Strategies for Human Tumor Virus Discoveries: From Microscopic Observation to Digital Transcriptome Subtraction

Ezra D. Mirvish[1] and Masahiro Shuda[2]*

[1] *Department of Dermatology, University of Pittsburgh Medical Center, Pittsburgh, PA, USA,* [2] *Cancer Virology Program, University of Pittsburgh Cancer Institute, University of Pittsburgh, Pittsburgh, PA, USA*

Over 20% of human cancers worldwide are associated with infectious agents, including viruses, bacteria, and parasites. Various methods have been used to identify human tumor viruses, including electron microscopic observations of viral particles, immunologic screening, cDNA library screening, nucleic acid hybridization, consensus PCR, viral DNA array chip, and representational difference analysis. With the Human Genome Project, a large amount of genetic information from humans and other organisms has accumulated over the last decade. Utilizing the available genetic databases, Feng et al. (2007) developed digital transcriptome subtraction (DTS), an *in silico* method to sequentially subtract human sequences from tissue or cellular transcriptome, and discovered Merkel cell polyomavirus (MCV) from Merkel cell carcinoma. Here, we review the background and methods underlying the human tumor virus discoveries and explain how DTS was developed and used for the discovery of MCV.

Keywords: tumor virus discoveries, methods, history, digital transcriptome subtraction, Merkel cell polyomavirus

INTRODUCTION

Approximately 20% of human cancers worldwide are associated with infectious agents, including parasites, bacteria, and viruses (Parkin, 2006). In 12% of cancers, seven different viruses have been causally linked to human oncogenesis (**Table 1**): Epstein–Barr virus (EBV), hepatitis B virus (HBV), human T-lymphotropic retrovirus type 1 (HTLV-1), high-risk human papillomaviruses (HPV), hepatitis C virus (HCV), Kaposi's sarcoma herpesvirus (KSHV), and Merkel cell polyomavirus (MCV). The epidemiological and clinical information provides clues that indicate whether an infectious agent is involved in the development of cancer. Cancers that are related to immunosuppression, for example, are candidates for being caused by tumor viruses (Grulich et al., 2007). During the 20th century, various methods, ranging from the classical electron microscopic observation to the advanced molecular biology techniques, were used to identify cancer-causing viruses. Here, we will review the background and methods underlying the tumor virus discoveries during the past century as well as the newest virus discovery strategy, digital transcriptome subtraction (DTS) that we used to discover MCV.

TABLE 1 | Discovery methods of human tumor viruses.

Virus	Initial discovery method	Year	Associated human cancers	Reference
Epstein–Barr virus	Electron microscopic observation	1964	Burkitt's lymphoma, Hodgkin's lymphoma, post-transplantation lymphoproliferative disorder, and nasopharyngeal carcinoma, gastric carcinoma	Epstein et al., 1964
Hepatitis B	Immunology	1967	Hepatocellular carcinoma (HCC)	Blumberg et al., 1967
HTLV-1	Cell culture	1980	Adult T cell leukemia	Poiesz et al., 1980
Papillomavirus (high risk 16 and 18)	Nucleotide acid hybridization	1983–1984	Cervical carcinoma, head and neck carcinoma, anogenital cancer	Durst et al., 1983; Boshart et al., 1984
Hepatitis C	cDNA cloning	1989	HCC	Choo et al., 1989
Kaposi's sarcoma associated herpesvirus	Representational difference analysis	1994	Kaposi's sarcoma, primary effusion lymphoma, Castleman's disease	Chang et al., 1994
Merkel cell polyomavirus	Digital transcriptome subtraction	2008	Merkel cell carcinoma	Feng et al., 2008

METHODS USED FOR THE HUMAN TUMOR VIRUS DISCOVERIES

Cell Culture and Electron Microscopy

By the early 1960s, many viruses were known to cause tumors in animals (Javier and Butel, 2008), but none had been identified in human cancers. Burkitt (1958) first described his eponymous lymphoma in 1958 as a sarcoma involving the jaws of African children and hypothesized that the cancer's geographic distribution might implicate some infectious agent as an etiologic factor in disease development. When the first Burkitt's lymphoma-derived cell line was established *in vitro* and examined by electron microscopy, obvious herpesvirus morphology was readily observed (Epstein et al., 1964). Biological examinations demonstrated that this was a new human herpesvirus, which became known as the EBV after the cell line in which it was discovered (Epstein et al., 1965).

