View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Apollo

Clinical and biological insights from viral genome sequencing

Authors: Charlotte J. Houldcroft[1§], Mathew A. Beale[2] & Judith Breuer[2,3*]

*corresponding author

Affiliations:

1. Infection, Immunity and Inflammation, Great Ormond Street Institute of Child Health, University College London, UK
2. Division of Infection and Immunity, University College London, UK
3. Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK.

§ Present address: Division of Biological Anthropology, University of Cambridge, UK

Introduction

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]), combined with the rapidly falling cost of high-throughput sequencing[4], genomics has become a major contributor to our understanding of human and pathogen biology. Multiple large scale systematic pathogen genome projects have been recently completed or are on-going (e.g. sequencing thousands of microbiomes and fungal genomes[5,6]); these projects are shaping our knowledge of the genetic variation present in human and pathogen populations, the nature of genetic changes that underlie disease, and the sheer diversity of microorganisms with which we share our environments.

The methods and data from whole genome sequencing are increasingly being applied to clinical medicine, both from a human[7] and pathogen perspective. For example whole-pathogen genome sequencing has been used to identify new routes of *Mycobacterium abscessus*[8] nosocomial transmission and to understand *Neisseria meningitidis* epidemics in Africa[9], while partial genome sequencing has been used to detect drug resistance in RNA viruses such as influenza[10] and DNA viruses such as human cytomegalovirus (HCMV)[11]. Viral genome sequencing has

gained considerable traction, often focused on research or epidemiology. Whole pathogen genome sequencing has the advantage of detecting all known drug resistance mutations in a single test while deep sequencing can identify low level drug resistance mutations early enough for clinical intervention[12,13]. Whole genomes also provide good data with which to identify linked infections for public health and infection control purposes[14,15]. Notwithstanding, progress in whole-genome sequencing (WGS) of viruses for clinical practice has been slow. In contrast whole-genome sequencing of bacteria is now well accepted particularly for outbreak tracking and for the management of nosocomial transmission of antimicrobial resistant bacteria[16,17].

This review will address the challenges and opportunities for making WGS, using modern next generation sequencing (NGS) methods, a standard part of clinical virology practice. We will discuss the strengths, weaknesses and technical challenges inherent to different viral WGS laboratory methods (Table 1). The importance of deeply sequencing certain viral pathogens will be addressed. 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 which clinical viral WGS presents.

Why sequence viruses in clinical practice?

For small viruses such as HIV, influenza, HBV and HCV, partial genome sequencing has been widely used for research purposes, but also has important clinical applications. For example, the management of highly active anti-retroviral therapy (HAART) for HIV relies heavily on viral sequencing for detection of mutations conferring drug resistance. HAART has dramatically improved survival of patients with HIV, but successful therapy requires long-term

suppression of viral replication with anti-retroviral drugs, which may be prevented by impaired host immunity, sub-optimal drug penetration to host tissue compartments and incomplete patient adherence to therapy[18]. Where viral replication continues to occur, the high mutation rate of HIV enables resistance variants to emerge. It has become standard practice in many parts of the world to sequence the HIV *pol* gene, which encodes the main viral enzymes, for mutations conferring resistance to inhibitors of reverse transcriptase, integrase and protease[19], particularly when patients are first diagnosed and when viral loads indicate treatment failure. Sequencing resistance mutations has allowed more targeted alterations in treatment with significantly 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 mutations to guide HIV treatment improves disease outcomes. Similar approaches have been taken for identifying HCV[22], HBV[23], and influenza[24] resistance mutations.

Why sequence whole genomes?

Limited sequencing of the small number of genes that are targeted by anti-viral agents, such as the HIV polymerase gene, has hitherto been the norm in clinical practice. For detecting a limited number of antiviral resistance mutations, WGS has been too costly and labour-intensive to justify. However, the increase in numbers of antivirals targeting genes that are located across the genome, coupled 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) encoded by more than 50% of the viral genome[25]. Separate targeted sequencing for each of these can be as expensive and time consuming as WGS[26]. Partial genome sequencing is particularly problematic for larger viral genomes most notably those of the herpesviruses HCMV[11], VZV[27], HSV-1[28] and -2[29]. These have traditionally been treated with drugs targeting the

protein/thymidine kinase and DNA polymerase genes. However the growing numbers of drugs in development that interact with different proteins encoded by viral genes scattered across the genome, means that the targeted sequencing of multiple genes required for resistance testing is costly and less tractable[30]. Sequencing the whole genome captures all resistance mutations simultaneously and obviates the need to design and optimise new PCR assays for detecting resistance to new drugs. A good example of this is HCMV, where WGS can simultaneously capture the genes with products targeted by the licensed therapies such as UL27 (unknown function), UL54 (DNA polymerase), and UL97 (protein kinase), as well as newer drugs such as letermovir which targets UL56 (terminase complex), enabling comprehensive anti-viral resistance testing in a single test[11]. At the same time WGS has the potential to provide information on epitopes, evolution of sequences within a patient over time[11], and evidence of recombination between HCMV strains[31]. WGS can highlight putative novel drug resistance mutations, or predicted changes to epitopes, although phenotypic testing of any findings in a model system is required to confirm clinical resistance (e.g. [32]) or to map epitope changes (e.g. [33]).

As pre-existing resistance to anti-viral drugs (for example, protease inhibitor-resistant HCV[34] and nucleoside analog reverse transcriptase inhibitor-resistant HBV[35]) increases, whole genome sequences will provide the comprehensive resistance data required for selecting appropriate treatment to achieve good patient outcomes. A complete knowledge of all resistance mutations can also support more radical management decisions. In a recent case report, identification of extensive genome-wide HCMV drug resistance within a patient supported the clinical decision to change to immunotherapeutic treatments, specifically autologous cytomegalovirus-specific T cells[36].

Whole genomes may also better identify transmission events and outbreaks, which is not always possible with sub-genomic fragments. For example, sequencing respiratory syncytial

virus (RSV) genomes demonstrated that variation was present outside the gene traditionally used for genotyping, and could be used to help track outbreaks within households, where there had been insufficient time for single genes to accumulate enough genetic variability to be used for transmission studies[37]. The increased number of phylogenetically informative variant sites obtained from generating full or near full length genomes has been shown to obviate the need for high quality sequences, allowing robust linking of Ebola cases and public health interventions in real time during the recent epidemic[38]. This also applies to Zika virus, and Box 2 explores the role of WGS in public health efforts to control the outbreak in South America.

The increased use of whole pathogen sequencing routinely for diagnostic purposes[39] is likely to have wider clinical and research benefits. For example HIV genome sequencing to identify resistance mutations, can also be used to explore questions related to viral evolution[40], public health[41] and viral genetic association with disease. This includes well-powered genotype-phenotype association studies or 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].

