PHILOSOPHICAL
TRANSACTIONS B

rstb.royalsocietypublishing.org

Review

Cite this article: Tang K-W, Larsson E. 2017 Tumour virology in the era of high-throughput genomics. *Phil. Trans. R. Soc. B* **372**: 20160265. http://dx.doi.org/10.1098/rstb.2016.0265

Accepted: 9 June 2017

One contribution of 14 to a theme issue 'Human oncogenic viruses'.

Subject Areas:
bioinformatics, genomics, microbiology

Keywords:
tumour virus, next-generation sequencing, virus integration

Author for correspondence:
Erik Larsson
e-mail: erik.larsson@gu.se

Tumour virology in the era of high-throughput genomics

Ka-Wei Tang[1] and Erik Larsson[2]

[1]Department of Infectious Diseases, and [2]Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Medicinaregatan 9A, 405 30 Gothenburg, Sweden

EL, 0000-0003-1400-0119

With the advent of massively parallel sequencing, oncogenic viruses in tumours can now be detected in an unbiased and comprehensive manner. Additionally, new viruses or strains can be discovered based on sequence similarity with known viruses. Using this approach, the causative agent for Merkel cell carcinoma was identified. Subsequent studies using data from large collections of tumours have confirmed models built during decades of hypothesis-driven and low-throughput research, and a more detailed and comprehensive description of virus–tumour associations have emerged. Notably, large cohorts and high sequencing depth, in combination with newly developed bioinformatical techniques, have made it possible to rule out several suggested virus–tumour associations with a high degree of confidence. In this review we discuss possibilities, limitations and insights gained from using massively parallel sequencing to characterize tumours with viral content, with emphasis on detection of viral sequences and genomic integration events.

This article is part of the themed issue 'Human oncogenic viruses'.

1. Introduction

Seven known human tumour viruses, discovered using a variety of techniques, are causative agents for a large fraction of human cancers [1]. Animal and human model tumour viruses have served as important tools for studies of tumorigenesis and were essential in establishing key concepts such as oncogenes and tumour suppressor genes [2,3]. Recently, the introduction of massively parallel sequencing, also known as next-generation sequencing (NGS), has revolutionized characterization of genomic and transcriptomic alterations in tumours. In addition, NGS-based approaches are now increasingly being applied to the study of viral nucleic acids in tumours.

The most recently discovered human tumour virus, Merkel cell polyomavirus (MCV), responsible for the majority of Merkel cell carcinomas, was identified using a pioneering bioinformatical method, digital transcriptome subtraction of sequences generated by NGS [4]. The principle was reminiscent of earlier molecular biological techniques for enrichment and sequencing of viral genetic material, which led to the discovery of hepatitis C virus (HCV) and Kaposi's sarcoma associated herpesvirus (KSHV or human herpesvirus 8 (HHV8)) [5,6]. However, the larger amounts of data generated by NGS now allowed the enrichment process to be performed *in silico* using bioinformatics, by first removing human sequences followed by unbiased detection of viral traces in the remaining data. Variants of this approach have since been used in many subsequent NGS-based studies. These have confirmed previously described virus–tumour associations and added several other insights, including rare associations, novel recurrent sites of virus integration, and rejection of previously proposed associations. Here we review some of the contributions NGS has made in this field, with the main focus on detection of viral sequences in cancer.

THE ROYAL SOCIETY
PUBLISHING

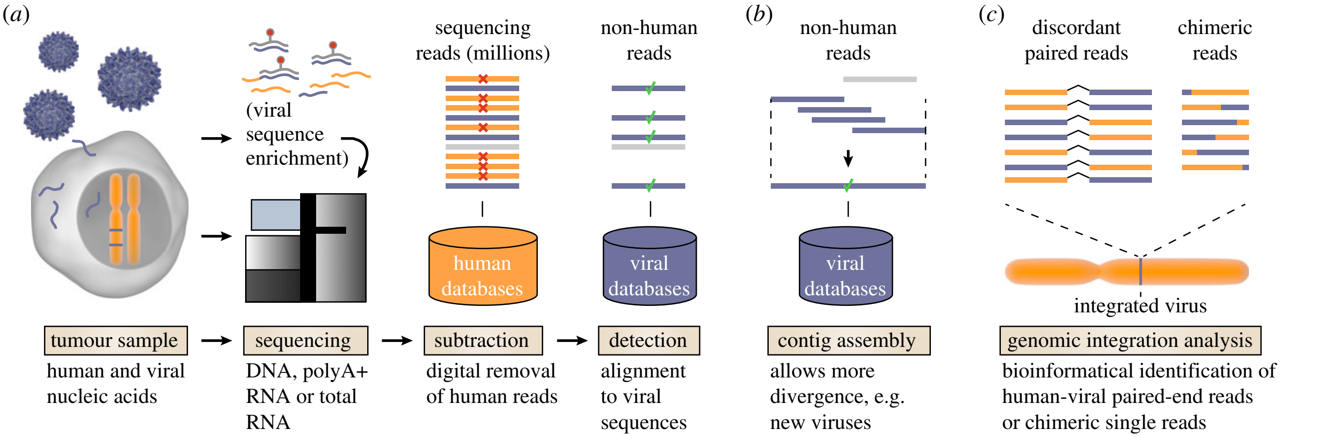
(a)
(b)
(c)
(viral sequence enrichment)
sequencing reads (millions)
non-human reads
non-human reads
discordant paired reads
chimeric reads
human databases
viral databases
viral databases
integrated virus
tumour sample
sequencing
subtraction
detection
contig assembly
genomic integration analysis
human and viral nucleic acids
DNA, polyA+ RNA or total RNA
digital removal of human reads
alignment to viral sequences
allows more divergence, e.g. new viruses
bioinformatical identification of human-viral paired-end reads or chimeric single reads

