Clinical and biological insights from viral genome sequencing

Charlotte J. Houldcroft[1,2], Mathew A. Beale[3,4] and Judith Breuer[3, 5]

1. Infection, Immunity and Inflammation, Great Ormond Street Institute of Child Health, University College London, UK
2. Division of Biological Anthropology, University of Cambridge, UK
3. Division of Infection and Immunity, University College London, UK
4. The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
5. Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK.

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

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 will be 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 used in virology and summarize their advantages and disadvantages. Although there are only partially addressed technical, financial and ethical issues with 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*[3]) and enabled by the rapidly falling cost of high-throughput sequencing[4], genomics has changed our understanding of human and pathogen biology. Several

large projects that aim to systematically analyse microbial genomes have been recently completed or are on-going (for example, sequencing thousands of microbiomes[5] and fungal genomes[5,6]); these projects are shaping our knowledge of the genetic variation present in pathogen populations, the genetic changes that underlie disease, and the diversity of microorganisms with which we share our environments.

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[7] and pathogens. For example, WGS has been used to identify new transmission routes of *Mycobacterium abscessus*[8] in healthcare facilities (nosocomial transmission) and to understand *Neisseria meningitidis* epidemics in Africa[9], whereas sequencing of partial genomes has been used to detect drug resistance in RNA viruses such as influenza virus[10] and DNA viruses such as human cytomegalovirus (HCMV)[11]. Viral genome sequencing is becoming ever more important, above all in clinical research and epidemiology. WGS of pathogens has the advantage of detecting all known drug resistance variants in a single test, whereas deep sequencing (that is, sequencing at high coverage) can identify low levels of drug resistance variants to allow intervention before resistance becomes clinically apparent[12,13]. Whole genomes also provide good data with which to identify linked infections for public health and infection control purposes[14,15]. 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[16,17].

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 causing encephalitis (Box 1); and 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.

[H1] Why sequence viruses in the clinic?

For small viruses such as human immunodeficiency virus (HIV), influenza virus, hepatitis B virus (HBV) and hepatitis C virus (HCV), sequencing of partial genomes has been widely used for research, but it also has important clinical applications. One of the main applications and reasons to sequence 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 resistance variants. HAART has dramatically improved survival of patients with HIV, but successful therapy requires long-term suppression of viral replication with antiretroviral drugs, which may be prevented by impaired host immunity, sub-optimal drug penetration in certain tissue compartments and incomplete adherence of patients to therapy[18]. When viral replication continues despite treatment the high mutation rate of HIV enables resistance 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[19], particularly when patients are first diagnosed and when viral loads indicate treatment failure. Sequencing resistance variants has allowed targeted changes in treatment, and this has resulted in greater reductions in virus loads compared with standard care (undetectable HIV load in 32% vs 14% of patients after six months)[20,21]. Thus, sequencing resistance variants to guide HIV treatment

improves disease outcomes. Similar approaches have been used to identify resistance variants in HCV[22], HBV[23], and influenza virus[24].

[H1] 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 detecting a limited number of antiviral resistance variants, WGS has been too costly and labour-intensive to use as compared with sequencing only the specific genes targeted by the drugs. However, the increasing numbers of resistance genes that are located across viral genome, together with falling 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, NS5B) and these genes encompass more than 50% of the viral genome[25]. Individually sequencing each of these genes can be as expensive and time-consuming as WGS[26]. Partial genome sequencing is particularly problematic for large viral genomes, in particular those of the herpesviruses HCMV[11], varicella zoster virus (VZV)[27], herpes simplex virus type 1 (HSV-1)[28] and HSV-2[29]. These viruses have traditionally been treated with drugs targeting the viral thymidine or serine-threonine protein kinases and DNA polymerase. However, the growing numbers of drugs in development that interact with different proteins encoded by viral genes scattered across the genome, means that targeted sequencing for resistance testing is costly, involves more PCR reactions increasing the chance of failure, requires more starting material, is more labour intensive and generally less tractable for diagnostic use[30]. Sequencing the whole genome simultaneously captures all resistance variants and obviates the need to design and optimise PCR assays for detecting 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 anti-viral resistance testing in a single test[11]. At the same time WGS can provide information on antigenic epitopes, viral evolution within a patient over time[11], and evidence of recombination between HCMV strains[31]. WGS can also detect putative novel drug resistance variants and predict changes to epitopes although phenotypic testing of variants is required to confirm clinical resistance[32] and to map epitope changes[33].

As pre-existing resistance to antiviral drugs increases (for example, HCV that is resistant to protease inhibitors[34] and HBV that is resistant to nucleoside analogue reverse-transcriptase inhibitor [35]), WGS will provide the comprehensive resistance data required for selecting appropriate treatment. The complete knowledge of all resistance variants can also support novel clinical management decisions; for example, identification of extensive genome-wide HCMV drug resistance within a patient supported the decision to treat the patient with autologous cytomegalovirus-specific T cells instead of antivirals[36].

WGS may also better identify transmission events and outbreaks, which is not always possible with sequences of sub-genomic fragments. For example, WGS of respiratory syncytial virus (RSV) identified variation outside the gene that is traditionally used for genotyping, and such information could be used to track outbreaks within households, when the genetic variability in single genes is too low for transmission studies[37]. The high number of phylogenetically informative variant sites that can be obtained from full-length or near full-length genomes obviates the need for high quality sequences, which allowed robust linking of Ebola cases and public health interventions in real time during the 2015 epidemic[38]. Another example of WGS supporting public health efforts is Zika virus (Box 2).

Using pathogen WGS routinely for diagnostic purposes[39] is likely to have wider clinical and research benefits. For example, Zika sequences generated for epidemiological purposes been used to inform public health decisions[40], whilst HIV genomes sequenced to identify antiviral

resistant variants have also be used to study viral evolution[41] 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[42,43].