Why do we need deep sequencing?

Modern methods which make use of massively parallel sequencing provide better opportunities to examine pathogen diversity through analysis of viral populations within or between hosts that contain nucleotide variants or haplotypes at low (sub-consensus, less than 50%) frequencies. Minority variant analysis is particularly powerful for RNA or retro-transcribing viruses, because they typically have high within-host nucleotide diversity. HIV is the classic example; the viral replication cycle utilises an error-prone reverse transcriptase enzyme that introduces mutations at an extremely high rate ($4.1 \pm 1.7 \times 10^{-3}$ per base per cell)[43]. This results in a given patient containing not one, but many closely related viruses each bearing subtly

different variants, sometimes described as a quasispecies or 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 alter the clinical phenotype of the virus, or 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 baseline patient sample can rapidly rise to become a majority (consensus) variant under the selective pressure of drug treatment[44]. Investigators have observed similar changes in frequency of resistance-associated alleles during treatment of viruses such as HBV[45], HCV[46], HCMV[11] and influenza[47].

Sensitive deep-sequencing of viruses is not only required to detect drug resistance: for HIV, it is also key in genotypic prediction of receptor tropism, which has clinical implications in treatment of HIV. HIV can be grouped genotypically by its cellular co-receptor usage as R5 (CCR5-using), X4 (CXCR4-using) or R5X4 (dual tropism). Maraviroc is a CCR5 receptor antagonist, blocking infection by R5-tropic HIV genotypes, but contraindicated in HIV+ individuals who have X4 or R5X4 HIV genotypes. Just a 2% frequency of X4 or R5X4 genotypes is predictive of maraviroc treatment failure[48]. Sub-consensus frequencies of X4 or R5X4 HIV are also important to the success[49] or failure[50] of bone marrow transplants from CCR5-deleted (CCR5-Δ32) donors. This may influence decisions to continue or stop anti-viral therapy in these patients[49].

Detection of minority variants and haplotype identification may also detect mixed infections. In HCMV mixed-genotype infections or super-infections[51] are associated with poor clinical outcomes - detection of these by WGS might support a decision to treat disease in these patients more aggressively.

Establishing the clinical associations of minority variants is clearly important, as with maraviroc treatment failure; Sanger sequencing of a virus population can detect minority

variants down to frequencies of between 10 and 40% (e.g. [52]), whilst NGS can sequence those same PCR amplicons to a much greater depth[53], and consequently capture more of the variability present. Thresholds of sensitivity and specificity established need to be specific to the virus in question, and reflect the potential biases of the sequencing methods used. Many studies of HIV drug resistance utilising deep-sequencing of PCR amplicons require minority variants to be present at >1%, to reduce the possibility of false positives[54,55]. This may lead to a failure to detect true drug resistance mutations at frequencies of 0.1%-1%, which may ultimately be associated with poor treatment outcome on some drug regimes[55]. While a 1-2% frequency threshold (or lower) may be clinically relevant to drug resistance in HIV, it is less clear whether the same degree of sensitivity would be required for monitoring vaccine escape in HBV or drug resistance in herpesvirues (discussed below). Large cohorts of patients will need to be followed with samples collected before, during and after treatment[44,48], to establish clinical significance thresholds for minority drug resistance[11] and vaccine escape variants for each virus.

Direct deep sequencing of clinical material, either by shotgun or RNAseq methods (so called metagenomic methods) also provides the opportunity for unbiased detection of pathogen sequences and thus primary diagnosis of viral and other infections, thereby providing an alternative to culture, electron microscopy and qPCR. This is discussed further below.

Practical considerations for sequencing virus genomes

As previously alluded to, sequencing viral nucleic acid whether cultured or directly from clinical specimens, is complicated by the presence of contaminating host DNA[56]. This makes it different from bacterial sequencing which is easily carried out using clinical isolates and thus sample preparation is relatively straightforward (Table 2). Currently, genome sequencing of viruses can be achieved by ultradeep sequencing or by enriching for viral nucleic acid prior to

sequencing either directly or through prior concentration of viral particles. All approaches have their own costs and complexities.

The three primary methods currently used for viral genome sequencing are summarised in Figure 1.

(i) Metagenomics - ultra deep sequencing

Metagenomic approaches have been extensively used for pathogen discovery and for characterising microbial and general pathogen diversity in environmental and clinical samples[57,58]. Total DNA and/or RNA from a sample, including from host, bacteria, viruses, fungi and other pathogens present are extracted, put through library preparation and sequenced by 'shotgun' or RNA-seq methods (see Box 1). These approaches have proven to be very powerful for detecting viral[59,60,61] and other causes[62] of encephalitis where other conventional methods such as PCR have failed. Box 1 explores the growing diagnostic applications for metagenomics and RNAseq, for example in encephalitis of unknown aetiology (e.g.[63-65]). In addition, a number of whole viral genomes have been sequenced in this manner, including Epstein-Barr virus (EBV)[66] and HCV[26]. However, these methods may be insensitive, because of the presence of contaminating host and commensal pathogen nucleic acid[56] (Table 2) in clinical specimens. For example on-target read yields (the proportion of reads matching the target genome) from metagenomic WGS of 0.008% (EBV genome from the blood of a healthy adult[67]), 0.0003% (lassa virus genomes from clinical samples[68]) and 0.3% (a filtration and centrifugation enriched Zika virus sample[69]) have been reported. The read depths obtained are often inadequate for robust resistance calling[26] and the cost is high. Thus the method has typically only been performed on a small number of samples for research purposes (e.g.[69,70]). To improve read depths, concentration of viral particles prior to sequencing (as for example in

the Zika case[69]), depletion of host material or ultra-deep sequencing have been employed, all of which add to the cost. Concentrating viral particles from clinical specimens by antibody-mediated pulldown (e.g. VIDISCA), filtration, or ultracentrifugation, to isolate a fragment size profile, and depletion of free nucleic acid[71-74] have all been tried. These host nucleic acid depletion methods may result in there being insufficient viral nucleic acid for sequencing library preparations. To overcome this, non-specific amplification methods (e.g. multiple displacement amplification; MDA) which make use of random primers and phi 29 polymerases may be effective in increasing DNA load. However, these approaches are time consuming, costly, and may increase the risk of biases, error and contamination without necessarily improving the sensitivity of sequencing[75,76]. Moreover, there are often still a high proportion of host reads present in treated samples[77].

Where metagenomic methods are used for pathogen discovery or diagnosis, appropriate bioinformatic tools and databases capable of evaluating whether detected pathogens sequences are truly likely to be the cause of infection, innocent bystanders or contaminants are critical. Bioinformatic analyses of large metagenomic datasets places an increased burden on high performance computational resources.

