasibility study. *Pediatr. Transplant.* **20**, 124–129 (2016).
71. Matranga, C. B. *et al.* Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol.* **15**, 519 (2014).
72. Calvet, G. *et al.* Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: a case study. *Lancet Infect. Dis.* **16**, 653–660 (2016).
73. Lei, H. *et al.* Epstein–Barr virus from Burkitt Lymphoma biopsies from Africa and South America share novel LMP-1 promoter and gene variations. *Sci. Rep.* **5**, 16706 (2015).
74. Kohl, C. *et al.* Protocol for metagenomic virus detection in clinical specimens. *Emerg. Infect. Dis.* **21**, 48–57 (2015).
75. Sauvage, V. & Eloit, M. Viral metagenomics and blood safety. *Transfus. Clin. Biol.* **23**, 28–38 (2016).
76. Lecuit, M. & Eloit, M. The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening. *Front. Cell. Infect. Microbiol.* **4**, 25 (2014).
77. Oude Munnink, B. B. *et al.* Autologous antibody capture to enrich immunogenic viruses for viral discovery. *PLoS ONE* **8**, e78454 (2013).
78. Sabina, J. & Leamon, J. H. Bias in whole genome amplification: causes and considerations. *Methods Mol. Biol.* **1347**, 15–41 (2015).