[H1] Why do we need deep sequencing?

Modern methods, which make use of massively parallel sequencing, allow better examination of diversity and analysis of viral 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, retro-transcribing DNA viruses and retroviruses, because they typically show high diversity, even in a single host. HIV is the classic example; the HIV reverse transcriptase is error-prone and introduces mutations at an extremely high rate ($4.1 \pm 1.7 \times 10^{-3}$ per base per cell)[44]. Not one, but 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 viral diversity. The presence of a mixed population of viruses introduces problems for determining the true consensus 'majority' sequence, but these minority (non-consensus) variants may also change the clinical phenotype of the virus, and can predict changes in genotype, tropism or drug resistance. For example, a minor variant conferring drug resistance in HIV present at 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[45]. Similar changes in frequency of resistance-associated alleles during treatment have been observed for HBV[46], HCV[47], HCMV[11] and influenza virus[48].

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 cellular co-receptor usage as R5 (C-C chemokine receptor type 5 (CCR5)-

using), X4 (C-X-C chemokine receptor type 4 (CXCR4)-using) or R5X4 (dual tropism). Maraviroc is a CCR5 antagonist, which blocks infection of R5-tropic HIV, but not of X4 and R5X4 HIV. Just a 2% frequency of X4 or R5X4 genotypes is predictive of maraviroc treatment failure[49]. Sub-consensus frequencies of X4 or R5X4 HIV are also important for the success[50] and failure[51] of bone marrow transplants from CCR5-deficient (CCR5-Δ32) donors and this information may influence the decision to stop antiviral therapy in these patients[50].

Minority variants and identification of haplotypes can also be used to detect mixed infections. Infections with different HCMV genotypes or super-infections[52] are associated with poor clinical outcomes and detection of such mixed infections by WGS might justify more aggressive treatment.

Sanger sequencing of a virus population can detect minority variants at frequencies of between 10 and 40%[53], whereas NGS can sequence those same PCR amplicons to a much greater depth[54], and consequently capture more of the variability present. Sensitivity and specificity are specific for the analysed virus and the sequencing method. Many studies of HIV drug resistance that use deep-sequencing of PCR amplicons require minority variants to be present at >1%, to reduce the possibility of false positives[55,56]. This may miss drug resistance mutations at frequencies of 0.1%-1% and lead to poor treatment outcome[56]. Although a 1-2% frequency threshold (or lower) may be clinically relevant to detect 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 herpesvirues (discussed below). Large cohorts of patients will need to be tested before, during and after treatment[45,49] to establish thresholds for minority drug resistance[11] and vaccine escape variants that are clinically relevant for each virus.

Direct deep sequencing of clinical material, either by shotgun or RNAseq methods (so called metagenomic methods), also allows the unbiased detection and diagnosis of pathogens and provides an alternative to culture, electron microscopy and qPCR (see below).

[H1] Practical considerations

Sequencing viral nucleic acids, whether from cultures or directly from clinical specimens, is complicated by the presence of contaminating host DNA[57]. By contrast, most bacterial sequencing is currently performed on clinical isolates (i.e. cultured), thus sample preparation is comparatively straightforward (Table 2 and reviewed elsewhere[58]). Currently, genome sequencing of viruses can be achieved by ultradeep sequencing or by enriching for viral nucleic acids before sequencing, either directly or by concentrating viral particles. All 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 (Figure 1).

[H3] Metagenomics. Metagenomic approaches have been extensively used for pathogen discovery and for characterising microbial diversity in environmental and clinical samples[59,60]. Total DNA and/or RNA, including from host, bacteria, viruses, fungi and other pathogens, are extracted from a sample, a library is prepared and sequenced by 'shotgun' or RNA-seq. Box 1 explores the diagnostic applications for metagenomics and RNAseq, for example in encephalitis of unknown aetiology[61-63], for which conventional methods such as PCR often are not diagnostic whereas metagenomics and RNAseq have detected viral infections[64,65,66] and other causes[67] of encephalitis. In addition, these methods have been used to sequence some whole viral genomes, including Epstein-Barr virus (EBV)[68] and HCV[26]. However, the presence of contaminating nucleic acids from the host and commensal microorganisms[57] (Table 2) in clinical specimens reduces sensitivity. The proportion of reads matching the viral target genome from metagenomic WGS is often low; for example, 0.008% for EBV in the blood of a

healthy adult[69], 0.0003% for lassa virus in clinical samples[70] and 0.3% for Zika virus in a sample that was enriched for viral particles by filtration and centrifugation[71]. The read depth is often inadequate to detect resistance[26] and the cost is high. Thus metagenomic sequencing has typically only been performed on a small number of samples for research purposes[71,72]. Concentration of viral particles (see Zika virus example above[71]), depletion of host material and/or sequencing to high read depth can increase the amount of viral sequence, but all of these methods add to the cost. Concentrating viral particles from clinical specimens by antibody-mediated pulldown (for example virus discovery based on cDNA-AFLP (amplified fragment length polymorphism); VIDISCA), filtration, ultracentrifugation, and depletion of free nucleic acids, which mostly come from the host have all been tried[73-76]; however, these methods may also reduce 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 phi 29 polymerases, can increase the DNA yield. However, these approaches are time consuming, costly, and may increase the risk of biases, error and contamination without necessarily improving sensitivity[77,78]. Moreover, the proportion of host reads often remains high[79].

When metagenomic methods are used for pathogen discovery or diagnosis, it is critical to have 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[26] as it allows 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[80]. For virus-associated cancers, metagenomics can inform clinical care, provide

information on cancer evolution and generate high coverage data of integrated virus genomes[68]. However, incidental findings, both in host and microbial sequences, may also present ethical and even diagnostic dilemmas for clinical metagenomics[81] (see below). A recent example was a cluster of cases of acute flaccid myelitis that were associated with enterovirus D68[82]. The metagenomic data from patient samples showed the presence of alternative pathogens, some of which are treatable, and was debated in formal[83] and informal scientific channels (http://omicsomics.blogspot.co.uk/2015/07/leaky-clinical-metagenomics-pipelines.html). Regulation and reporting frameworks will be important to resolve future issues of this kind.