The fact that metagenomics requires no prior knowledge of the viral genome, can be considered a strength[26] in that it allows novel viruses to be sequenced without the need for primer or probe design and synthesis. This is particularly apposite for rapid responses to emerging threats such as Zika[78]. Metagenomic viral genome sequencing may also 'piggy back' on projects to sequence virus-associated cancer genomes, which informs clinical care of the cancer or provides further information on cancer evolution, while generating high coverage of integrated virus genomes as part of the process[66]. However, the presence of incidental findings (human genome sequences with potential disease associations, pathogens which were not part of the question that prompted the initial sequencing) may also present ethical (and even diagnostic)

dilemmas for some applications of clinical metagenomics (discussed below and reviewed in [79]). A recent case in point was a cluster of acute flaccid myelitis cases associated with enterovirus D68[80]. The analysis of the metagenomic datasets derived from patients was the subject of discussion through formal[81] and informal scientific channels (http://omicsomics.blogspot.co.uk/2015/07/leaky-clinical-metagenomics-pipelines.html), with different groups disagreeing over the interpretation of the same data, especially as some of the alternative pathogens detected can cause treatable bacterial disease. Regulation and reporting frameworks will be important to resolve future issues of this kind.

(ii) PCR amplicon enrichment,

An alternative to metagenomic approaches is to enrich for the specific viral genome prior to sequencing. PCR amplification of hundreds to thousands of base pairs of viral genetic material using primers that are complementary to a known nucleotide sequence has been the most common approach to enriching for small viral genomes such as HIV and influenza, prior to NGS sequencing for diagnostic and public health purposes. Recent examples of this approach being applied for public health include sequencing measles virus by PCR-WGS to provide maximum phylogenetic resolution of an outbreak at the 2010 Winter Olympics[82], sequencing of Ebola virus genome to study epidemic dynamics[38], and Zika virus genome sequencing (explored in Box 2). PCR whole-genome sequencing of norovirus (7.5kb genome) has been used to understand norovirus transmission in community[83] and hospital[84] settings. For example, this research showed that some cases within a hospital with plausible epidemiological linkage were in fact independent introductions of the pathogen; but that other cases were the result of transmission, despite infection control practices being in place[84]. Other PCR-based deep sequencing studies have generated multiple whole genomes for influenza[85] (~13.5kb), dengue[86] (~11kb), and HCV[87] (9.6kb). This method is feasible (as with PCR and Sanger sequencing)

because these viruses all have relatively small genomes, requiring only a small number of PCR amplicons to assemble whole genome sequences. RNA virus heterogeneity may however necessitate the use of multiple overlapping primer sets to ensure comprehensive amplification of all genotypes, for example HCV[26], norovirus[83], rabies[88] and RSV[37]. PCR amplicon sequencing is also more successful for WGS of samples with low virus loads than metagenomic methods[26], although other methods such as target enrichment of viral sequences may work equally well in low copy number samples (e.g. low copy norovirus samples[89]).

Overlapping PCRs combined with NGS have been used to sequence the whole genomes of larger viruses such as HCMV[90], but this overlapping amplicon method has limited scalability, since many primers are needed[90] and a greater amount of starting DNA to allow for each additional PCR, which may not be available from clinical samples. This limits the number of suitable samples available and also the genomes which can be studied with this method. A molecular epidemiology study of the relatively small Ebola genome required between 8 and 19 PCR products to amplify the genome for MinION nanopore sequencing[38], whilst 14[83] and 22 pairs[84] of primers were needed to amplify and Illumina sequence norovirus genomes. This becomes less practical in a clinical rather than research setting 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 in very variable genomes and the increasing labour and consumables cost associated with multiple PCR reactions[91]. 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 sequencing of sequencing viral genomes beyond 20 - 50 Kb impractical with current technologies, particularly with regards to large multi-sample studies or routine diagnostics. Another consideration is that increasing PCR reactions require a corresponding

increase in available sample, and this is not always possible where clinical specimens are limited. Improvements in microfluidic technologies may help to overcome some of these barriers to PCR-based methods, for example Fluidigm, RainDance and other 'droplet' sequencing technologies. Microfluidics-based PCR and pooling of multiple amplicons have been used successfully to sequence multiple anti-microbial resistance loci, for example[92], and can also applied to viral genomes, potentially down to the single-cell sequencing level.

PCR may encounter problems in amplifying highly variable pathogens such as HCV[93] and norovirus where there are many different genotypes, with some genotypes encountering primer amplification issues[26,89], or where there is insufficient characterisation of intra-genotypic diversity, leading to primer mis-matches[83]. Careful design of degenerate primers may help to mitigate these problems, but novel variants still present a risk to detection and amplification.

(iii) <u>Target Enrichment methods</u>

Target enrichment (TE) methodologies (also known as pulldown, capture or specific enrichment methods) represent one solution to problems of PCR or metagenomic sequencing of virus genomes. We and a number of other groups have been developing methods that can be used to sequence whole viral genomes directly from clinical samples without the need for prior culture or PCR[94-96]. These methods typically involve small RNA/DNA probes designed to be complementary to the pathogen reference sequence (or panel of references). Unlike in specific PCR amplicon based methods, the entire genome can be covered by a single tube of overlapping probes which are used in a hybridisation reaction to capture or 'pull down' complementary DNA sequences bound to a solid phase (e.g. streptavidin-labelled magnetic beads) from the total nucleic acids present in a sample, followed by sequencer-specific (e.g. Illumina) adaptor ligation and a small number of PCR cycles to enrich for successfully ligated fragments. This has been used successfully to characterise large and small clinically relevant viruses such as HCV[26], HSV1[97], VZV[96], EBV[98], CMV[66], HHV6[99] and HHV7[100]. The reaction

is performed in a single well and, like microfluidics-based PCR reactions, is amenable to high throughput automation[98]. The lack of a culture step means that the sequences obtained are more representative of original virus than cultured viral isolates, with fewer mutations than observed in PCR amplified templates[66,96]. The success of this method is in part based on the number of available reference sequences for the virus of interest: specificity increases when baits are designed against a larger panel of reference sequences, leading to better capture of the breadth of within and between sample diversity. TE probe design allows for limited mismatching between template and probe, but whilst PCR requires only knowledge of flanking regions of a target region, TE requires knowledge of the internal sequence in order to design baits. This is balanced by the fact that TE is less vulnerable to a single amplicon failure due to mismatch as internal and overlapping regions may still be captured even if one probe fails[66,96]. As such TE is not suitable for characterisation of novel viruses with low homology to known viruses, where metagenomics (or in some cases, PCR using degenerate primers), may be more appropriate.