79. Jensen, R. H. *et al.* Target-dependent enrichment of virions determines the reduction of high-throughput sequencing in virus discovery. *PLoS ONE* **10**, e0122636 (2015).
80. Denesvre, C., Dumarest, M., Remy, S., Gourichon, D. & Eloit, M. Chicken skin virome analyzed by high-throughput sequencing shows a composition highly different from human skin. *Virus Genes* **51**, 209–216 (2015).
81. Mlakar, J. *et al.* Zika virus associated with microcephaly. *N. Engl. J. Med.* **374**, 951–958 (2016).
82. Hall, R. J., Draper, J. L., Nielsen, F. G. & Dutilh, B. E. Beyond research: a primer for considerations on using viral metagenomics in the field and clinic. *Front. Microbiol.* **6**, 224 (2015).
83. Greninger, A. L. *et al.* A novel outbreak enterovirus D68 strain associated with acute flaccid myelitis cases in the USA (2012–2014): a retrospective cohort study. *Lancet Infect. Dis.* **15**, 671–682 (2015).
84. Breitwieser, F. P., Pardo, C. A. & Salzberg, S. L. Re-analysis of metagenomic sequences from acute flaccid myelitis patients reveals alternatives to enterovirus D68 infection. *F1000Res.* **4**, 180 (2015).
85. Gardy, J. L. *et al.* Whole-genome sequencing of measles virus genotypes H1 and D8 during outbreaks of infection following the 2010 Olympic Winter Games reveals viral transmission routes. *J. Infect. Dis.* **212**, 1574–1578 (2015).
86. Cotten, M. *et al.* Deep sequencing of norovirus genomes defines evolutionary patterns in an urban tropical setting. *J. Virol.* **88**, 11056–11069 (2014).
87. Kundu, S. *et al.* Next-generation whole genome sequencing identifies the direction of norovirus transmission in linked patients. *Clin. Infect. Dis.* **57**, 407–414 (2013).
88. Watson, S. J. *et al.* Molecular epidemiology and evolution of influenza viruses circulating within European swine between 2009 and 2013. *J. Virol.* **89**, 9920–9931 (2015).
89. Parameswaran, P. *et al.* Genome-wide patterns of intrahuman dengue virus diversity reveal associations with viral phylogenetic clade and interhost diversity. *J. Virol.* **86**, 8546–8558 (2012).
90. Newman, R. M. *et al.* Whole genome pyrosequencing of rare hepatitis C virus genotypes enhances subtype classification and identification of naturally occurring drug resistance variants. *J. Infect. Dis.* **208**, 17–31 (2013).
91. Jakava-Viljanen, M. *et al.* Evolutionary trends of European bat lyssavirus type 2 including genetic characterization of Finnish strains of human and bat origin 24 years apart. *Arch. Virol.* **160**, 1489–1498 (2015).
92. Brown, J. R. *et al.* Norovirus whole genome sequencing by SureSelect target enrichment: a robust and sensitive method. *J. Clin. Microbiol.* **54**, 2530–2537 (2016).
93. Renzette, N., Bhattacharjee, B., Jensen, J. D., Gibson, L. & Kowalik, T. F. Extensive genome-wide variability of human cytomegalovirus in congenitally infected infants. *PLoS Pathog.* **7**, e1001344 (2011).
94. Bialasiewicz, S. *et al.* Detection of a divergent parainfluenza 4 virus in an adult patient with influenza like illness using next-generation sequencing. *BMC Infect. Dis.* **14**, 275 (2014).
95. Johnson, T. A. *et al.* Clusters of antibiotic resistance genes enriched together stay together in swine agriculture. *mBio* **7**, e02214-15 (2016).
96. Ocwieja, K. E. *et al.* Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic Acids Res.* **40**, 10345–10355 (2012).
97. Bonsall, D. *et al.* ve-SEQ: Robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens. *F1000Res.* **4**, 1062 (2015).
98. Wylie, T. N., Wylie, K. M., Herter, B. N. & Storch, G. A. Enhanced virome sequencing using targeted sequence capture. *Genome Res.* **25**, 1910–1920 (2015).
99. Tsangaras, K. *et al.* Hybridization capture using short PCR products enriches small genomes by capturing flanking sequences (CapFlank). *PLoS ONE* **9**, e109101 (2014).
100. Depledge, D. P. *et al.* Deep sequencing of viral genomes provides insight into the evolution and pathogenesis of varicella zoster virus and its vaccine in humans. *Mol. Biol. Evol.* **31**, 397–409 (2014).
101. Ebert, K., Depledge, D. P., Breuer, J., Harman, L. & Elliott, G. Mode of virus rescue determines the acquisition of VHS mutations in VP22-negative herpes simplex virus 1. *J. Virol.* **87**, 10389–10393 (2013).
102. Palser, A. L. *et al.* Genome diversity of Epstein–Barr virus from multiple tumor types and normal infection. *J. Virol.* **89**, 5222–5237 (2015).
103. Tweedy, J. *et al.* Complete genome sequence of the human herpesvirus 6A strain AJ from Africa resembles strain GS from North America. *Genome Announc.* **3**, e01498-14 (2015).
104. Donaldson, C. D., Clark, D. A., Kidd, I. M., Breuer, J. & Depledge, D. D. Genome sequence of human herpesvirus 7 strain UCL-1. *Genome Announc.* **1**, e00830-13 (2013).
105. Kamperschroer, C., Gosink, M. M., Kumpf, S. W., O'Donnell, L. M. & Tartaro, K. R. The genomic sequence of lymphocryptovirus from cynomolgus macaque. *Virology* **488**, 28–36 (2016).
106. Briese, T. *et al.* Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *mBio* **6**, e01491-15 (2015).
107. Naccache, S. N. *et al.* Distinct Zika virus lineage in Salvador, Bahia, Brazil. *Emerg. Infect. Dis.* **22**, 1788–1792 (2016).
108. Hofmann, B. Incidental findings of uncertain significance: to know or not to know — that is not the question. *BMC Med. Ethics* **17**, 13 (2016).
109. Public Health England. Good laboratory practice when performing molecular amplification assays. *Public Health England* https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/344076/Q_4i4.4.pdf (2013).
110. Viana, R. V. & Wallis, C. L. in *Wide Spectra of Quality Control* Ch. 3 (ed Akyar, I.) (InTech, 2011).
111. Blomquist, T., Crawford, E. L., Yeo, J., Zhang, X. & Willey, J. C. Control for stochastic sampling variation and qualitative sequencing error in next generation sequencing. *Biomol. Detect. Quantif.* **5**, 30–37 (2015).
112. Houldcroft, C. J. & Breuer, J. Tales from the crypt and coral reef: the successes and challenges of identifying new herpesviruses using metagenomics. *Front. Microbiol.* **6**, 188 (2015).
113. Munro, A. C. & Houldcroft, C. Human cancers and mammalian retroviruses: should we worry about bovine leukemia virus? *Future Virol.* **11**, 163–166 (2016).
114. Hue, S. *et al.* Disease-associated XMRV sequences are consistent with laboratory contamination. *Retrovirology* **7**, 111 (2010).
115. Erlwein, O. *et al.* DNA extraction columns contaminated with murine sequences. *PLoS ONE* **6**, e23484 (2011).
116. Rosseel, T., Pardon, B., De Clercq, K., Ozhelvaci, O. & Van Borm, S. False-positive results in metagenomic virus discovery: a strong case for follow-up diagnosis. *Transbound. Emerg. Dis.* **61**, 293–299 (2014).
117. Naccache, S. N. *et al.* The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J. Virol.* **87**, 11966–11977 (2013).
118. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).
119. Lipkin, W. I. A. Vision for investigating the microbiology of health and disease. *J. Infect. Dis.* **212** (Suppl. 1), S26–S30 (2015).
120. Shafer, R. W. Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis.* **194** (Suppl. 1), S51–S58 (2006).
121. Gnaneshan, S., Ijaz, S., Moran, J., Ramsay, M. & Green, J. HepSEQ: international public health repository for hepatitis B. *Nucleic Acids Res.* **35**, D367–D370 (2007).
122. Rhee, S. Y. *et al.* Hepatitis B virus reverse transcriptase sequence variant database for sequence analysis and mutation discovery. *Antiviral Res.* **88**, 269–275 (2010).
123. Kuiken, C., Yusim, K., Boykin, L. & Richardson, R. The Los Alamos hepatitis C sequence database. *Bioinformatics* **21**, 379–384 (2005).
124. Hong, L. Z. *et al.* BAsE-Seq: a method for obtaining long viral haplotypes from short sequence reads. *Genome Biol.* **15**, 517 (2014).
125. Schmidt, K. *et al.* Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J. Antimicrob. Chemother.* **72**, 104–114 (2017).
126. Kilianski, A. *et al.* Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *Gigascience* **4**, 12 (2015).
127. Eckert, S. E., Chan, J. Z.-M., Houniet, D., Breuer, J. & Speight, G. Enrichment of long DNA fragments from mixed samples for Nanopore sequencing. Preprint at *bioRxiv* http://dx.doi.org/10.1101/048850 (2016).
128. Greninger, A. L. *et al.* Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med.* **7**, 99 (2015).
129. Mathijs, E., Vandenbussche, F., Verpoest, S., De Regge, N. & Van Borm, S. Complete genome sequence of pseudorabies virus reference strain NIA3 using single-molecule real-time sequencing. *Genome Announc.* **4**, e00440-16 (2016).
130. Bull, R. A. *et al.* A method for near full-length amplification and sequencing for six hepatitis C virus genotypes. *BMC Genomics* **17**, 247 (2016).
131. Kimberlin, D. W. & Whitley, R. J. Antiviral resistance: mechanisms, clinical significance, and future implications. *J. Antimicrob. Chemother.* **37**, 403–421 (1996).
132. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
133. Li, C. *et al.* INC-Seq: accurate single molecule reads using nanopore sequencing. *Gigascience* **5**, 34 (2016).
134. Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S. & Turner, S. W. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* **38**, e159 (2010).
135. Karamitros, T. & Magiorkinis, G. A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits. *Nucleic Acids Res.* **43**, e152 (2015).
136. Guan, H. *et al.* Detection of virus in CSF from the cases with meningoencephalitis by next-generation sequencing. *J. Neurovirol.* **22**, 240–245 (2016).
137. Morfopoulou, S. *et al.* Deep sequencing reveals persistence of cell-associated mumps vaccine virus in chronic encephalitis. *Acta Neuropathol.* http://dx.doi.org/10.1007/s00401-016-1629-y (2016).
138. Brown, J. R. *et al.* Astrovirus VA1/HMO-C: an increasingly recognized neurotropic pathogen in immunocompromised patients. *Clin. Infect. Dis.* **60**, 881–888 (2015).
139. Fremond, M. L. *et al.* Next-generation sequencing for diagnosis and tailored therapy: a case report of astrovirus-associated progressive encephalitis. *J. Pediatric Infect. Dis. Soc.* **4**, e53–e57 (2015).
140. Barjas-Castro, M. L. *et al.* Probable transfusion-transmitted Zika virus in Brazil. *Transfusion* **56**, 1684–1688 (2016).
141. Musso, D. *et al.* Potential for Zika virus transmission through blood transfusion demonstrated during an outbreak in French Polynesia, November 2013 to February 2014. *Euro Surveill.* **19**, 20761 (2014).
142. Ellison, D. W. *et al.* Complete genome sequences of Zika virus strains isolated from the blood of patients in Thailand in 2014 and the Philippines in 2012. *Genome Announc.* **4**, e00359-16 (2016).
143. Faria, N. R. *et al.* Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.* **8**, 97 (2016).
144. Chitsaz, H. *et al.* Efficient *de novo* assembly of single-cell bacterial genomes from short-read data sets. *Nat. Biotechnol.* **29**, 915–921 (2011).
145. Feehery, G. R. *et al.* A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PLoS ONE* **8**, e76096 (2013).