Figure 1. Detecting viruses in tumour samples using high-throughput sequencing. (*a*) RNA (polyA+ or total) or DNA prepared from tumour tissue using standard protocols is subjected to high-throughput sequencing, producing millions to billions of short sequencing reads. Alternative protocols allow for enrichment of viral nucleic acids prior to sequencing. Typically, human sequencing reads are then bioinformatically subtracted and the remaining data are compared against known viral reference sequences, such as available complete viral genomes. (*b*) Bioinformatical assembly of non-human reads into longer sequences (contigs) prior to comparison to viral references allows for detection of more distant evolutionary relationships including new viral species. (*c*) Viral genomic integrations can be revealed by identifying discordant read pairs from paired-end sequencing where one mate aligns to human and the other to viral reference sequences. Individual chimeric human-viral reads allow fine-mapping of genomic integration breakpoints.

2. Detection of viruses in tumours using high-throughput sequencing

NGS-based characterization of viral sequences in tumour material generally presents few experimental challenges, since standard protocols for nucleic acid extraction and sequencing are applicable. Sequencing libraries generated for general transcriptomic or genomic analyses of tumours can therefore be repurposed for viral analyses, which has enabled large cancer cohorts to be screened for viral content solely using bioinformatics and publicly available sequencing data [7–9].

Typically, tumour tissue is flash frozen after harvesting, and pathological tissue slides are prepared and analysed to define the borders of the tumour. Samples with high tumour content are then lysed and nucleic acids are purified. The type of nucleic acid analyte analysed (typically mRNA, total RNA or DNA) will determine what types of viruses can be detected and what kinds of studies may be performed. For example, DNA-based analyses can reveal integrated and latent non-expressed viruses, and may enable quantification of absolute viral load per human cell [10]. Transcriptome sequencing, on the other hand, may reveal non-retrotranscribed RNA viruses that will go undetected in DNA data, and also provides crucial insight into viral and host gene activities. Sequencing libraries are typically prepared by fragmentation into appropriate sizes and by adding specific adaptors to the fragments. Next, the fragments are immobilized to two-dimensional surfaces on flowcells, where they are amplified by solid phase PCR and subjected to a sequencing-by-synthesis reaction using fluorescently labelled nucleotides. With current technology, in the order of tens to hundreds of millions of short sequencing reads will be produced for a single transcriptome, and more still in the case of whole genome sequencing (WGS) (figure 1*a*).

The bioinformatical analysis generally involves removal of low-fidelity reads followed by matching against human reference sequences. Remaining non-human reads are finally matched against a viral genome database (figure 1*a*) [7–9,11–19]. Variability in sequencing depth is typically accounted for by normalizing to the total number of obtained reads, for example by stating viral expression levels as 'reads per kilobase and million base pairs sequenced' (RPKM) or in parts-per-million (ppm) of total library reads. Greater sensitivity for detecting highly diverged viral strains or new viruses can be obtained by first assembling non-human sequences into longer contiguous segments (contigs), followed by searches for homology to known viral reference sequences (figure 1*b*) [7,9,12,13,15,16,18]. Furthermore, sites of viral genomic integration can be bioinformatically pinpointed by identification of discordant paired reads or chimeric human-viral sequences (figure 1*c*, discussed below). Several software packages are now available to simplify these tasks, reducing the expertise required [7,14,17,18].

3. Reference results from known virus-associated tumours

Early NGS-based studies of tumour viruses were limited by the relatively low sequencing depth available at the time. Bioinformatical processing was carried out meticulously with every sequence read analysed and categorized [19]. Transcriptomic sequencing of four Merkel cell carcinomas using pyrosequencing yielded less than four-hundred thousand high-fidelity reads of which two unknown transcripts led to the discovery of the Merkel cell polyomavirus [4]. Later analyses of larger patient cohorts using more current sequencing methodologies have established NGS as an efficient method for detection of viral mRNA [8,9,11,20–23]. In particular, studies of tumours with known viral aetiology have been important in establishing a point of reference against which novel observations can be compared. With very few exceptions, these studies show that a single type of virus is predominant in each tumour.

Analysis of 85 cervical squamous cell carcinomas and endocervical adenocarcinomas, and 43 head and neck squamous cell carcinomas, infected with various types of high-risk human papillomaviruses (HPV), revealed an average of approximately 200 ppm of viral mRNA reads in relation to the total library size, ranging from 11 to 598 ppm for cervical squamous cell carcinoma and endocervical adenocarcinoma, and 22 to 848 ppm for head and neck squamous cell carcinoma. Similarly, the average hepatitis B virus (HBV) mRNA content in 11 liver hepatocellular carcinomas was also nearly 200 ppm, ranging from 2 to 854 ppm with three tumours containing less than 10 ppm viral mRNA [9]. The low proportion of HBV reads in some samples likely reflects the fact that some HBV-initiated liver tumours are able to proliferate in the absence of the viral genome.