As with all methods, the technique is also subject to constraints with regard to starting viral load. We have shown that although capable of sequencing virus from viral loads as low as 2000 IU/ml (HCV) or 2500 IU/ml (HCMV), targets could only be enriched so much, leading to reduced depth of coverage in sequencing data at lower viral concentration[26,66]. With metagenomics, the proportion of sequencing data mapping to the pathogen genome (the on-target read percentage) that can be expected from unenriched sequencing of clinical samples is small. Depending upon the starting pathogen load in a sample, TE can enrich percentage on-target viral reads from 0.01% up to 80% or more[66]. This allows a higher degree of multiplexing than unenriched metagenomics, and brings an accompanying decrease in the price of sequencing, albeit with a relative increase in the cost of library preparation. There are alternative approaches to enriching viral reads which include pulse-field gel electrophoresis[101],

which separates large viral genomes from smaller host DNA fragments, allowing for sequencing libraries composed of a smaller proportion of contaminating host DNA.

Enrichment techniques which make use of degenerate RNA or DNA probes to hundreds of viral species to pull viral nucleic acid out of samples and sequence them, e.g.the VirCapSeq method, have also been developed[102]. This method is designed for detection of both known and novel viruses, although its performance remains to be evaluated.

Comparison of all three 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, yielding the lowest genome coverage for comparable sequencing effort and were more prone to yield incomplete genome assemblies. The PCR method was the least tractable and most labour intensive, requiring repeated amplification and was the most likely to miss mixed infections, but where reactions were successful, yielded the most consistent read depth, whereas metagenomics and TE yielded read depths in proportion to virus copy number. Some HCV genotypes (particularly genotype 2) were more prone to generate incomplete sequences when PCR was used instead of metagenomics or TE. Targeted enrichment was the most consistent method, achieving full genomes and identical consensus sequences. The ease of library preparation for metagenomic and TE sequencing of HCV was considered a major advantage for clinical sequencing, but PCR may still be appropriate for very low virus load samples.

Similar results were achieved in a study comparing norovirus sequencing from PCR amplicons and target enrichment[89]. TE generated 100% genome coverage in 164/164 samples, while PCR-

based capsid sequencing was only possible in 158/164 samples, with PCR failures attributable to low virus titres and PCR primer mismatches, suggesting TE is more sensitive than PCR for norovirus sequencing and better accommodates between-strain sequence heterogeneity[89]. TE has also been used as a fall-back method for samples with lower virus loads which do not give WGS after metagenomic sequencing[103]. Both metagenomic and TE methods have the advantage that they are applicable to all size pathogen genomes, whereas PCR based methods are less tractable for sequencing larger viral genomes or for non-viral (e.g. bacterial, fungal, parasite) pathogen genomes.

These direct comparisons of different methods[26,89] will be important in demonstrating the situations in which each method should be used, based on their sensitivity and specificity, as well as factors which are relevant to clinical diagnostic labs such as cost, scalability and turn-around time (summarised in Table 1).

Challenges of analysis and interpretation

Beyond the technical challenges of method choice for viral WGS, 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 proliferation of analysis options.

Ethical issues and incidental findings

In many clinical tests (e.g. MRI scans, host genome sequencing), there is a risk of detecting a disease association that was not part of the original investigation yet may have clinical significance for the individual or their family. These so called ‘incidental findings’ remain a

topic of intense medical ethical debate[104]. The risk of incidental findings in pathogen sequencing (e.g. discovery of HIV infection during metagenomic sequencing for other pathogens) is not unique and has been resolved in clinical virology laboratories, where multiplex PCRs are used and only one of the tests has been requested. In these cases it is the practice of the laboratory to suppress the result that has not been requested (personal communication, J Breuer). In UK laboratories, 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 (e.g. detection of variants that predispose to cancer risk) from a pathogen metagenomics study are not reported to the individual in the UK, because this reporting is only permissible with patient consent. In regard to both host and virus incidental findings, targeted enrichment and PCR have an advantage as they target only the 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 by Hall and colleagues[79] and represent a separate challenge.

Regulatory challenges

Regulation, as well as helping to address some of the concerns addressed above, 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 CLIA in the USA, or accreditation against medical laboratory quality and competence standardisation criteria for ISO 15189) will support the development of viral WGS standards if there is sufficient pressure from hospitals, journals and funding agencies.

Lessons learned from the use of PCR in diagnostics may be useful here, beginning with ensuring good clinical laboratory and molecular practices[105,106]. This will mean including

406 negative samples in every sequencing run, to assess contamination thresholds, spiking samples
407 with a known virus to provide a sensitivity threshold and including positive controls and
408 controls for batch-to-batch variation[1077], all of which will increase sequencing costs and are
409 likely to deter adoption of pathogen genome sequencing by laboratories sequencing small
410 batches of samples. The result may be to drive centralisation of virus WGS to ensure adequate
411 standards are kept, ensure large batches of samples and keep costs down.
412 The issues of sensitivity and contamination are especially important in WGS because of the
413 risk of both false-negative and false-positive detection of pathogens. Highly sensitive
414 sequencing (whether metagenomic, PCR or TE based) may detect low-level contaminating
415 viral nucleic acid (reviewed in[108,109]). For example murine leukaemia virus[110,111] and
416 parvovirus-like sequences[112,113] are just two of many contaminants that have been recognised
417 to come from common laboratory reagents such as nucleic acid extraction columns[114]. As with
418 other highly sensitive technologies, robust laboratory practices and protocols are needed to
419 minimize contamination. It is also important to remember that detection of viral nucleic acid
420 does not necessarily identify the cause of illness, and it is good practice when using NGS
421 methods for diagnosis of viral infections to confirm the findings with alternative, independent
422 methods which do not rely on nucleic acid testing. For example in cases of encephalitis of
423 unknown origin, positive NGS findings can be confirmed by immunohistochemical analysis of
424 the affected tissue[59,115], or identification of the virus by electron microscopy or tissue culture[79].
425 The standardisation of methods, including bioinformatics approaches will be key to the
426 successful use of NGS and WGS in clinical virology. Software packages that use a graphical
427 user interface (GUI) rather than requiring command-line expertise, with strict version control
428 of software and analysis pipelines to make results reproducible, best practices easily shareable,
429 and to allow accreditation of analysis software will be necessary, whilst retaining an
430 appreciation that best-practice analysis methods are continually evolving and prematurely

standardising in an overly rigid manner may inhibit innovation. Commercialisation and regulation may help, providing financial and regulatory incentives to ensure that analysis tools and technologies meet the needs of clinical sequencing for virology. Finally for drug resistance, the development of well curated databases of which mutations are truly indicative of drug resistance will be critical for accurate clinical interpretation. Such databases have already been created for HIV[116], HBV[117,118] and HCV[119], but without recognition of their value by funding agencies, and corresponding centralised funding to ensure their continued maintenance and upkeep, tools may become swiftly outdated or unusable.