[H3] 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 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[84] and tracking the recent Ebola virus[38] and Zika virus (Box 2) epidemics. PCR amplicon WGS of norovirus, which has a genome size of 7.5kb, has been used to understand transmission in community[85] and hospital[86] settings, which revealed both independent introductions of the pathogen to the hospital and nosocomial transmission despite measures to control infection[86]. Other PCR-based deep sequencing studies have generated multiple whole genomes of influenza virus[87] (~13.5kb), dengue virus[88] (~11kb), and HCV[89] (9.6kb). This was feasible because these viruses all have relatively small genomes, requiring only few PCR amplicons to assemble whole-genome sequences. However, the heterogeneity of RNA viruses, such as HCV[26], norovirus[85], rabies virus[90] and RSV[37] may necessitate the use of multiple overlapping sets of primers to ensure amplification of all genotypes. PCR amplicon sequencing is more successful for WGS from samples with low virus

concentrations than metagenomic methods[26], although other methods such as target enrichment of viral sequences may work equally well in such samples, as shown for norovirus samples[91]. Overlapping PCRs combined with NGS have been used to sequence the whole genomes of larger viruses such as HCMV[92], but this method has limited scalability, as many primers and a relatively large amount of starting DNA are needed[92]. This limits the number of suitable samples available and also the genomes which can be studied with this method. For example, 8 to 19 PCR products were needed to amplify the genome of Ebola virus[38], and two studies of norovirus needed 14 and 22 PCR products respectively[85,86]. For clinical applications this is problematic because of the high laboratory workload associated with large numbers of discrete PCR reactions, the necessity for individually normalising concentrations of different PCR amplicons prior to pooling, the increasing probability of reaction failure due to primer mismatch, particularly for very variable viruses, and the high costs of labour and consumables[93]. 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 beyond 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 pooling of multiple amplicons have been used successfully to sequence multiple antimicrobial resistance loci, for example from the microbiome of pigs[94], and can also be used for viral genomes, potentially down to the single-genome level [95].

Highly variable pathogens, particularly those with widely divergent genetic lineages or genotypes, such as HCV[96] and norovirus cause problems for PCR amplification, such as primer

amplification[26,91] and primer mismatches[85]. Careful primer design may help to mitigate these problems, but novel variants remain problematic.

[H3] Target enrichment. Methods of target enrichment (also known as pulldown, 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[97-99]. These methods typically involve small RNA or DNA probes that are complementary to the pathogen reference sequence (or panel of reference sequences). Unlike in 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 hybridisation reaction, the probes, which are bound to a solid phase (for example, streptavidin-labelled magnetic beads) capture or 'pull down' complementary DNA sequences from the total nucleic acids 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 characterise small and large, clinically relevant viruses such as HCV[26], HSV1[100], VZV[99], EBV[101], CMV[68], human herpesvirus 6 (HHV6)[102] and HHV7[103]. The reaction is performed in a single well and, like microfluidics-based PCR, is amenable to high throughput automation[101]. The lack of a culture step means that the sequences obtained are more representative of original virus than cultured viral isolates, and there are fewer mutations than in PCR amplified templates[68,99]. 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 within and between samples. Target enrichment is possible despite small mismatches between template and probe, but, 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[68,99]. Target enrichment is not suitable for characterisation of novel viruses with 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 starting viral concentration. Although virus could be sequenced from samples with viral loads as low as 2000 International Units (IU)/ml (HCV) or 2500 IU/ml (HCMV), there was reduced depth of coverage in sequencing data at lower viral concentration[26,68]. With metagenomics, the proportion of sequencing data mapping to the pathogen from unenriched of clinical samples is small. Target enrichment can increase the percentage of on-target viral reads from 0.01% up to 80% or more[68]. The improvement in quality and depth of sequence that results allows more samples to be sequenced per run compared to unenriched metagenomic libraries for equivalent on-target sequencing performance, and decreases the price of sequencing, although the cost of library preparation is increased. There are alternative approaches for enriching viral reads, including pulse-field gel electrophoresis[104], which separates large viral genomes from smaller fragments of host DNA fragments.

Enrichment techniques that use degenerate RNA or DNA probes to capture hundreds of viral species, for example, VirCapSeq[105], have also been developed. This method is designed for detection of both known and novel viruses, although its performance remains to be evaluated.

[H3] 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 HCV sequencing[26]. 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, compared to metagenomics and target enrichment, generated more incomplete sequences for some HCV genotypes (particularly genotype 2). 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 very low virus load samples.

Similar results were achieved in a study that compared norovirus PCR amplicon and target enrichment sequencing of norovirus[91]. With target enrichment, the whole viral genome could be sequenced in 164/164 samples, whereas PCR-based capsid sequencing was only possible in 158/164 samples, owing to low virus titres and PCR primer mismatches, which suggests that target enrichment is more sensitive than PCR for norovirus sequencing and better accommodates between-strain sequence heterogeneity[91]. Target enrichment has also been used for samples with low virus loads and incomplete genome coverage in metagenomic sequencing[106]. Both metagenomics and target enrichment 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[26,91] 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).

[H1] Analysis and interpretation challenges

Beyond the technical challenges of viral WGS that are mentioned above, there are a number of other roadblocks which 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.