The human T-lymphotropic virus type 1 (HTLV-1), the first known human retrovirus, was also identified in cell culture (Poiesz et al., 1980). Retrovirus particles with type C morphology were observed in thin-section electron micrographs of fixed, pelleted cellular materials from two T-cell lymphoblastoid cell lines (HUT102 and CTCL-3) and peripheral blood lymphocytes from a patient with a cutaneous T-cell lymphoma. Mature particles were 100–110 nm in diameter, and consisted of an electron-dense core separated by an electron-lucent region from an outer membrane. Similar type C particles were identified in MT-1 cells derived from Japanese patients with adult T cell leukemia (ATL), which is now known to be caused by HTLV-1 (Hinuma et al., 1981). Indirect immunofluorescence staining was also used in the studies of ATL. A unique antigen was identified in MT-1 cells derived from ATL that was not found in other human lymphoid cell lines. All 44 examined ATL patients had positive serological reactions for this antigen.

In general, it is very rare that human cancer arises directly from the acute consequences of viral infection. Tumorigenesis usually occurs after a latency period of 15–40 years. However, a special exception is the X-chromosome-linked lymphoproliferative disorder (XLP) found in some EBV-infected patients. Mutated immune response genes leave susceptible males unable to respond to interferon (IFN) signaling, which can in turn lead to the development of an acute lymphoproliferative disease as a consequence of EBV replication (Kutok and Wang, 2006). Most viruses do not replicate efficiently in common laboratory cell lines due to the innate immune system, which controls IFN signaling. By targeting molecules in the IFN signaling pathway through knockout technology or RNA interference, it is possible to produce cells that are permissive for viral replication. Cell culture models with a defective IFN signaling pathway are useful for novel virus identification and characterization, especially in monitoring for cytopathic effects (CPE). The robust replication of a virus often gives rise to CPE, which involves morphological changes in the host cell, such as cell rounding, disorientation, swelling or shrinking, detachment from the culture surface, and cell death, and may indicate that the cells being studied are a good model in which to amplify the genome of the putative virus. This strategy was successfully used in acute infection with African green monkey kidney cells (Vero, a IFN-deficient cell) to identify previously uncharacterized human coronavirus (HCoV-NL63; van der Hoek et al., 2004). Humanized mice have also been developed to study viral infection in a systemic setting (Lassnig et al., 2005). In this study, viral replication was markedly enhanced in the offspring of mice with human viral receptor genes that were crossed with IFN unresponsive Stat1$^{-/-}$ mice. Such cell and animal models, which take advantage of advanced techniques in molecular biology and genetic engineering, have the potential to provide novel insights into the biology of otherwise unculturable infectious agents and to identify the tumorigenic potential of a virus.

Immunologic Methods to Detect Viral Antigen

Tissue culture, animal inoculation, and other virological methods were all applied unsuccessfully in early attempts to identify HBV (Blumberg, 1977). Blumberg et al. (1965) performed a systematic study of the sera collected from leukemia and hemophilia patients who had experienced frequent blood transfusions in order to detect precipitating iso-antibodies against unknown

antigens found in donor serum. The 'Australia antigen' was identified in the serum of an Australian Aborigine as being a target of precipitating antibodies generated in the serum of transfusion patients (Blumberg et al., 1965). It was found to be rare in normal Americans, but was common in normal people from Africa, Asia, and Oceania. These data led some to hypothesize a relationship between leukemia, the unknown agent, and genetic polymorphism(s) in Australia. Subsequently, the Australian antigen was revealed to be an antigen shared by multiple hepatitis patients (Blumberg et al., 1967; Prince, 1968). Virus-like particles containing 'Australian antigen' on their surface were eventually found by electronic microscopy in the blood of three patients with 'serum hepatitis' (Dane et al., 1970; Jokelainen et al., 1970). These findings were the first indication that the disease now known as hepatitis B is caused by a virus (HBV).