Financial barriers to the use of viral WGS in a clinical setting

While there are good reasons for sequencing whole genomes, and the general use of NGS, if diagnostic or hospital-based laboratories are to be persuaded to make the transition away from sequencing sub-genomic fragments, they need to see not only that the additional information gained from WGS is really of benefit to patient care; but that WGS is (or will become) as scalable and automatable as sub-genomic fragment sequencing, that the regulatory framework is suitable and that the price of sequencing whole genomes is competitive with sequencing fragments.

Currently the costs of sequencing viral genomes, notwithstanding their small size, remain generally higher than sequencing of sub-genomic target resistance genes. Equally, whole genome information may provide important additional knowledge, as discussed above. 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.

What does the future hold? Long-read sequencing and host depletion

Current generation NGS technologies based around Illumina, 454, Ion Torrent or Sanger methodologies as described above have the ubiquitous problem of generating short-read data which presents challenges for haplotype phasing of intra-host minor variants, which aims to identify whether a set of genetic variants occur on the same genetic background (clonal population) or on related, highly-similar but different genetic backgrounds within the same population (sometimes called a viral swarm or cloud); as well as sequencing across repetitive, recombinatorial or mobile genetic regions which are more difficult to resolve using short reads due to problems such as mapping ambiguities. The clinical implications of understanding whether, for example, multi-drug resistance occurs on a clonal genetic background or in a mixed population of viruses with different drug resistance profiles is currently unclear.

While there are computational tools (e.g.[120]) to help resolve these issues, especially of interest to researchers, there are also new technologies available. Newer single molecule sequencers such as PacBio (Pacific Biosciences) and MinION (Oxford Nanopore) are capable of extremely long read sequencing, and in some cases whole viral sequences (for example viruses with genomes under 20kb, such as Ebola virus, norovirus and influenza A) could theoretically be obtained from single reads. The MinION also has the advantage of being very fast, taking in some cases as little as four hours to go from sample receipt to reporting of analysed data[121]. Data on viral read lengths achieved from MinION sequencing have been relatively modest (e.g. mean read lengths of: 751bp (Modified Vaccinia Ankara), 758bp (cowpox virus)[122], 455bp [range 126–1477] (chikungunya virus), 358bp [220–672] (Ebola virus), 1576bp[123] or 6895bp (HCMV)[personal communication, M Beale] and 572bp [range 318–792] (HCV)[124]). Results from the better-established PacBio technology are more promising, including a recent report of a pseudorabies virus genome sequenced with a mean read length of 12,777bp (against a double-stranded DNA genome ~142kp in length)[125], and 9.2kb reads have been achieved in

PacBio HCV genome sequencing (where only 9.2kb of the 9.6kb had been pre-amplified by PCR)[126].

A drawback of both NGS and single-molecule sequencing however is the need for high coverage to minimize the impact of sequencing errors, particularly in the context of 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[127]. This can be a challenge where the amount of viral genome is dwarfed by the presence of host DNA, and when the error profile of a technology makes point mutations particularly hard to detect accurately[121]. At the time of writing, MinION sequencing (R9 pore chemistry) has raw 2D read error rates of ~5% [personal communication, Josh Quick], which compares unfavourably with Illumina (<0.1%), Ion Torrent (~1%) and PacBio (13% single pass, <1% with circular consensus read) error rates[128].

However, demonstration of the potential for using these long read technologies with target enrichment provides a potential way forward[123,129], as ambiguities can be resolved if sufficient depth of sequence is achieved for the target pathogen, and errors rates for all methodologies may be reduced with technological and analytical improvements. Products or methods which can deplete the host genetic background but not the viral nucleic acids within a sample would be an alternative solution, meaning a higher proportion of virus reads would be recovered from each sequencing run. While there are already solutions in place for bacterial sequencing (e.g. human ribosome RNA or mitochondrial depletion, selective depletion of DNA with a certain methylation pattern), there are no dedicated products for viral sequencing.

Conclusion

Whole virus genome sequencing is of growing importance in a clinical context, for diagnosis, disease management and molecular epidemiology (including infection control). There are a

number of methods available to achieve WGS of viruses from clinical samples. Currently the choice of methodology (amplicon sequencing, target enrichment or metagenomics) 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 works well where viral genomes are short and diversity in primer binding sites is low, while 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: firstly for methods that can effectively deplete host DNA whilst preserving viral DNA, and secondly for further development in the long-read technology market in order to achieve the range of flexibility and competitive pricing that exists in the short-read market. New technologies are needed to unite the strengths of these different methods and allow healthcare providers to invest in a single technology which is suitable for all viral WGS applications.

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.

Funding

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

529 Centre. The authors acknowledge infrastructure support for the UCL Pathogen Genomics Unit
530 from the UCL MRC Centre for Molecular Medical Virology and the UCLH/UCL National
531 Institute for Health Research Biomedical Research Centre. The funders had no role in study
532 design, data collection and interpretation, or the decision to submit the work for publication.

533

534 **Box 1. RNA-seq and metagenomics diagnostics.**

BOX:RNASEQ AND METAGENOMICS DIAGNOSTICS

In cases of encephalitis of unknown origin, metagenomic techniques are becoming increasingly promising diagnostic tools. There are a variety of protocols in use, but the clearest distinction is between RNA-seq and metagenomics. RNA-seq is the sequencing of either the total RNA or a subset of RNA extracted from a sample (cerebrospinal fluid or brain biopsy, for example), converted to cDNA and sequenced. Metagenomics is generally used to describe the same procedure for DNA, but may also include simultaneous sequencing of DNA and RNA by incorporating a cDNA synthesis step. RNA-seq methodologies may improve detection of pathogenic viruses, as many viruses have RNA genomes; the expression of viral genes in the CSF or brain is indicative of 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[64].

Both methods have successfully identified new or known viral pathogens implicated in encephalitis of unknown origin. Metagenomics has been used to aid in diagnosis and characterisation of enterovirus D68 in cases of acute flaccid paralysis[80]. Metagenomics identified herpesviruses in the CSF of four patients with suspected viral meningoencephalitis[130]. RNA-seq also successfully identified HSV1 in an encephalitis case, although the use of a DNAse I digestion (intended to lower the amount of host nucleic acid

in the subsequent sequencing library) lowered the number of HSV1 reads[64]. Mumps vaccine virus has also been detected a chronic encephalitis case using RNAseq [Morfopoulou, S. Deep sequencing reveals persistence of cell-associated mumps vaccine virus in chronic encephalitis. Acta Neuropathologica (In Press.)].