Results from six AIDS-associated lymphomas containing Epstein–Barr virus (EBV/HHV4) revealed viral expression ranging from 145 to 8857 ppm, with an average of 2750 ppm EBV transcripts [24]. Interestingly, 24 EBV-positive gastric adenocarcinoma showed on average only 88 ppm viral mRNA, ranging from 4 to 300 ppm [22]. The significantly higher values in the AIDS-associated lymphomas may conceivably reflect the absence of humoral or cell-mediated surveillance.

A limit of 10 ppm viral mRNA reads (corresponding to 100 viral reads at a sequencing depth of 10 million reads) has been suggested as an approximate divider for tumours with clonal presence of expressed viruses, since most virally induced tumours were found to surpass this level [9]. Lower levels can however not always be disregarded, exemplified by the Merkel cell polyomavirus which was discovered at 5 ppm [4]. Additionally, results presented above mostly derive from studies using polyA enrichment prior to sequencing, which theoretically excludes certain viruses such as HCV of the Flavivirus family. Nevertheless, HCV sequences have still been detected at very low levels in polyA enriched libraries [9,11], and there may thus be technical reasons as to why even weaker signals should be considered in some cases.

Recent studies employing small RNA sequencing, whole exome sequencing (WXS) and WGS data for identification of viruses in tumours have produced results that are largely consistent with transcriptomic analyses [10,21–23,25]. It should be noted that the viral signals seen in WXS-based analyses are typically weak, sometimes with only single viral reads observed even in known virus-associated tumours [22], which is explained by the host sequence enrichment step inherent to the methodology. WGS, in contrast to WXS and transcriptome sequencing, produces a constant host genome background that can be useful to estimate absolute viral genome copy number per cell [10]. As an example, at one EBV (170 kb genome) per cell (6.4 Mb diploid genome), and assuming 100% tumour cell content, one would expect approximately 27 ppm of total library reads to be of viral origin.

4. Low-level detection and contamination

The sensitivity and unbiased nature of NGS gives rise to a new type of problem, where trace amounts of human as well as non-human viruses are often detected in tumours and control tissues [26–28]. These signals can arise for several reasons, one being infiltration of virus-positive lymphocytes in tumour tissue. This has been shown, for example, in AIDS-associated lymphomas, where low levels of EBV transcripts detected by NGS were confirmed by *in situ* hybridization to be due to infiltration of latently EBV-infected lymphocytes [24]. Beta (HHV5/CMV, HHV6 and HHV7) and gammaherpesviruses (HHV4 and HHV8) as well as HIV-1 are also known to infect and establish latency in haematopoietic progenitor cells and lymphocytes [29]. This likely explains why low levels of viral transcripts from these agents have been detected by NGS in tumours as well as healthy control tissue from several cancers [9,11,25].

Viral signals may also arise from infected tissue surrounding the tumour. Primary herpes simplex virus 1 (HSV1) infection occurs predominantly in the oropharyngeal area with ensuing cold sores [29]. Possibly, this explains high levels of HSV1 detected in one head and neck squamous cell carcinoma, which could not be confirmed by immunohistochemistry [21]. HSV1 has also been detected in several oesophageal carcinomas [11], but it is not clear whether these signals originate from tumour cells or surrounding cells.

An additional challenge is contaminants, which may be introduced at all steps during sample preparation or downstream processing [30]. Silica membranes in some nucleic acid extraction kits have been shown to contain algae viruses, which were mistakenly classified as a new hepatitis virus [31,32]. Other reagents and components of the laboratory environment can also contain contaminants, which may be of human, animal, invertebrate, plant, fungi and bacterial origin [26,27,33,34]. Unexpected microbial detections in NGS libraries have sometimes been linked to specific sequencing centres and timepoints, further supporting that they represent contaminants [34,35].

Sequences from non-human viruses are typically present only in low amounts when detected by NGS in tumours [9,26]. Although possibly explained by zoonotic or environmental infections, most of these signals likely arise from contamination during sample processing or sequencing, or environmental exposure at the tumour site. The association of the murine XMRV with human prostate cancer mislead the scientific community for many years. Thorough investigation, including reanalysis of the original tissue sample, finally revealed this to be due to contamination [36]. Several NGS-based studies have since confirmed the absence of XMRV in large prostate cancer cohorts [8,9,11,13,37,38].

A frequently found synthetic viral contaminant is the immediate early promoter of the human cytomegalovirus (HHV5 or CMV) used in many mammalian expression plasmids [9–11,25]. Additionally, intentional phage spiking of sequencing libraries may cause confusion during downstream analyses [39]. Another possible source of synthetic viral sequences are cell lines where viral agents such as HPV, adenoviruses, EBV, retroviruses and SV40 have been used for transformation [12]. HeLa cells naturally harbour HPV18 and have been known to cross-contaminate cell lines throughout the world. Recently, it was shown that low levels of HPV18 detected in some colorectal tumours sequenced by The Cancer Genome Atlas (TCGA) were due to HeLa contamination, as evidenced by an identical HeLa-specific HPV genotype in these samples [35]. The cutaneous HPV38 has also been suggested to be present as a contaminant in endometrial cancer RNA sequencing libraries from TCGA [40]. Also in TCGA data, a single clear cell renal cell carcinoma was found to contain HBV mRNA [9]. However, closer analysis also revealed weak but consistent expression of liver marker mRNAs, supporting contamination by HBV-positive liver tumour mRNA.