The discovery of HCV was an extraordinary achievement, because isolation of the viral genome did not rely on previously visualized viral particles, growth of the virus *in vitro*, or development of an immunological assay. Patients who had transfusion-associated hepatitis were described, despite being negative for hepatitis A and B antigens, which suggested the existence of another, unidentified infectious agent (Feinstone et al., 1975). Although several immunologic and serologic assays were developed, none was a specific, reliable serologic marker for this unknown hepatitis antigen. However, administration of the putative infectious reagent induced non-A, non-B hepatitis in chimpanzees, which confirmed that an infectious agent was likely responsible for the disease (Tabor et al., 1978). A decade after the development of the chimpanzee disease model, Alter and Houghton (2000), utilizing novel methodologies in molecular biology, conducted a blind immunoscreen using a recombinant phage expression library. In searching for an unknown virus in these non-A, non-B hepatitis patients, a randomly primed cDNA library was constructed from the plasma of chimpanzees that had been injected with serum from non-A, non-B hepatitis patients. cDNA was inserted into the bacteriophage λgt11 and expressed in *Escherichia coli*. The expressed proteins were then screened against serum from an infected non-A, non-B hepatitis patient. Among millions of the clones screened, one clone encoded an antigen that showed specific seroreactivity in non-A, non-B hepatitis patients. The putative virus was shown to contain a positive-strand RNA molecule of about 10,000 nucleotides that belongs to the *Flaviviridae* family (Choo et al., 1989). Novel virus discovery by screening recombinant expression libraries from infected tissues against patient sera represented a significant breakthrough. However, library screenings are also laborious, as hundreds of millions of bacterial cDNA clones had to be screened (Alter and Houghton, 2000).

Cross-hybridization to Identify Related Viruses

The *Papillomaviridae* comprise a super-family with several 100 members, most of which were identified in birds and mammals (de Villiers et al., 2004). Papillomaviruses are highly host-tropic and tissue-specific, causing small benign tumors known

as papillomas or warts. HPV particles were observed by electron microscopy in cervical dysplasia (Della Torre et al., 1978; Meisels et al., 1981), whereas in cervical intraepithelial lesions and cervical cancer specimens, virus particles are not generally observed due to the viral integration into host genome. High-risk HPV 16 was isolated from an invasive cervical cancer by cross-hybridization with HPV type 11 DNA under non-stringent conditions (Durst et al., 1983). Using virus-specific probes, HPV 16 presence was confirmed in precursor lesion of malignant tumors (Ikenberg et al., 1983). High-risk HPV18 was subsequently identified by low stringency hybridization with HPV 8, 9, 10, and 11-related sequences (Boshart et al., 1984). zur Hausen's (1999) research group identified both HPV16 and 18, which together are responsible for approximately 70% of invasive cervical cancers (de Sanjose et al., 2010). His group further demonstrated that HPV DNA is integrated into the host genome in cervical cancer cell lines and that viral E6 and E7 genes are expressed in cervical cancer tumors (Schwarz et al., 1985).

Nucleic acid hybridization with a single or few probes is restricted by its ability to detect only a limited number of candidate viruses. To obviate this problem, Virochip was invented to detect a broad spectrum of viruses in a single analysis (Wang et al., 2002). The Virochip was successfully used to identify a previously uncharacterized coronavirus isolated from a SARS patient during the outbreak of the severe acute respiratory syndrome (SARS; Rota et al., 2003). For the tumor virus identification, the Virochip was used to screen RNA samples from prostate tumors. A novel gammaretrovirus, Xenotropic murine leukemia virus-related virus (XMRV), which is closely related to the xenotropic murine leukemia viruses (MuLVs), was found in patients with prostate cancers (Urisman et al., 2006), and chronic fatigue syndrome (Lombardi et al., 2009). XMRV is detected in malignant prostate epithelium by using quantitative PCR assay and immunohistochemistry with an anti-XMRV specific antiserum, raising the possibility of the virus may indirectly support tumorigenesis (Schlaberg et al., 2009). However, subsequent reports disprove the etiologic connection between XMRV infection and these two diseases (Fischer et al., 2008; Hohn et al., 2009; Lee et al., 2012). XMRV was found to have arisen during the passage of a human prostate cancer in mice, as a result of recombination between two endogenous MuLVs from the mouse cells. The resulting XMRV infected human prostate cancer cells due to its xenotropic host range (Paprotka et al., 2011).