RNA-seq has been very successful in identifying encephalitis caused by astroviruses[131,132] and coronaviruses[59]. The deaths of three squirrel breeders from encephalitis was linked to a novel squirrel bornavirus through the use of a metagenomic protocol in which DNA and RNA were separately extracted and sequenced as discrete libraries, providing complementary data[63]. Ultimately, metagenomics provides more information about the virus genome present in a sample than PCR alone, which may be important for molecular epidemiology, while RNA-seq has the power to selectively capture information on which sequences are present, as well as informing researchers about viral gene expression of relevance to pathology.

Box 2: The role of whole-genome sequencing in Zika virus epidemiology and infection control

Box 2: The role of whole-genome sequencing in Zika virus epidemiology and infection control

Zika whole-genome sequencing is being used to understand the epidemiology of the outbreak (where did the virus come from? When did it enter Brazil?); to understand the connection between the virus and microcephaly; and to inform control measures, by stopping importation or interrupting transmission from a reservoir, and informing blood safety measures in hospitals, for example by demonstrating transfusion transmissibility of the virus.

Whole-genome (or near whole-genome) sequence is required from flavivirus genomes to give molecular epidemiology studies sufficient power[41]. WGS, phylogenetic analysis and molecular clock dating, combined with other epidemiological data, were useful in excluding

hypotheses about the introduction of Zika virus to South America[41]. 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[41].

WGS is also central to understanding Zika virus pathogenesis, and could be used to interrogate the whole genome of Zika virus for changes associated with microcephaly, as not enough of the virus's biology is currently understood to allow studies to limit themselves to smaller regions of the genome. It's likely that a wide sample of Zika whole genome sequences, from around the world and from microcephaly and asymptomatic cases, will be needed to be give confidence to any studies linking particular mutations to the birth defects seen in the recent Zika virus outbreak. No changes in the Zika virus genome have yet been unambiguously associated with microcephaly[41,69,78].

Whole-genome and fragment sequencing were used to identify a case of probably transfusion transmission of Zika virus through a platelet donation. This has significant public health and infection control relevance as it suggests asymptomatic donors are capable of transmitting the virus to immunocompromised individuals, although PCR-based testing had already established the presence of Zika virus in the blood supply in a previous outbreak, in this case without molecular epidemiology to demonstrate cases of Zika virus in blood product recipients[133]. Blood products may need to be routinely screened for Zika virus[134].

Finally, whole-genome sequencing of Zika isolates has found sequence polymorphisms within primer-binding sites[135], which may make PCR-based diagnosis and virus load quantification more difficult. This highlights the need to characterise population-level diversity, especially in epidemics, where the locally circulating virus sequence may have diverged significantly from related sequences from other locations or time periods. A number of projects are underway to achieve these goals, including the ZIBRA mobile laboratory

project[136], employing portable metagenomic sequencing of Zika virus (http://zibraproject.github.io/) and real-time reporting of results[103].

Figure 1: Major 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 at the cost of high sequencing and data analysis/storage costs. PCR amplicon sequencing uses many discrete PCR reactions to enrich the viral genome, significantly increasing the workload for large genomes, but reducing the sequencing costs. Target enrichment sequencing uses virus-specific nucleotide probes bound to a solid phase to enrich the viral genome in a single reaction, reducing workload, but increasing library cost (relative to PCR).

Table 1. Advantages and disadvantages of different viral sequencing sample preparation approaches

	Advantages	Disadvantages
Metagenomics	• Simple, cost-effective sample preparation • Can sequence novel/poorly characterised genomes • Effective in 'pathogen fishing' approaches to identify potential underlying pathogen • Low number of PCR cycles limits introduction of amplification mutations • Preservation of minor variant frequencies	• High sequencing cost to obtain sufficient pathogen sequence • Relatively low sensitivity to target pathogen, and coverage proportional to viral load • High proportion of non-pathogen reads increases computational challenges • Incidental sequencing of human and off-target pathogens raises ethical/diagnostic issues

	reflects *in vivo* variation • No primer/probe design enables rapid response to novel pathogens or sequence variants.	
PCR	• Tried and trusted – large technical resource of well-established methods and trained staff • Highly specific – most sequencing reads will be pathogen, reducing sequencing costs • Highly sensitive, with good coverage achievable even at low pathogen load • Relatively straightforward to introduce new primer designs 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 • Increased specificity over metagenomics reduces sequencing costs • Overlapping tiling of probes increases tolerance for individual primer mismatches	• High cost and technical expertise for sample preparation • Unable to sequence novel pathogens and requires well characterised reference genomes for probe design • Sensitivity is comparable to PCR but coverage is proportional to pathogen load – low pathogen load yields

	• Reduced number of PCR cycles (relative to PCR) limits introduction of amplification mutations • Preservation of minor variant frequencies reflects *in vivo* variation	low/incomplete coverage • Cost and time to generate new probe sets limits rapid response to emerging/novel sequences

550

551 **Table 2. Limitations of viral sequencing**

	Bacteria	**Viruses.**	**Challenges**
Genome	dsDNA	dsDNA, ssDNA, partially dsDNA, ssRNA, dsRNA	Different extraction protocols for different viruses, use of cDNA synthesis in RNA viruses or second strand synthesis for ssDNA viruses,
Gene Conservation	Bacteria have highly conserved genes essential for life (e.g. 16s) 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 virome.
Culture	Often straightforward to culture and obtain pure, highly enriched bacterial DNA/RNA	Challenging to culture, and requires a host cell for replication	Cultured virus is heavily contaminated with host cell genome/transcriptome, reducing equivalent 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	Viruses are intracellular pathogens, and cannot easily be separated from clinical samples prior to extraction	Clinical specimens are heavily contaminated with host genome/transcriptome, reducing equivalent viral sequencing output

	prior to extraction		
Bacterial methylation patterns	Bacteria are prokaryotes and use different methylation patterns from eukaryotes. Host DNA can be depleted post-extraction using restriction endonucleases directed against CpG methylation	DNA viruses are often methylated by 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 post-extraction

References

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, doi:10.1038/35057062 (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, doi:10.1038/507294a (2014).

5 Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449**, 804-810, doi:10.1038/nature06244 (2007).