The examples described here stress the need to maintain a critical standpoint towards novel virus–tumour associations detected by NGS. Negative control samples and complementary laboratory assays such as *in situ* hybridization, immunohistochemistry and PCR of tumour and healthy tissue are useful to confirm initial findings [4,21,24,41]. Additionally, lack of viral genetic diversity in between different samples may indicate presence of a common contaminant [35,42]. Finally, even confirmed presence of a virus naturally does not imply causation, and overlapping epidemiologies for different viruses may further complicate interpretation. Conversely, viral presence is not obligate in tumours initiated by chronic inflammation caused by viruses. Specific criteria for defining virus–tumour associations are therefore discouraged, and we should not rely solely on one method but find multiple biological indicators that together convincingly can justify the virus as a causative agent [43].

5. Rare virus–tumour associations

Viruses that have co-evolved with humans as their main host are typically highly selective in terms of the cell types they can infect [44]. Detections beyond this preferred range of cell types are therefore uncommon and can indeed often be explained by contaminations, as discussed above. However, some rare associations detected by NGS warrant further consideration.

HPV16 is one of the few viruses that have been associated with tumours outside the primary sites of infection in the ano-genital region. In addition to expected detections in head and neck and cervical tumours, HPV16 has been found in single uterus, lung and bladder carcinomas sequenced by TCGA [8,9,11]. While supported by some earlier studies [45–47], this still represents a very small fraction of tumours from these locations. Additionally, a recent NGS-based study reported HPV16 in 3 out of 530 low grade gliomas [11,25]. Further verification using *in situ* hybridization or immunohistochemistry is needed to confirm these observations.

EBV is another agent implicated in a wide range of cancers including Burkitt's lymphoma, nasopharyngeal carcinoma, Hodgkin's lymphoma and gastric adenocarcinoma. Additionally, transcriptomic analysis recently revealed high levels of EBV in 2 out of 105 diffuse large B-cell lymphomas (432 and 37 ppm) [48]. Two tumours were also positive for HHV6 in this cohort (99 and 19 ppm), in one case coinciding with EBV infection. Analysis of viral gene expression patterns further supported a causal role for EBV, while HHV6 was suggested to be due to disease-related immunosuppression.

A single bladder urothelial carcinoma, out of 316 characterized using transcriptome sequencing by TCGA, was found to contain BK polyomavirus (BKV) [8,9,20]. Earlier reports of this virus in bladder carcinoma using low-throughput diagnostic methods have been contrasting [49]. The oncogenic BKV T-antigen was expressed at high levels (318 ppm) and the viral genome was shown to be integrated into the host genome [9]. Although this supports a functional contribution from BKV in rare cases, it should be noted that BK as well as other polyomaviruses often cause asymptomatic infections and are ubiquitous in humans [50].

6. Non-detection

The literature contains a large number of proposed virus–tumour associations that are controversial. These are typically based on traditional viral diagnostic techniques, including PCR, immunohistochemistry, *in situ* hybridization and western blotting, all of which are prone to false positive detections. Modern genomic approaches, which allow unbiased screening of large tumour cohorts, have the potential to bring clarity to some of these proposed associations.

During the 1960s and 70s, it was widely believed that HSV2, which causes genital herpes, was the causative agent for cervical carcinoma. High-risk HPV types were later identified in these tumours and years of disputes followed [51] before high-risk HPV was finally established as the *de facto* causative agent [52]. Today, NGS-based studies of large cohorts confirm that more than 90% of cervical carcinomas express high levels of high-risk HPV, while no HSV2 sequences can be detected [9,11,23]. Hence, this could have been clarified faster had high-throughput sequencing been available at the time.

Breast cancer is the most common invasive cancer in the world, and has been extensively studied. Several viruses including EBV, HPV and MMTV (mouse mammary tumour virus) have been implicated [53]. Frequent clonal presence and expression of EBV or HPV can be ruled out, considering that transcriptomic data from more than 800 breast tumours have now been analysed without any significant levels of these viruses being detected [8,9,54]. A small number of reads aligning to MMTV (9 out of more than 1.5 billion) were detected in the same cohort [9]. These trace amounts are suggestive of contamination, but silent genomic integration could still be possible and WGS-based analysis is warranted to resolve this.

The role of CMV in human cancer has been highly controversial. After the initial claim that CMV DNA and protein was found clonally in the majority of gliomas [55], CMV has now been associated with a wide variety of other cancers in the literature. Ubiquitous presence of CMV has been proposed in most types of brain tumours, but this has been contested in other reports [56]. All NGS-based studies of non-enriched glioma material, in total more than 700 samples, have concluded that CMV RNA cannot be detected [8,9,11,16,25]. Likewise, analysis of deep coverage WGS data from 34 glioblastoma multiforme tumours failed to reveal CMV [10]. The examples discussed here show that NGS-based approaches have great potential to bring clarity to debated virus–tumour associations.

7. Viral genomic integration

Retroviruses such as HTLV-1 (human T-lymphotropic virus type 1) establish chronic infection by genome integration, which causes the virus to be propagated in the host for long periods of time [57,58]. Integration of DNA viruses into tumour genomes appear to be random events, although these may be facilitated by disruption of DNA repair pathways by viral gene products [59,60]. The integration of MCV appears to be obligate for tumorigenesis of Merkel cell carcinoma, while certain types of HPV display a low rate of integration [9,11,61]. Genomic viral integration may contribute to cellular transformation by insertion of strong viral promoters near oncogenes or by disruption of tumour suppressors [60,62]. By using high-throughput sequencing, it is now possible to study these events in great detail on a genome-wide scale.