With the widespread usage of PCR and availability of extensive viral sequences, it is possible to combine short-primer hybridization and the power of PCR to amplify potential viral sequences. This method is commonly known as consensus sequence-base PCR or degenerate PCR. Primer sequences are designed based on a conserved region of a viral genome. Human hepatitis G virus was identified by amplifying a segment of a putative helicase gene with consensus primers from hepatitis A, B, and C virus (Simons et al., 1995a). As with genomic nucleic acid hybridization, consensus PCR/degenerate PCR relies on known viral sequences, and as such has limitations in identifying novel viruses. Moreover, false-positives due to non-specific

amplification and PCR contamination may introduce costly and fruitless downstream analyses. Thus, careful optimization of reaction and primer annealing temperatures and the use of nested or semi-nested strategies, as well as the use of known viruses as positive controls, are necessary.

Differential Display Strategy in Viral Pathogen Discovery

The recent molecular techniques for virus discovery are solely based on nucleic acids—rather than conventional viral particle morphology, *in vitro* cell culture, or serological assays—but must be used in conjunction with strong epidemiologic evidence (Gao and Moore, 1996). These methods identify discrepancies in nucleotide sequence between disease-associated tissue and normal tissue by subtractive strategies. Representational difference analysis (RDA) combines physical subtractive hybridization with gene amplification to detect differences between tumor genomic DNA and that of normal cells (Lisitsyn et al., 1993). Epidemiologic evidence indicated that Kaposi's sarcoma may be caused by a pathogen other than human immunodeficiency virus (HIV; Beral et al., 1990, 1992). By using RDA, novel viral fragments were identified in AIDS patients and further characterized as human Kaposi's sarcoma associated herpesvirus (KSHV; Chang et al., 1994; Moore et al., 1996). RDA has also been used to successfully identify hepatitis G virus (Simons et al., 1995b) and Torque Teno virus (TTV) in non-A–E hepatitis patients (Nishizawa et al., 1997). While RDA was used to successfully identify these viruses, this technique relies heavily on the non-diseased tissue controls and is not quantitative.

In Silico Subtraction Approach Using Nucleic Acid Database

With initiation of the Human Genome Project, the nucleic acid databases for humans and many other organisms rapidly expanded. Taking advantage of this new wealth of data, Weber et al. (2002) developed an *in silico* subtraction approach capable of identifying non-human sequences in 'human' expressed-sequence tag (EST) libraries. In this computational subtraction technique, sequences with high similarity to the human genome are subtracted from cDNA libraries, such that the remaining data are enriched for sequences of non-human origin. Weber et al. (2002) directly identified multiple human tumor viruses erroneously deposited as human ESTs in GenBank, including EBV, HBV, HPV type 16 and 18, HCV, and KSHV. Viral sequences were also found in data deposited from hepatocellular carcinoma (HCC) and cervical carcinoma cell lines: HBV sequences represented 0.1% of deposited HCC data, while HPV18 sequences accounted for 0.03% of the cervical carcinoma library. The same group also generated a cDNA library from a patient with post-transplant lymphoproliferative disorder (PTLD; Xu et al., 2003). A total of 27,840 cDNA sequences were generated and filtered by computational subtraction against known human sequences, leaving 32 non-matching sequences. Among these sequences, 10 (accounting for 0.03% of total sequences analyzed) were from EBV viruses. Taken together, these results suggest that