[H3] Ethical issues and incidental findings. In many clinical tests (for example, MRI scans and sequencing of the patient's genome), there is a risk of detecting a disease association that was 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[107]. 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 (personal communication, J.B.). In the UK, the clinical virologist who interprets the test results is part of the team managing the patient and as such may decide to discuss an unexpected result with the physician-in-charge. Incidental host genetic findings (for example, detection of variants that predispose to cancer development) in a pathogen metagenomic analysis are not reported to the individual in the UK, because this is only permissible with patient consent. In regard to both host and virus incidental findings, target enrichment and PCR have the advantage of only providing results about pathogen of interest. The ethical and privacy concerns associated with the presence of host genetic data in publically available metagenomic datasets have been well reviewed[81] and represent a separate challenge.

[H3] Regulatory challenges. Regulation, as well as helping to address some of the ethical concerns, will also be important in standardising WGS of viruses. The framework required to make viral WGS sufficiently robust and reproducible in clinical practice will come from a number of areas.

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

Lessons learned from the use of PCR in diagnostics may be useful here, beginning with ensuring good clinical laboratory and molecular practices[108,109]. 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[110], all of which will increase sequencing costs and are likely to deter adoption of pathogen genome sequencing by laboratories, that are sequencing only small batches of samples. Centralisation of virus WGS can help to ensure keeping adequate standards, 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[111,112]. For example, murine leukaemia virus[113,114] and parvovirus-like sequences[115,116] are just two of many contaminants that can come from common laboratory reagents such as nucleic acid extraction columns[117]. As with other highly sensitive technologies, robust laboratory practices and protocols are needed to minimize

contamination. It is also important to remember that detection of viral nucleic acid does not necessarily identify the cause of illness, and it is good practice when using NGS methods for diagnosis of viral infections to confirm the findings with alternative, independent methods which do not rely on testing for nucleic acids. For example, in cases of encephalitis of unknown origin, positive NGS findings can be confirmed by immunohistochemical analysis of the affected tissue[64,118], or identification of the virus by electron microscopy or tissue culture[81].

The standardisation 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 needed to make results reproducible, to make best practices easily shareable, and to allow accreditation of analysis software. However, best-practice analysis methods are continually evolving and prematurely standardising in an overly rigid manner may inhibit innovation. Commercialisation and regulation may help, as they provide financial and regulatory incentives to ensure that analysis tools and technologies meet the clinical needs. Finally, the development of well curated databases that show which variants are truly indicative of drug resistance will be critical for accurate clinical interpretation. Such databases have already been created for HIV[119], HBV[120,121] and HCV[122], but without recognition of their value by funding agencies, and corresponding centralised funding to ensure their continued maintenance and upkeep, these databases and associated tools may become swiftly outdated or unusable.

[H3] 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 sub-genomic 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 sub-genomic 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 sub-genomic 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**.** Additionally, whole genome information may provide important additional knowledge, as discussed above.

[H1] What does the future hold?

Current generation NGS technologies based around Illumina, 454, Ion Torrent or Sanger methodologies all generate short-read data, which presents challenges for haplotype phasing, i.e. determining whether genetic variants (whether inter or intra-host) occur on the same genetic background (single viral genome, clonal) or on related, highly-similar but different genetic backgrounds within 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, multi-drug resistance variants occur together on a single viral genome or distributed between a mixed population of viruses, each with different drug resistance profiles, is currently unclear.

Although there are computational tools[123] 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 with genomes under 20kb, 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[124]. So far, viral read lengths achieved by MinION sequencing have been relatively modest; examples of mean read lengths are 751bp for modified Vaccinia Ankara virus, 758bp for cowpox virus[125], 455bp (range 126–1477bp) for chikungunya virus, 358bp (range 220–672bp) for Ebola virus, 1576bp[126] (personal communication M.A.B: 6895bp) for HCMV and 572bp (range 318–792bp) for HCV[127]. Results from the better-established PacBio technology are more promising, including a recent report of a mean read length of 12,777bp for pseudorabies virus[128], which has a double-stranded DNA genome ~142kp in length. 9.2kb reads have been achieved with PacBio for HCV, although 9.2kb of the 9.6kb genome had been pre-amplified by PCR[129].

A drawback of both NGS and single-molecule sequencing is the need for high coverage to minimize the impact of sequencing errors. This is particularly problematic for drug resistance studies, as drug resistance most frequently results from single nucleotide mutations or small deletions (1-3 bases), especially in lower-fidelity RNA viruses[130]. Achieving the high coverage necessary to ensure accurate variant typing is challenging when there is a lot of host DNA compared to viral sequences, and when the error profile of a technology makes point mutations particularly hard to detect[124]. At the time of writing, MinION sequencing (R9 pore chemistry) has raw high quality (so called '2D reads') read error rates of ~5% (personal communication, Josh Quick), which compares unfavourably with error rates of other technologies (Illumina (<0.1%), Ion Torrent (~1%), but not PacBio (13% single pass)[131], although accuracy can been improved using circular consensus read sequencing[132,133].

However, combining these long-read technologies with target enrichment provides a potential way forward[126,134], as ambiguities can be resolved if sufficient depth of sequence is achieved for the target pathogen, and error rates for all methodologies may be reduced by further technological and analytical improvements. Depleting the host nucleic acids is an alternative

solution, as a higher proportion of virus reads would be recovered from each sequencing run. Whereas 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.

[H1] Conclusions

Viral WGS is of growing clinical importance for diagnosis, disease management, molecular epidemiology and infection control. There are a number of methods 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 characterised 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 with diverse but well characterised genomes. Two obvious areas of innovation currently exist: methods that can effectively deplete host DNA without affecting viral DNA and further development of long-read technologies to achieve the flexibility and competitive pricing of short-read technologies. New technologies are needed to unite the strengths of these different methods and allow healthcare providers to invest in a single technology that is suitable for all viral WGS applications.

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 Fraser, C. M. *et al.* The minimal gene complement of Mycoplasma genitalium. *Science* **270**, 397-403 (1995).