Two main principles, applicable to both DNA and RNA sequencing data, are used for detection of viral integrations (figure 1*c*) [7–9,11,13,14,16,17,40,63–65]. The first involves identification of discordant viral-human read pairs in paired-end sequencing data, where both tails of DNA fragments are sequenced. Challenges include false chimeric pairs that may arise for technical reasons [9]. The second approach takes advantage of individual reads that map partly to human and partly to viral sequences. This offers base-pair resolution, but such reads are more rare and higher sequencing depth is therefore required. Performance can be improved by applying a combination of both aforementioned techniques, by enrichment of viral sequences prior to sequencing (figure 1*a*), and by combining transcriptome and WGS data [62,66].

HPV integrations have been studied for decades, and methods for enrichment and sequencing of integration points have been developed both for fusion transcripts and integrated DNA [60]. Recent studies using NGS have verified previously described integration breakpoint hotspots in both the viral and host sequence [11,67]. Several studies have shown that integrations often coincide with copy number gains and cellular genes showing elevated expression [9,11,21,23]. Possibly, initiation of replication of integrated viral DNA and subsequent activation of DNA damage responses can contribute to copy number gains in these regions [68].

Likewise, HBV integration has been thoroughly studied, and recurrent viral integrations have been detected in specific genomic regions [69]. During primary HBV infection the virus spreads throughout the whole liver. Interestingly, NGS-based analyses suggest that the preferred locations for integration may differ between tumours and surrounding liver, since integrations in the *FN1* gene were frequent only in non-cancerous tissue [11,69–71].

EBV integration has previously been described in cell lines, but not in tumours. Surprisingly, transcriptomic analysis of 24 EBV-positive gastric adenocarcinomas revealed several reads that supported integration in one single tumour, suggesting that EBV integration may occur also in clinical samples [22].

Comprehensive mapping of HTLV-1 integrations by enrichment of insertion sites before sequencing has enabled monitoring of latently infected CD4+ T-lymphocyte populations before the onset of adult T-cell leukaemia, and has revealed that malignant cells most often originate from single clones and not through oligoclonal proliferation as previously proposed [57,65]. As exemplified here, NGS-based approaches have been useful for confirming as well as extending our knowledge of viral integrations in tumours genomes.

8. Future perspectives

NGS was successfully used in the discovery of MCV in Merkel cell carcinoma [4], but it should be noted that subsequent attempts to identify novel oncogenic viruses in large patient cohorts using genomic approaches have been unsuccessful [8,9,11]. However, current *in silico* methods are highly dependent on existing databases of viral sequences, and many new viruses, including potential human pathogens [72,73], have been uncovered in recent years. While unlikely, it is possible that a new, sequence-unique, family of tumour viruses awaits discovery, possibly in rare tumours or in immunosuppressed patients, and it is almost certain that NGS would play a key role in such a finding.

Sequencing technologies continue to improve at a rapid pace. Each new generation of machines promises longer reads, shorter run times, and lower per-base-pair cost. While current technology is more than sufficient for the detection of clonally present viruses in tumours, many benefits can still be expected as the amount of data continues to increase. For example, more widespread use of total RNA transcriptomic analyses, rather than the polyA enriched protocols that dominate today, will provide a more comprehensive description of cellular RNA content including non-polyA viruses. High-coverage WGS, which is becoming increasingly accessible, gives improved possibilities for mapping of viral integration breakpoints, quantifying viral load, and better statistical strength for rejecting controversial virus–tumour associations. Larger cohorts will increase the chances of finding new viruses or rare associations. Furthermore, single-cell sequencing approaches will be able to shed light on heterogeneity in cell populations, including the interplay between virus-infected and surrounding cells [74].

The discovery of human tumour viruses has had a profound effect on the prevention of human cancer. Efficient vaccines and antiviral treatment have completely shifted the aetiological causes for cancer in developed countries. It is therefore of great importance for the scientific community to remain vigilant in the search for new virus–tumour associations. NGS is now proven to be extremely efficient for characterization of tumour viral content, and will soon be the primary tool for discovering, confirming, as well as rejecting virus–tumour associations.

Data accessibility. This article has no additional data.

Authors' contributions. K.-W.T. and E.L. wrote the article.

Competing interests. We have no competing interests.

Funding. K.-W.T. was supported by grants from ALF Foundation at Sahlgrenska University Hospital. E.L. was supported by grants from the Knut and Alice Wallenberg Foundation, the Swedish Foundation for Strategic Research, the Swedish Medical Research Council, the Swedish Cancer Society, the Åke Wiberg foundation, and Region Västra Götaland.

Acknowledgements. We thank the Editors of this special issue of Philosophical Transactions B for the invitation to contribute this article.