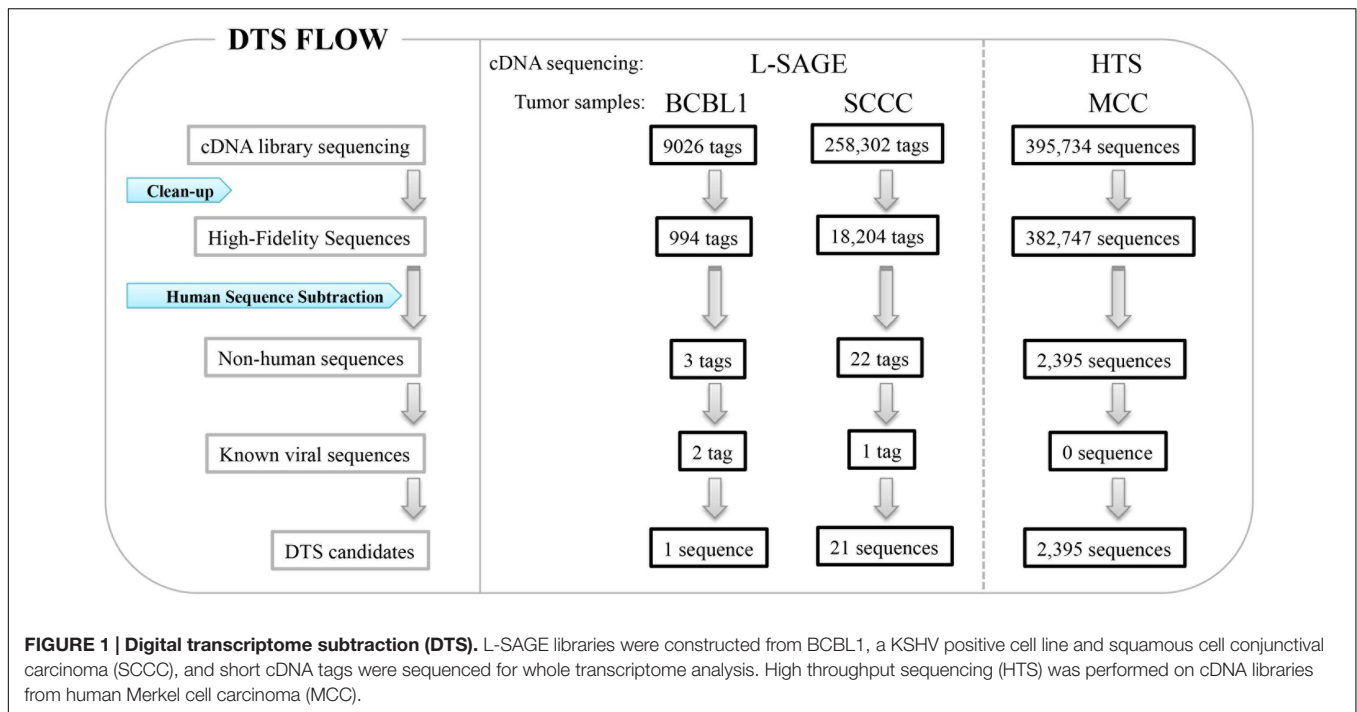it is possible to identify viral sequences by sampling fewer than 10,000 sequences.

DEVELOPMENT OF DIGITAL TRANSCRIPTOME SUBTRACTION

Feng et al. (2007) developed DTS as a quantitative means to search for viral infection in human cancers (**Figure 1**) (Feng et al., 2007). High quality sequences are extracted with a high stringency score (Phred > = 20) to avoid sequencing errors, and all sequences artificially used in the cDNA library preparation, such as sequencing primers, are removed. All sequences matching human databases are computationally subtracted to leave 'non-human' candidates. There are approximately 200,000 mRNA transcripts in a cell (Bishop et al., 1974; Velculescu et al., 1999), which correspond to five transcripts per million (TPM) for a single transcript per cell. In theory, one can analyze the whole transcriptome to search for viral etiology in human cancers. Thus, if no viral transcript is found, the possibility of a virus expressing transcript in a given tumor can be excluded above a certain threshold level.

Long serial analysis of gene expression (L-SAGE) was used to sample the cellular transcriptome for the presence of viral transcripts. L-SAGE quantitatively concatenates ~21 bp cDNA tags from the 3′ end of mRNA transcripts, allowing for the measurement of gene expression via high-throughput sequencing (**Figure 1**) (Saha et al., 2002). When they performed a pilot analysis of DTS on 9,026 SAGE tags from an expression library of BCBL-1 cells infected with KSHV, only three candidate sequences were identified as being of non-human origin: two of these sequences belonged to KSHV transcripts and the third belonged to an unannotated human expression sequence tag, as confirmed by additional experiments. Overall, 0.24% of transcripts from this cell line were of viral origin. DTS was then applied to 258,302 SAGE tags from squamous cell conjunctival carcinoma (SCCC), a cancer strongly associated with immunosuppression (**Figure 1**). Only 21 of these sequences did not align to human databases, excluding one KSHV tag as an internal control. All 21 candidates were ruled out as viral sequences by further experimental examination (Feng et al., 2007). This analysis shows that it is unlikely that distinguishable viral transcripts are present in SCCC at 20 TPM or higher, which is the equivalent of approximately four transcripts per cell.