Acknowledgements

The authors thank J. Brown and K. Gilmour (Great Ormond Street Hospital (GOSH), London, UK) and R. Doyle (University College London (UCL), UK) for their helpful discussions, and J. Quick (University of Birmingham, UK) for sharing unpublished MinION statistics. C.J.H. was funded by Action Medical Research (grant GN2424). M.A.B. was funded through the European Union's Seventh Programme for research, technological development and demonstration under grant agreement number 304875 held by J.B. This work was supported by the UK National Institute for Health Research Biomedical Research Centre at GOSH National Health Service (NHS) Foundation Trust and UCL. J.B. receives funding from the University College London Hospitals (UCLH)/UCL National Institute for Health Research Biomedical Research Centre. The authors acknowledge infrastructure support for the UCL Pathogen Genomics Unit, from the UCL UK Medical Research Council (MRC) Centre for Molecular Medical Virology and the UCLH/UCL National Institute for Health Research Biomedical Research Centre. The funders had no role in study design, data collection and interpretation, or the decision to submit work for publication.

Competing interests statement

The authors declare no competing interests.

FURTHER INFORMATION

Omicsomics blogspot article: http://omicsomics.blogspot.co.uk/2015/07/leaky-clinical-metagenomics-pipelines.html

ALL LINKS ARE ACTIVE IN THE ONLINE PDF