References

1. Moore PS, Chang Y. 2010 Why do viruses cause cancer? Highlights of the first century of human tumour virology. *Nat. Rev. Cancer* **10**, 878–889. (doi:10.1038/nrc2961)
2. Martin GS. 2001 The hunting of the Src. *Nat. Rev. Mol. Cell Biol.* **2**, 467–475. (doi:10.1038/35073094)
3. Whyte P, Buchkovich KJ, Horowitz JM, Friend SH, Raybuck M, Weinberg RA, Harlow E. 1988 Association between an oncogene and an anti-oncogene: the adenovirus E1A proteins bind to the retinoblastoma gene product. *Nature* **334**, 124–129. (doi:10.1038/334124a0)
4. Feng H, Shuda M, Chang Y, Moore PS. 2008 Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* **319**, 1096–1100. (doi:10.1126/science.1152586)
5. Chang Y, Cesarman E, Pessin MS, Lee F, Culpepper J, Knowles DM, Moore P. 1994 Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* **266**, 1865–1869. (doi:10.1126/science.7997879)
6. Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M. 1989 Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science* **244**, 359–362. (doi:10.1126/science.2523562)
7. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M. 2011 PathSeq: software to

identify or discover microbes by deep sequencing of human tissue. *Nat. Biotechnol.* **29**, 393–396. (doi:10.1038/nbt.1868)