4 Hayden, E. C. Technology: The $1,000 genome. *Nature* **507**, 294-295, (2014).

5 Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449**, 804-810, (2007).

6 Grigoriev, I. V. *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res* **42**, D699-704, (2014).
7 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).
8 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).
9 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**, e01974-01914, (2014).
10 Zaraket, H. *et al.* Genetic makeup of amantadine-resistant and oseltamivir-resistant human influenza A/H1N1 viruses. *J Clin Microbiol* **48**, 1085-1092, (2010).
11 Houldcroft, C. J. *et al.* Detection of Low Frequency Multi-Drug Resistance and Novel Putative Maribavir Resistance in Immunocompromised Pediatric Patients with Cytomegalovirus. *Frontiers in Microbiology* **7**, doi:10.3389/fmicb.2016.01317 (2016).
12 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).
13 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).
14 Smith, G. J. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122-1125, (2009).
15 Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369-1372, (2014).
16 Koser, C. U. *et al.* Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog* **8**, e1002824, (2012).
17 Cartwright, E. J., Koser, C. U. & Peacock, S. J. Microbial sequences benefit health now. *Nature* **471**, 578, (2011).
18 Paredes, R. & Clotet, B. Clinical management of HIV-1 resistance. *Antivir Res* **85**, 245-265, (2010).
19 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).
20 Durant, J. *et al.* Drug-resistance genotyping in HIV-1 therapy: the VIRADAPT randomised controlled trial. *Lancet* **353**, 2195-2199 (1999).
21 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).
22 Khudyakov, Y. Molecular surveillance of hepatitis C. *Antivir Ther* **17**, 1465-1470, (2012).
23 Kim, J. H., Park, Y. K., Park, E. S. & Kim, K. H. Molecular diagnosis and treatment of drug-resistant hepatitis B virus. *World journal of gastroenterology* **20**, 5708-5720, (2014).
24 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).
25 Pawlotsky, J. M. Hepatitis C Virus Resistance to Direct-Acting Antiviral Drugs in Interferon-Free Regimens. *Gastroenterology* **151**, 70-86, (2016).

26 Thomson, E. *et al.* Comparison of next generation sequencing technologies for the comprehensive assessment of full-length hepatitis C viral genomes. *J Clin Microbiol*, doi:10.1128/JCM.00330-16 (2016).
27 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).
28 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).
29 Piret, J. & Boivin, G. Antiviral drug resistance in herpesviruses other than cytomegalovirus. *Rev Med Virol* **24**, 186-218, (2014).
30 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).
31 Lassalle, F. *et al.* Islands of linkage in an ocean of pervasive recombination reveals two-speed evolution of human cytomegalovirus genomes. *Virus Evolution* **2**, doi:10.1093/ve/vew017 (2016).
32 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).
33 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).
34 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 e591, (2014).
35 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).
36 Pierucci, P. *et al.* Novel autologous T-cell therapy for drug-resistant cytomegalovirus disease after lung transplantation. *J Heart Lung Transplant*, doi:10.1016/j.healun.2015.12.031 (2016).
37 Agoti, C. N. *et al.* Local evolutionary patterns of human respiratory syncytial virus derived from whole-genome sequencing. *J Virol* **89**, 3444-3454, (2015).
38 Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228-232, (2016).
39 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**, doi:10.1128/mBio.00444-16 (2016).
40 Faria, N. R. *et al.* Zika virus in the Americas: Early epidemiological and genetic findings. *Science* **352**, 345-349, (2016).
41 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).
42 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).
43 Power, R. A. *et al.* Genome-Wide Association Study of HIV Whole Genome Sequences Validated using Drug Resistance. *PLoS One* **11**, e0163746, (2016).

44 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).
45 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).
46 Zhou, B. *et al.* Composition and Interactions of Hepatitis B Virus Quasispecies Defined the Virological Response During Telbivudine Therapy. *Sci Rep* **5**, 17123, (2015).
47 Itakura, J. *et al.* Resistance-Associated NS5A Variants of Hepatitis C Virus Are Susceptible to Interferon-Based Therapy. *PLoS One* **10**, e0138060, (2015).
48 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**, doi:10.1128/mBio.02464-14 (2015).
49 Swenson, L. C., Daumer, M. & Paredes, R. Next-generation sequencing to assess HIV tropism. *Curr Opin HIV AIDS* **7**, 478-485, (2012).
50 Hutter, G. *et al.* Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *N Engl J Med* **360**, 692-698, (2009).
51 Kordelas, L. *et al.* Shift of HIV tropism in stem-cell transplantation with CCR5 Delta32 mutation. *N Engl J Med* **371**, 880-882, (2014).
52 Coaquette, A. *et al.* Mixed cytomegalovirus glycoprotein B genotypes in immunocompromised patients. *Clin Infect Dis* **39**, 155-161, (2004).
53 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).
54 Chou, S. *et al.* Improved detection of emerging drug-resistant mutant cytomegalovirus subpopulations by deep sequencing. *Antimicrob Agents Chemother* **58**, 4697-4702, (2014).
55 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).
56 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).
57 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).
58 Loman, N. J. & Pallen, M. J. Twenty years of bacterial genome sequencing. *Nature reviews. Microbiology* **13**, 787-794, (2015).
59 Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74, (2004).
60 Mulcahy-O'Grady, H. & Workentine, M. L. The Challenge and Potential of Metagenomics in the Clinic. *Frontiers in immunology* **7**, 29, (2016).
61 Hoffmann, B. *et al.* A Variegated Squirrel Bornavirus Associated with Fatal Human Encephalitis. *N Engl J Med* **373**, 154-162, (2015).
62 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).
63 Duncan, C. J. *et al.* Human IFNAR2 deficiency: Lessons for antiviral immunity. *Sci Transl Med* **7**, 307ra154, (2015).