6 Grigoriev, I. V. *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res* **42**, D699-704, doi:10.1093/nar/gkt1183 (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, doi:10.1097/GIM.0b013e3182088158 (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, doi:10.1016/S0140-6736(13)60632-7 (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, doi:10.1128/mBio.01974-14 (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, doi:10.1128/JCM.01532-09 (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, doi:10.1128/JCM.02993-14 (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, doi:10.1086/596736 (2009).
14 Smith, G. J. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122-1125, doi:10.1038/nature08182 (2009).
15 Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369-1372, doi:10.1126/science.1259657 (2014).
16 Koser, C. U. *et al.* Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog* **8**, e1002824, doi:10.1371/journal.ppat.1002824 (2012).
17 Cartwright, E. J., Koser, C. U. & Peacock, S. J. Microbial sequences benefit health now. *Nature* **471**, 578, doi:10.1038/471578d (2011).
18 Paredes, R. & Clotet, B. Clinical management of HIV-1 resistance. *Antivir Res* **85**, 245-265, doi:10.1016/j.antiviral.2009.09.015 (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, doi:10.1016/j.coviro.2015.06.001 (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, doi:10.3851/IMP2476 (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, doi:10.3748/wjg.v20.i19.5708 (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, doi:10.1016/j.jcv.2016.03.005 (2016).
25 Pawlotsky, J. M. Hepatitis C Virus Resistance to Direct-Acting Antiviral Drugs in Interferon-Free Regimens. *Gastroenterology* **151**, 70-86, doi:10.1053/j.gastro.2016.04.003 (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, doi:10.1128/AAC.05115-14 (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, doi:10.1371/journal.pone.0157600 (2016).
29 Piret, J. & Boivin, G. Antiviral drug resistance in herpesviruses other than cytomegalovirus. *Rev Med Virol* **24**, 186-218, doi:10.1002/rmv.1787 (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, doi:10.2147/IDR.S79131 (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, doi:10.1093/infdis/jiw073 (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, doi:10.1128/JVI.01522-07 (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, doi:10.1053/j.gastro.2014.05.010 (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, doi:10.1128/JCM.02149-10 (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, doi:10.1128/JVI.03391-14 (2015).
38 Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228-232, doi:10.1038/nature16996 (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 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, doi:10.1093/cid/civ393 (2015).
41 Faria, N. R. *et al.* Zika virus in the Americas: Early epidemiological and genetic findings. *Science* **352**, 345-349, doi:10.1126/science.aaf5036 (2016).
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, doi:10.7554/eLife.01123 (2013).

43 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, doi:10.1371/journal.pbio.1002251 (2015).
44 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, doi:10.1371/journal.pone.0086771 PONE-D-13-40833 [pii] (2014).
45 Zhou, B. *et al.* Composition and Interactions of Hepatitis B Virus Quasispecies Defined the Virological Response During Telbivudine Therapy. *Sci Rep* **5**, 17123, doi:10.1038/srep17123 (2015).
46 Itakura, J. *et al.* Resistance-Associated NS5A Variants of Hepatitis C Virus Are Susceptible to Interferon-Based Therapy. *PLoS One* **10**, e0138060, doi:10.1371/journal.pone.0138060 (2015).
47 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).
48 Swenson, L. C., Daumer, M. & Paredes, R. Next-generation sequencing to assess HIV tropism. *Curr Opin HIV AIDS* **7**, 478-485, doi:10.1097/COH.0b013e328356e9da (2012).
49 Hutter, G. *et al.* Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *N Engl J Med* **360**, 692-698, doi:10.1056/NEJMoa0802905 (2009).
50 Kordelas, L. *et al.* Shift of HIV tropism in stem-cell transplantation with CCR5 Delta32 mutation. *N Engl J Med* **371**, 880-882, doi:10.1056/NEJMc1405805 (2014).
51 Coaquette, A. *et al.* Mixed cytomegalovirus glycoprotein B genotypes in immunocompromised patients. *Clin Infect Dis* **39**, 155-161, doi:10.1086/421496 (2004).
52 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, doi:10.1128/JVI.02011-08 (2009).
53 Chou, S. *et al.* Improved detection of emerging drug-resistant mutant cytomegalovirus subpopulations by deep sequencing. *Antimicrob Agents Chemother* **58**, 4697-4702, doi:10.1128/AAC.03214-14 (2014).
54 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, doi:10.1016/j.jcv.2015.11.009 (2015).
55 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, doi:10.1128/AAC.00038-16 (2016).
56 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, doi:10.1128/JVI.00823-11 (2011).
57 Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74, doi:10.1126/science.1093857 (2004).
58 Mulcahy-O'Grady, H. & Workentine, M. L. The Challenge and Potential of Metagenomics in the Clinic. *Frontiers in immunology* **7**, 29, doi:10.3389/fimmu.2016.00029 (2016).
59 Morfopoulou, S. *et al.* Human Coronavirus OC43 Associated with Fatal Encephalitis. *N Engl J Med* **375**, 497-498, doi:10.1056/NEJMc1509458 (2016).

60 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, doi:10.1093/cid/ciu912 (2015).
61 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, doi:10.1038/srep15223 (2015).
62 Wilson, M. R. *et al.* Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med* **370**, 2408-2417, doi:10.1056/NEJMoa1401268 (2014).
63 Hoffmann, B. *et al.* A Variegated Squirrel Bornavirus Associated with Fatal Human Encephalitis. *N Engl J Med* **373**, 154-162, doi:10.1056/NEJMoa1415627 (2015).
64 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, doi:10.1016/j.jviromet.2015.09.010 (2015).
65 Duncan, C. J. *et al.* Human IFNAR2 deficiency: Lessons for antiviral immunity. *Sci Transl Med* **7**, 307ra154, doi:10.1126/scitranslmed.aac4227 (2015).
66 Depledge, D. P. *et al.* Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS One* **6**, e27805, doi:10.1371/journal.pone.0027805 (2011).
67 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).
68 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, doi:10.1186/PREACCEPT-1698056557139770 (2014).
69 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).
70 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, doi:10.1038/srep16706 (2015).
71 Kohl, C. *et al.* Protocol for metagenomic virus detection in clinical specimens. *Emerg Infect Dis* **21**, 48-57, doi:10.3201/eid2101.140766 (2015).
72 Sauvage, V. & Eloit, M. Viral metagenomics and blood safety. *Transfusion clinique et biologique : journal de la Societe francaise de transfusion sanguine* **23**, 28-38, doi:10.1016/j.tracli.2015.12.002 (2016).
73 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, doi:10.3389/fcimb.2014.00025 (2014).
74 Oude Munnink, B. B. *et al.* Autologous antibody capture to enrich immunogenic viruses for viral discovery. *PLoS One* **8**, e78454, doi:10.1371/journal.pone.0078454 (2013).
75 Sabina, J. & Leamon, J. H. Bias in Whole Genome Amplification: Causes and Considerations. *Methods in molecular biology* **1347**, 15-41, doi:10.1007/978-1-4939-2990-0_2 (2015).
76 Jensen, R. H. *et al.* Target-dependent enrichment of virions determines the reduction of high-throughput sequencing in virus discovery. *PLoS One* **10**, e0122636, doi:10.1371/journal.pone.0122636 (2015).
77 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).