8. Khoury JD *et al.* 2013 Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. *J. Virol.* **87**, 8916–8926. (doi:10.1128/JVI.00340-13)
9. Tang KW, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E. 2013 The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat. Commun.* **4**, 2513. (doi:10.1038/ncomms3513)
10. Tang KW, Hellstrand K, Larsson E. 2015 Absence of cytomegalovirus in high-coverage DNA sequencing of human glioblastoma multiforme. *Int. J. Cancer* **136**, 977–981. (doi:10.1002/ijc.29042)
11. Cao S *et al.* 2016 Divergent viral presentation among human tumors and adjacent normal tissues. *Sci. Rep.* **6**, 1127. (doi:10.1038/srep28294)
12. Borozan I *et al.* 2012 CaPSID: a bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. *BMC Bioinformatics* **13**, 206. (doi:10.1186/1471-2105-13-206)
13. Chen Y, Wei J. 2015 Identification of pathogen signatures in prostate cancer using RNA-seq. *PLoS ONE* **10**, e0128955. (doi:10.1371/journal.pone.0128955)
14. Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X. 2013 VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics* **29**, 266–267. (doi:10.1093/bioinformatics/bts665)
15. Cosset E *et al.* 2014 Comprehensive metagenomic analysis of glioblastoma reveals absence of known virus despite antiviral-like type I interferon gene response. *Int. J. Cancer* **135**, 1381–1389. (doi:10.1002/ijc.28670)
16. Schelhorn SE, Fischer M, Tolosi L, Altmuller J, Nurnberg P, Pfister H, Lengauer T, Berthold F, Wilke CO. 2013 Sensitive detection of viral transcripts in human tumor transcriptomes. *PLoS Comput. Biol.* **9**, e1003228. (doi:10.1371/journal.pcbi.1003228)
17. Wang Q, Jia P, Zhao Z. 2015 VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. *Genome Med.* **7**, 2. (doi:10.1186/s13073-015-0126-6)
18. Xu G, Strong MJ, Lacey MR, Baribault C, Flemington EK, Taylor CM. 2014 RNA CoMPASS: a dual approach for pathogen and host transcriptome analysis of RNA-seq datasets. *PLoS ONE* **9**, e89445. (doi:10.1371/journal.pone.0089445)
19. Feng H, Taylor JL, Benos PV, Newton R, Waddell K, Lucas SB, Chang Y, Moore PS. 2007 Human transcriptome subtraction by using short sequence tags to search for tumor viruses in conjunctival carcinoma. *J. Virol.* **81**, 11 332–13 340. (doi:10.1128/JVI.00875-07)
20. Cancer Genome Atlas Research Network. 2014 Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322. (doi:10.1038/nature12965)
21. Cancer Genome Atlas Network. 2015 Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582. (doi:10.1038/nature14129)
22. Cancer Genome Atlas Research Network. 2014 Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209. (doi:10.1038/nature13480)
23. Ojesina AI *et al.* 2014 Landscape of genomic alterations in cervical carcinomas. *Nature* **506**, 371–375. (doi:10.1038/nature12881)
24. Arvey A *et al.* 2015 The tumor virus landscape of AIDS-related lymphomas. *Blood* **125**, e14–e22. (doi:10.1182/blood-2014-11-599951)
25. Strong MJ, Blanchard 4th E, Lin Z, Morris CA, Baddoo M, Taylor CM, Ware ML, Flemington EK. 2016 A comprehensive next generation sequencing-based virome assessment in brain tissue suggests no major virus–tumor association. *Acta Neuropathol. Commun.* **4**, 71. (doi:10.1186/s40478-016-0338-z)
26. Andrade MA, Perez-Iratxeta C, Ponting CP. 2001 Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* **134**, 117–131. (doi:10.1006/jsbi.2001.4392)
27. Lusk RW. 2014 Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS ONE* **9**, e110808. (doi:10.1371/journal.pone.0110808)
28. Moustafa A *et al.* 2017 The blood DNA virome in 8,000 humans. *PLoS Pathog.* **13**, e1006292. (doi:10.1371/journal.ppat.1006292)
29. Lieberman PM. 2016 Epigenetics and genetics of viral latency. *Cell Host Microbe* **19**, 619–628. (doi:10.1016/j.chom.2016.04.008)
30. Ballenghien M, Faivre N, Galtier N. 2017 Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC Biol.* **15**, 25. (doi:10.1186/s12915-017-0366-6)
31. Naccache SN, Hackett Jr J, Delwart EL, Chiu CY. 2014 Concerns over the origin of NIH-CQV, a novel virus discovered in Chinese patients with seronegative hepatitis. *Proc. Natl Acad. Sci. USA* **111**, E976. (doi:10.1073/pnas.1317064111)
32. Xu B *et al.* 2013 Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing. *Proc. Natl Acad. Sci. USA* **110**, 10 264–10 269. (doi:10.1073/pnas.1303744110)
33. Friis-Nielsen J *et al.* 2016 Identification of known and novel recurrent viral sequences in data from multiple patients and multiple cancers. *Viruses* **8**, 53. (doi:10.3390/v8020053)
34. Strong MJ *et al.* 2014 Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS Pathog.* **10**, e1004437. (doi:10.1371/journal.ppat.1004437)
35. Cantalupo PG, Katz JP, Pipas JM. 2015 HeLa nucleic acid contamination in the cancer genome atlas leads to the misidentification of human papillomavirus 18. *J. Virol.* **89**, 4051–4057. (doi:10.1128/JVI.03365-14)
36. Paprotka T *et al.* 2011 Recombinant origin of the retrovirus XMRV. *Science* **333**, 97–101. (doi:10.1126/science.1205292)
37. Lee D *et al.* 2012 In-depth investigation of archival and prospectively collected samples reveals no evidence for XMRV infection in prostate cancer. *PLoS ONE* **7**, e44954. (doi:10.1371/journal.pone.0044954)
38. Smelov V, Bzhalava D, Arroyo Muhr LS, Eklund C, Komyakov B, Gorelov A, Dillner J, Hultin E. 2016 Detection of DNA viruses in prostate cancer. *Sci. Rep.* **6**, 25235. (doi:10.1038/srep25235)
39. Strong MJ, O'Grady T, Lin Z, Xu G, Baddoo M, Parsons C, Zhang K, Taylor CM, Flemington EK. 2013 Epstein-Barr virus and human herpesvirus 6 detection in a non-Hodgkin's diffuse large B-cell lymphoma cohort by using RNA sequencing. *J. Virol.* **87**, 13 059–13 062. (doi:10.1128/JVI.02380-13)
40. Kazemian M, Ren M, Lin JX, Liao W, Spolski R, Leonard WJ. 2015 Possible human papillomavirus 38 contamination of endometrial cancer RNA sequencing samples in the Cancer Genome Atlas Database. *J. Virol.* **89**, 8967–8973. (doi:10.1128/JVI.00822-15)
41. Arroyo Muhr LS *et al.* 2015 Human papillomavirus type 197 is commonly present in skin tumors. *Int. J. Cancer* **136**, 2546–2555. (doi:10.1002/ijc.29325)
42. Hue S *et al.* 2010 Disease-associated XMRV sequences are consistent with laboratory contamination. *Retrovirology* **7**, 111. (doi:10.1186/1742-4690-7-111)
43. Moore PS, Chang Y. 2014 The conundrum of causality in tumor virology: the cases of KSHV and MCV. *Semin. Cancer Biol.* **26**, 4–12. (doi:10.1016/j.semcancer.2013.11.001)
44. Schneider-Schaulies J. 2000 Cellular receptors for viruses: links to tropism and pathogenesis. *J. Gen. Virol.* **81**, 1413–1429. (doi:10.1099/0022-1317-81-6-1413)
45. Shigehara K *et al.* 2011 Etiologic role of human papillomavirus infection in bladder carcinoma. *Cancer* **117**, 2067–2076. (doi:10.1002/cncr.25777)
46. Yanagawa N, Wang A, Kohler D, Santos Gda C, Sykes J, Xu J, Pintilie M, Tsao M-S. 2013 Human papilloma virus genome is rare in North American non-small cell lung carcinoma patients. *Lung Cancer* **79**, 215–220. (doi:10.1016/j.lungcan.2012.11.018)
47. Olesen TB, Svahn MF, Faber MT, Duun-Henriksen AK, Junge J, Norrild B, Kjaer SK. 2014 Prevalence of human papillomavirus in endometrial cancer: a systematic review and meta-analysis. *Gynecol. Oncol.* **134**, 206–215. (doi:10.1016/j.ygyno.2014.02.040)
48. Strong MJ, O'Grady T, Lin Z, Xu G, Baddoo M, Parsons C, Zhang K, Taylor CM, Flemington EK. 2016 Correction for Strong *et al.*, Epstein-Barr virus and human herpesvirus 6 detection in a non-Hodgkin's diffuse large B-cell lymphoma cohort by using RNA sequencing. *J. Virol.* **90**, 1152. (doi:10.1128/JVI.02772-15)
49. Roberts IS, Besarani D, Mason P, Turner G, Friend PJ, Newton R. 2008 Polyoma virus infection and urothelial carcinoma of the bladder following renal

transplantation. *Br. J. Cancer* **99**, 1383–1386. (doi:10.1038/sj.bjc.6604711)