64 Morfopoulou, S. *et al.* Human Coronavirus OC43 Associated with Fatal Encephalitis. *N Engl J Med* **375**, 497-498, (2016).
65 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).
66 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).
67 Wilson, M. R. *et al.* Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med* **370**, 2408-2417, (2014).
68 Depledge, D. P. *et al.* Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS One* **6**, e27805, (2011).
69 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*, doi:10.1111/petr.12610 (2015).
70 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).
71 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*, doi:10.1016/S1473-3099(16)00095-5 (2016).
72 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).
73 Kohl, C. *et al.* Protocol for metagenomic virus detection in clinical specimens. *Emerg Infect Dis* **21**, 48-57, (2015).
74 Sauvage, V. & Eloit, M. Viral metagenomics and blood safety. *Transfusion clinique et biologique : journal de la Societe francaise de transfusion sanguine* **23**, 28-38, (2016).
75 Lecuit, M. & Eloit, M. The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening. *Frontiers in cellular and infection microbiology* **4**, 25, (2014).
76 Oude Munnink, B. B. *et al.* Autologous antibody capture to enrich immunogenic viruses for viral discovery. *PLoS One* **8**, e78454, (2013).
77 Sabina, J. & Leamon, J. H. Bias in Whole Genome Amplification: Causes and Considerations. *Methods in molecular biology* **1347**, 15-41, (2015).
78 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).
79 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, doi:10.1007/s11262-015-1231-8 (2015).
80 Mlakar, J. *et al.* Zika Virus Associated with Microcephaly. *N Engl J Med* **374**, 951-958, (2016).
81 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).
82 Greninger, A. L. *et al.* A novel outbreak enterovirus D68 strain associated with acute flaccid myelitis cases in the USA (2012-14): a retrospective cohort study. *Lancet Infect Dis* **15**, 671-682, (2015).

83 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).
84 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).
85 Cotten, M. *et al.* Deep sequencing of norovirus genomes defines evolutionary patterns in an urban tropical setting. *J Virol* **88**, 11056-11069, (2014).
86 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).
87 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).
88 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).
89 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).
90 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).
91 Brown, J. R. *et al.* Norovirus whole genome sequencing by SureSelect target enrichment: a robust and sensitive method. *J Clin Microbiol*, doi:10.1128/JCM.01052-16 (2016).
92 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).
93 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).
94 Johnson, T. A. *et al.* Clusters of Antibiotic Resistance Genes Enriched Together Stay Together in Swine Agriculture. *MBio* **7**, e02214-02215, (2016).
95 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).
96 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).
97 Wylie, T. N., Wylie, K. M., Herter, B. N. & Storch, G. A. Enhanced virome sequencing using targeted sequence capture. *Genome Res*, doi:10.1101/gr.191049.115 (2015).
98 Tsangaras, K. *et al.* Hybridization capture using short PCR products enriches small genomes by capturing flanking sequences (CapFlank). *PLoS One* **9**, e109101, (2014).
99 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).
100 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).

101 Palser, A. L. *et al.* Genome diversity of Epstein-Barr virus from multiple tumor types and normal infection. *J Virol* **89**, 5222-5237, (2015).
102 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**, (2015).
103 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**, (2013).
104 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).
105 Briese, T. *et al.* Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. *MBio* **6**, e01491-01415, (2015).
106 Naccache, S. N. *et al.* Distinct Zika Virus Lineage in Salvador, Bahia, Brazil. *Emerg Infect Dis* **22**, 1788-1792, doi:10.3201/eid2210.160663 (2016).
107 Hofmann, B. Incidental findings of uncertain significance: To know or not to know--that is not the question. *BMC Med Ethics* **17**, 13, (2016).
108 England, P. H. Good Laboratory Practice when Performing Molecular Amplification Assays. Q 4 Issue 4.4. *UK Standards for Microbiology Investigations.* **Q 4** (2013).
109 Viana, R. V. & Wallis, C. L. *Good Clinical Laboratory Practice (GCLP) for Molecular Based Tests Used in Diagnostic Laboratories*. (INTECH Open Access Publisher, 2011).
110 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).
111 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).
112 Munro, A. C. & Houldcroft, C. Human cancers and mammalian retroviruses: should we worry about bovine leukemia virus? *Future Virology* **11**, 163-166, (2016).
113 Hue, S. *et al.* Disease-associated XMRV sequences are consistent with laboratory contamination. *Retrovirology* **7**, 111, (2010).
114 Erlwein, O. *et al.* DNA extraction columns contaminated with murine sequences. *PLoS One* **6**, e23484, (2011).
115 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).
116 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).
117 Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* **12**, 87, (2014).
118 Lipkin, W. I. A Vision for Investigating the Microbiology of Health and Disease. *J Infect Dis* **212 Suppl 1**, S26-30, (2015).
119 Shafer, R. W. Rationale and uses of a public HIV drug-resistance database. *J Infect Dis* **194 Suppl 1**, S51-58, (2006).
120 Gnaneshan, S., Ijaz, S., Moran, J., Ramsay, M. & Green, J. HepSEQ: International Public Health Repository for Hepatitis B. *Nucleic Acids Res* **35**, D367-370, (2007).
121 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).