78 Mlakar, J. *et al.* Zika Virus Associated with Microcephaly. *N Engl J Med* **374**, 951-958, doi:10.1056/NEJMoa1600651 (2016).
79 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, doi:10.3389/fmicb.2015.00224 (2015).
80 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, doi:10.1016/S1473-3099(15)70093-9 (2015).
81 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, doi:10.12688/f1000research.6743.2 (2015).
82 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, doi:10.1093/infdis/jiv271 (2015).
83 Cotten, M. *et al.* Deep sequencing of norovirus genomes defines evolutionary patterns in an urban tropical setting. *J Virol* **88**, 11056-11069, doi:10.1128/JVI.01333-14 (2014).
84 Kundu, S. *et al.* Next-generation whole genome sequencing identifies the direction of norovirus transmission in linked patients. *Clin Infect Dis* **57**, 407-414, doi:10.1093/cid/cit287 (2013).
85 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, doi:10.1128/JVI.00840-15 (2015).
86 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, doi:10.1128/JVI.00736-12 (2012).
87 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, doi:10.1093/infdis/jis679 (2013).
88 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, doi:10.1007/s00705-015-2424-0 (2015).
89 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).
90 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, doi:10.1371/journal.ppat.1001344 (2011).
91 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, doi:10.1186/1471-2334-14-275 (2014).
92 Johnson, T. A. *et al.* Clusters of Antibiotic Resistance Genes Enriched Together Stay Together in Swine Agriculture. *MBio* **7**, e02214-02215, doi:10.1128/mBio.02214-15 (2016).
93 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, doi:10.12688/f1000research.7111.1 (2015).

94 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).

95 Tsangaras, K. *et al.* Hybridization capture using short PCR products enriches small genomes by capturing flanking sequences (CapFlank). *PLoS One* **9**, e109101, doi:10.1371/journal.pone.0109101 (2014).

96 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, doi:10.1093/molbev/mst210 (2014).

97 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, doi:10.1128/JVI.01654-13 (2013).

98 Palser, A. L. *et al.* Genome diversity of Epstein-Barr virus from multiple tumor types and normal infection. *J Virol* **89**, 5222-5237, doi:10.1128/JVI.03614-14 (2015).

99 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**, doi:10.1128/genomeA.01498-14 (2015).

100 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**, doi:10.1128/genomeA.00830-13 (2013).

101 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, doi:10.1016/j.virol.2015.10.025 (2016).

102 Briese, T. *et al.* Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. *MBio* **6**, e01491-01415, doi:10.1128/mBio.01491-15 (2015).

103 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).

104 Hofmann, B. Incidental findings of uncertain significance: To know or not to know--that is not the question. *BMC Med Ethics* **17**, 13, doi:10.1186/s12910-016-0096-2 (2016).

105 England, P. H. Good Laboratory Practice when Performing Molecular Amplification Assays. Q 4 Issue 4.4. *UK Standards for Microbiology Investigations.* **Q 4** (2013).

106 Viana, R. V. & Wallis, C. L. *Good Clinical Laboratory Practice (GCLP) for Molecular Based Tests Used in Diagnostic Laboratories*. (INTECH Open Access Publisher, 2011).

107 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, doi:10.1016/j.bdq.2015.08.003 (2015).

108 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, doi:10.3389/fmicb.2015.00188 (2015).

109 Munro, A. C. & Houldcroft, C. Human cancers and mammalian retroviruses: should we worry about bovine leukemia virus? *Future Virology* **11**, 163-166, doi:10.2217/fvl.16.5 (2016).

110 Hue, S. *et al.* Disease-associated XMRV sequences are consistent with laboratory contamination. *Retrovirology* **7**, 111, doi:10.1186/1742-4690-7-111 (2010).

111 Erlwein, O. *et al.* DNA extraction columns contaminated with murine sequences. *PLoS One* **6**, e23484, doi:10.1371/journal.pone.0023484 (2011).

112 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, doi:10.1111/tbed.12251 (2014).
113 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, doi:10.1128/JVI.02323-13 (2013).
114 Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* **12**, 87, doi:10.1186/s12915-014-0087-z (2014).
115 Lipkin, W. I. A Vision for Investigating the Microbiology of Health and Disease. *J Infect Dis* **212 Suppl 1**, S26-30, doi:10.1093/infdis/jiu649 (2015).
116 Shafer, R. W. Rationale and uses of a public HIV drug-resistance database. *J Infect Dis* **194 Suppl 1**, S51-58, doi:10.1086/505356 (2006).
117 Gnaneshan, S., Ijaz, S., Moran, J., Ramsay, M. & Green, J. HepSEQ: International Public Health Repository for Hepatitis B. *Nucleic Acids Res* **35**, D367-370, doi:10.1093/nar/gkl874 (2007).
118 Rhee, S. Y. *et al.* Hepatitis B virus reverse transcriptase sequence variant database for sequence analysis and mutation discovery. *Antiviral Res* **88**, 269-275, doi:10.1016/j.antiviral.2010.09.012 (2010).
119 Kuiken, C., Yusim, K., Boykin, L. & Richardson, R. The Los Alamos hepatitis C sequence database. *Bioinformatics* **21**, 379-384, doi:10.1093/bioinformatics/bth485 (2005).
120 Hong, L. Z. *et al.* BAsE-Seq: a method for obtaining long viral haplotypes from short sequence reads. *Genome Biol* **15**, 517, doi:10.1186/PREACCEPT-6768001251451949 (2014).
121 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).
122 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).
123 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).
124 Greninger, A. L. *et al.* Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med* **7**, 99, doi:10.1186/s13073-015-0220-9 (2015).
125 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).
126 Bull, R. A. *et al.* A method for near full-length amplification and sequencing for six hepatitis C virus genotypes. *BMC Genomics* **17**, 247, doi:10.1186/s12864-016-2575-8 (2016).
127 Kimberlin, D. W. & Whitley, R. J. Antiviral resistance: mechanisms, clinical significance, and future implications. *J Antimicrob Chemother* **37**, 403-421 (1996).
128 Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333-351, doi:10.1038/nrg.2016.49 (2016).

129 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, doi:10.1093/nar/gkv773 (2015).
130 Guan, H. *et al.* Detection of virus in CSF from the cases with meningoencephalitis by next-generation sequencing. *J Neurovirol* **22**, 240-245, doi:10.1007/s13365-015-0390-7 (2016).
131 Brown, J. R. *et al.* Astrovirus VA1/HMO-C: an increasingly recognized neurotropic pathogen in immunocompromised patients. *Clin Infect Dis* **60**, 881-888, doi:10.1093/cid/ciu940 (2015).
132 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, doi:10.1093/jpids/piv040 (2015).
133 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).
134 Barjas-Castro, M. L. *et al.* Probable transfusion-transmitted Zika virus in Brazil. *Transfusion*, doi:10.1111/trf.13681 (2016).
135 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).
136 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).