50. Egli A, Infanti L, Dumoulin A, Buser A, Samaridis J, Stebler C, Gosert R, Hirsch HH. 2009 Prevalence of polyomavirus BK and JC infection and replication in 400 healthy blood donors. *J. Infect. Dis.* **199**, 837–846. (doi:10.1086/597126)
51. Durst M, Gissmann L, Ikenberg H, zur Hausen H. 1983 A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. *Proc. Natl Acad. Sci. USA* **80**, 3812–3815. (doi:10.1073/pnas.80.12.3812)
52. Smith JS *et al.* 2002 Herpes simplex virus-2 as a human papillomavirus cofactor in the etiology of invasive cervical cancer. *J. Natl. Cancer Inst.* **94**, 1604–1613. (doi:10.1093/jnci/94.21.1604)
53. Salmons B, Gunzburg WH. 2013 Revisiting a role for a mammary tumor retrovirus in human breast cancer. *Int. J. Cancer* **133**, 1530–1535. (doi:10.1002/ijc.28210)
54. Fimereli D, Gacquer D, Fumagalli D, Salgado R, Rothe F, Larsimont D, Sotiriou C, Detours V. 2015 No significant viral transcription detected in whole breast cancer transcriptomes. *BMC Cancer* **15**, 147. (doi:10.1186/s12885-015-1176-2)
55. Cobbs CS, Harkins L, Samanta M, Gillespie GY, Bharara S, King PH, Nabors LB, Cobbs CG, Britt WJ. 2002 Human cytomegalovirus infection and expression in human malignant glioma. *Cancer Res.* **62**, 3347–3350.
56. Baumgarten P *et al.* 2014 Human cytomegalovirus infection in tumor cells of the nervous system is not detectable with standardized pathologico-virological diagnostics. *Neuro Oncol.* **16**, 1469–1477. (doi:10.1093/neuonc/nou167)
57. Cook LB, Melamed A, Niederer H, Valganon M, Laydon D, Foroni L, Taylor GP, Matsuoka M, Bangham CRM. 2014 The role of HTLV-1 clonality, proviral structure, and genomic integration site in adult T-cell leukemia/lymphoma. *Blood* **123**, 3925–3931. (doi:10.1182/blood-2014-02-553602)
58. Kassiotis G. 2014 Endogenous retroviruses and the development of cancer. *J. Immunol.* **192**, 1343–1349. (doi:10.4049/jimmunol.1302972)
59. Weitzman MD, Lilley CE, Chaurushiya MS. 2010 Genomes in conflict: maintaining genome integrity during virus infection. *Annu. Rev. Microbiol.* **64**, 61–81. (doi:10.1146/annurev.micro.112408.134016)
60. Wentzensen N, Vinokurova S, von Knebel Doeberitz M. 2004 Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Res.* **64**, 3878–3884. (doi:10.1158/0008-5472.CAN-04-0009)
61. Wendzicki JA, Moore PS, Chang Y. 2015 Large T and small T antigens of Merkel cell polyomavirus. *Curr. Opin. Virol.* **11**, 38–43. (doi:10.1016/j.coviro.2015.01.009)
62. Parfenov M *et al.* 2014 Characterization of HPV and host genome interactions in primary head and neck cancers. *Proc. Natl Acad. Sci. USA* **111**, 15 544–15 549. (doi:10.1073/pnas.1416074111)
63. Forster M *et al.* 2015 Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. *Sci. Rep.* **5**, 11534. (doi:10.1038/srep11534)
64. Lee E *et al.* 2012 Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971. (doi:10.1126/science.1222077)
65. Gillet NA *et al.* 2011 The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. *Blood* **117**, 3113–3122. (doi:10.1182/blood-2010-10-312926)
66. Miyazato P, Katsuya H, Fukuda A, Uchiyama Y, Matsuo M, Tokunaga M, Hino S, Nakao M, Satou Y. 2016 Application of targeted enrichment to next-generation sequencing of retroviruses integrated into the host human genome. *Sci. Rep.* **6**, 28324. (doi:10.1038/srep28324)
67. Bodelon C, Untereiner ME, Machiela MJ, Vinokurova S, Wentzensen N. 2016 Genomic characterization of viral integration sites in HPV-related cancers. *Int. J. Cancer* **139**, 2001–2011. (doi:10.1002/ijc.30243)
68. Kadaja M, Isok-Paas H, Laos T, Ustav E, Ustav M. 2009 Mechanism of genomic instability in cells infected with the high-risk human papillomaviruses. *PLoS Pathog.* **5**, e1000397. (doi:10.1371/journal.ppat.1000397)
69. Sung WK *et al.* 2012 Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat. Genet.* **44**, 765–769. (doi:10.1038/ng.2295)
70. Ding D *et al.* 2012 Recurrent targeted genes of hepatitis B virus in the liver cancer genomes identified by a next-generation sequencing-based approach. *PLoS Genet.* **8**, e1003065. (doi:10.1371/journal.pgen.1003065)
71. Fujimoto A *et al.* 2012 Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat. Genet.* **44**, 760–764. (doi:10.1038/ng.2291)
72. Palacios G *et al.* 2008 A new arenavirus in a cluster of fatal transplant-associated diseases. *N. Engl. J. Med.* **358**, 991–998. (doi:10.1056/NEJMoa073785)
73. Phan TG *et al.* 2016 A new protoparvovirus in human fecal samples and cutaneous T cell lymphomas (mycosis fungoides). *Virology* **496**, 299–305. (doi:10.1016/j.virol.2016.06.013)
74. Wu L *et al.* 2015 Full-length single-cell RNA-seq applied to a viral human cancer: applications to HPV expression and splicing analysis in HeLa S3 cells. *Gigascience* **4**, 51. (doi:10.1186/s13742-015-0091-4)