122 Kuiken, C., Yusim, K., Boykin, L. & Richardson, R. The Los Alamos hepatitis C sequence database. *Bioinformatics* **21**, 379-384, (2005).
123 Hong, L. Z. *et al.* BAsE-Seq: a method for obtaining long viral haplotypes from short sequence reads. *Genome Biol* **15**, 517, (2014).
124 Schmidt, K. *et al.* Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J Antimicrob Chemother*, doi:10.1093/jac/dkw397 (2016).
125 Kilianski, A. *et al.* Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *Gigascience* **4**, 12, doi:10.1186/s13742-015-0051-z (2015).
126 Eckert, S. E., Chan, J. Z.-M., Houniet, D., Breuer, J. & Speight, G. Enrichment of long DNA fragments from mixed samples for Nanopore sequencing. *bioRxiv*, doi:10.1101/048850 (2016).
127 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).
128 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**, doi:10.1128/genomeA.00440-16 (2016).
129 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).
130 Kimberlin, D. W. & Whitley, R. J. Antiviral resistance: mechanisms, clinical significance, and future implications. *J Antimicrob Chemother* **37**, 403-421 (1996).
131 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).
132 Li, C. *et al.* INC-Seq: accurate single molecule reads using nanopore sequencing. *Gigascience* **5**, 34, doi:10.1186/s13742-016-0140-7 (2016).
133 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).
134 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).
135 Guan, H. *et al.* Detection of virus in CSF from the cases with meningoencephalitis by next-generation sequencing. *J Neurovirol* **22**, 240-245, (2016).
136 Morfopoulou, S. *et al.* Deep sequencing reveals persistence of cell-associated mumps vaccine virus in chronic encephalitis. *Acta Neuropathol*, doi:10.1007/s00401-016-1629-y (2016).
137 Brown, J. R. *et al.* Astrovirus VA1/HMO-C: an increasingly recognized neurotropic pathogen in immunocompromised patients. *Clin Infect Dis* **60**, 881-888, (2015).
138 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**, e53-57, (2015).
139 Barjas-Castro, M. L. *et al.* Probable transfusion-transmitted Zika virus in Brazil. *Transfusion*, doi:10.1111/trf.13681 (2016).
140 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** (2014).

141 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**, doi:10.1128/genomeA.00359-16 (2016).
142 Faria, N. R. *et al.* Mobile real-time surveillance of Zika virus in Brazil. *Genome Med* **8**, 97, doi:10.1186/s13073-016-0356-2 (2016).
143 Chitsaz, H. *et al.* Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nature biotechnology* **29**, 915-921, (2011).
144 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 would like to thank Julianne Brown and Kimberly Gilmour (GOSH) and Ronan Doyle (UCL) for their helpful discussions, and Josh Quick (University of Birmingham) for sharing unpublished MinION statistics. CJH was funded by Action Medical Research grant GN2424. MAB was funded through the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 304875 held by JB. This work was supported by the National Institute for Health Research Biomedical Research Centre at Great Ormond Street Hospital for Children NHS Foundation Trust and University College London. JB receives funding from the UCLH/UCL National Institute for Health Research Biomedical Research Centre. The authors acknowledge infrastructure support for the UCL Pathogen Genomics Unit from the UCL 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 the work for publication.

Conflicts of interest

The authors declare no conflicts of interest.

Box 1. RNA-seq and metagenomic diagnostics.

In cases of encephalitis of unknown origin, metagenomic techniques are promising diagnostic tools. There are a variety of protocols in use, but the main methods used are RNA-seq and metagenomics. For RNA-seq, the total RNA or a subset of RNA is extracted from a sample (cerebrospinal fluid or brain biopsy, for example), converted to cDNA and sequenced. Metagenomics generally describes the same procedure for DNA, but may also include simultaneous sequencing of DNA and RNA by incorporating a cDNA synthesis step. RNA-seq may improve 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[62].

Both methods have successfully identified new or known viral pathogens in encephalitis of unknown origin. Metagenomics has been used to aid the diagnosis and characterisation of enterovirus D68 in cases of acute flaccid paralysis[82]. Metagenomics identified herpesviruses in the CSF of four patients with suspected viral meningoencephalitis[135]. RNA-seq also identified HSV1 in an encephalitis case, although the use of a DNAse I digestion (intended to lower the amount of host nucleic acid) reduced the number of HSV1 reads[62]. Mumps vaccine virus has also been detected a chronic encephalitis case using RNAseq[136] .

RNA-seq has been very successful in identifying encephalitis caused by astroviruses[137,138] and coronaviruses[64]. The deaths of three squirrel breeders from encephalitis was linked to a novel squirrel bornavirus, which was identified by separate metagenomic sequencing of DNA and RNA[61]. 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.

Box 2: WGS of Zika virus

Whole-genome sequencing (WGS) of Zika virus can help to understand the epidemiology of the outbreak, 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 cases or interrupting 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 give molecular epidemiology studies sufficient power[40]. WGS, phylogenetic analysis and molecular clock dating, combined with other epidemiological data, were useful to study the introduction of Zika virus to South America[40]. For example, the most recent common ancestor of strains circulating in Brazil predates the 2014 football World Cup, making it highly unlikely that this event was responsible for introducing Asian-lineage Zika virus to South America[40].

WGS is also central for understanding Zika virus pathogenesis, for example, by trying to identify sequence changes that are associated with microcephaly, as it's currently unclear which genome regions determine pathogenesis. It is likely that numerous whole genome sequences of Zika virus from around the world and from microcephaly and asymptomatic cases are needed to link particular mutations to birth defects. So far, no changes in the Zika virus genome have been unambiguously associated with microcephaly[40,71,80].

WGS and fragment sequencing were used to identify a case of Zika virus transmission through platelet transfusion[139]. 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[140]. Based on this new case, blood products may need to be screened routinely for Zika virus[139].

Finally, WGS of Zika virus isolates has identified sequence polymorphisms in primer-binding sites[141], which may make PCR-based diagnosis and virus load quantification more difficult.

This highlights the need to characterise population-level diversity, especially in epidemics, in which the locally circulating virus may have diverged from viruses from other locations or time periods. A number of projects are underway to determine population-level diversity, including the Zika in Brazil real time analysis (ZIBRA) mobile laboratory project[142], which uses portable metagenomic sequencing of Zika virus and real-time reporting of results[106].

Figure 1: Methods for sequencing viral genomes from clinical specimens.

All specimens originally comprise a mix of host (in blue) and pathogen (in red) sequences. Direct metagenomic sequencing provides an accurate representation of the sequences within the sample although at high sequencing and data analysis/storage costs. PCR amplicon sequencing uses many discrete PCR reactions to enrich the viral genome, which increases the workload for large genomes substantially, but reduces the costs. Target enrichment sequencing uses virus-specific nucleotide probes 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 to PCR.

Table 1. Advantages and disadvantages of different viral sequencing methods

Method	Advantages	Disadvantages
Metagenomics	• Simple, cost-effective sample preparation • Can sequence novel or poorly characterised genomes • Effective in 'fishing' approaches to identify a potential underlying pathogen	• High sequencing cost to obtain sufficient data • Relatively low sensitivity to target pathogen • Coverage proportional to viral load • High proportion of non-pathogen reads increases

	• Low number of PCR cycles causes few amplification mutations • Preservation of minor variant frequencies reflects *in vivo* variation • No primer or probe design needed, which enables rapid response to novel pathogens or sequence variants	computational challenges • Incidental sequencing of human and off-target pathogens raises ethical and diagnostic issues
PCR amplification	• Tried and trusted – well-established methods and trained staff • Highly specific – most sequencing reads will be pathogen-specific, reducing 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 subject to primer mismatch, particularly in poorly characterised 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	• Single tube sample preparation suited to high throughput automation and sequencing of large genomes • Higher specificity than metagenomics	• High cost and technical expertise for sample preparation • Unable to sequence novel pathogens and requires well-characterised reference

	reduces sequencing costs • Overlapping probes increases tolerance for individual primer mismatches • Fewer PCR cycles (than PCR amplification) limits introduction of amplification mutations • Preservation of minor variant frequencies reflects *in vivo* variation	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 limits rapid response to emerging and novel viruses

Table 2. Limitations of viral sequencing compared to bacterial sequencing

Feature	Bacteria	Viruses	Challenges
Genome	dsDNA	dsDNA, ssDNA, partially dsDNA, ssRNA, 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) allowing 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 virus is heavily contaminated with host cell nucleic acids, reducing viral sequencing output
Clinical specimens	Hardy bacterial cells with cell walls can often be separated from human cells in	Viruses are intracellular pathogens, and although separation	Clinical specimens are heavily contaminated with host nucleic acids, reducing viral sequencing output

	clinical specimens using differential lysis methods or flow cytometry[143] prior to extraction	from host is possible (e.g. filtration, antibody pulldown), virus cannot easily be separated from clinical samples prior to extraction	
Methylation patterns	Bacteria use different methylation patterns from eukaryotes; host DNA can be depleted post-extraction using restriction endonucleases directed against CpG methylation[144]	DNA viruses are often methylated by the host intracellular machinery, and may possess similar methylation patterns	DNA digestion according to methylation patterns is less effective as a means of host-depletion for viral sequencing

Author biographies

Charlotte Houldcroft

Charlotte Houldcroft is an unestablished lecturer in the Division of Biological Anthropology, University of Cambridge, UK. She received her PhD in molecular biology in 2014 from the Wellcome Trust Sanger Institute, UK. Her first postdoctoral position was at University College London's Institute of Child Health, researching the role of genomic variation in pathology caused by the three biggest infectious killers of paediatric transplant recipients: Epstein-Barr virus, cytomegalovirus and adenovirus. She also has an interest in which infectious diseases afflicted the Neanderthals. Charlotte Houldcroft's homepage: http://www.bioanth.cam.ac.uk/directory/ch504

Mathew Beale

Mat Beale is a Postdoctoral Fellow at Wellcome Trust Sanger Institute (Cambridge, UK). He received his PhD in Virology from University College London (UCL) in 2013 for work conducted at Public Health England, followed by postdoctoral roles in pathogen genomics at Imperial College, St George's University of London (joint role) and UCL. He has expertise in genomic laboratory methods and bioinformatic analysis of pathogen genomes, and works with viral, bacterial and fungal human pathogens. His current research focuses on the use of genome sequencing and population genomics to address questions of inter- and intra-host evolutionary dynamics and population structure, host adaptation and acquisition of virulence. More details available at orcid.org/0000-0002-4740-3187

Judith Breuer

Judith Breuer is Consultant Clinical Virologist at Great Ormond Street Hospital for Children, London, Professor of Virology and Director of the Division of Infection and Immunity at University College London. She qualified in Medicine from the University of London, undertaking postgraduate medical training in London followed by a period in research at the Medical Research Council's National Institute for Medical Research. Her research centres on the application of pathogen genome sequencing to understanding the pathogenesis of viral infections, for which purpose she has pioneered the use of novel methods to sequence pathogen genomes directly from clinical material. She is currently using whole genome sequencing of viruses and hard-to-culture bacteria to improve the detection of drug resistant mutations and to understand the epidemiology of virus transmission for patient management. She leads an international Wellcome Trust funded consortium to use whole genome sequencing and other approaches to investigate the mechanisms underlying pandemic spread of norovirus infections.

Subject terms

Biological sciences / Biological techniques / Microbiology techniques

[URI /631/1647/2234]

Biological sciences / Biotechnology / Sequencing / Next-generation sequencing

[URI /631/61/514/2254]

Biological sciences / Microbiology / Microbial communities / Clinical microbiology

[URI /631/326/2565/107]

Biological sciences / Microbiology / Virology / Metagenomics

[URI /631/326/596/2142]

Health sciences / Health care / Diagnosis

[URI /692/700/139]

ToC blurb

Sequencing viral DNA and RNA is an important part of clinical practice, although so far mostly subgenomic fragements are sequenced. In this Opinion article, Houldcroft, Beale and Breuer, highlight the potential that sequencing whole viral genomes has for clinical applications.