MERKEL CELL POLYOMAVIRUS DISCOVERY BY DTS

We further performed high-throughput sequencing on cDNA libraries constructed from Merkel cell carcinoma (MCC), a malignant skin tumor frequently found among immunosuppressed, transplant, and AIDS patients (Feng et al., 2008). A total of 395,734 sequences were subtracted through DTS analysis, leaving 2,395 'non-human' sequences (**Figure 1**). One transcript was similar to but distinct from African green monkey lymphotropic polyomavirus (LPV) and human BK

FIGURE 1 | Digital transcriptome subtraction (DTS). L-SAGE libraries were constructed from BCBL1, a KSHV positive cell line and squamous cell conjunctival carcinoma (SCCC), and short cDNA tags were sequenced for whole transcriptome analysis. High throughput sequencing (HTS) was performed on cDNA libraries from human Merkel cell carcinoma (MCC).

polyomavirus T antigen sequences, defining a new polyomavirus, MCV. Subsequently, we sequenced the complete closed circular genome of MCV (5,387 bp), which encodes a T antigen locus, late gene cassette for viral capsids, and replication origin sequences. MCV is the first human member of the murine polyomavirus subgroup, and shares highest homology with LPV and a new polyomavirus recently isolated from orangutans (Groenewoud et al., 2010). Overall, DTS is a simple screening method to discover novel viral nucleic acids in a quantitative manner with high throughput sequencing. We found MCV transcripts in MCC at a level of 10 TPM or approximately five transcripts per cell and can exclude distinguishable viral transcripts in SCCC at 20 TPM or higher, which is the equivalent of approximately four transcripts per cell. This provides, for the first time, quantitative evidence against some classes of viral etiology when no viral transcripts are found, thereby reducing the uncertainty involved in new pathogen discovery.

MCV IS THE FIRST HUMAN POLYOMAVIRUS CAUSING MCC

Virus discovery is only the first step to determining the etiology behind a disease. The detection of nucleic acid is not sufficient to prove causality. Our initial discovery of MCV demonstrated that the MCV genome is clonally integrated in ∼80% of MCC tumor cell DNA (Feng et al., 2008). Similar to other polyomaviruses, the MCV T antigen locus encodes for three major overlapping transcripts: large T (LT), small T (sT), and 57 kT antigens (Shuda et al., 2008). With respect to other well-studied polyomavirus T antigens, MCV LT has conserved N-terminus DnaJ and Rb binding domains, as well as C-terminus origin binding (OBD)

and helicase domains, which are required for viral replication. MCV sT has a conserved PP2A binding domain (Shuda et al., 2008). Our viral sequence analysis in MCC tumors revealed that integrated MCV LT antigens are C-terminally truncated by tumor-specific mutations, which results in the loss of helicase function, whereas MCV sT remains intact (Shuda et al., 2008). This indicates that MCV is replication-defective in tumors and not a passenger virus that secondarily infects MCC tumors. We developed monoclonal antibodies, which specifically recognize MCV LT or MCV sT and carried out immunohistochemistry using these antibodies in MCC tissue microarrays. We identified that approximately ∼70% of MCC tumors express the truncated LT and sT proteins consistent with their role as directly transforming oncoproteins in MCC cancers (Shuda et al., 2009, 2011).

We identified three MCC cell lines harboring clonally integrated MCV, expressing both truncated LT and sT proteins (Houben et al., 2010; Guastafierro et al., 2013). If MCV is a direct viral carcinogen that causes most cases of MCC, we anticipate that its viral oncogene expression is necessary for the growth of MCC cells. The panT antigen knockdown (both LT and sT) by shRNA results in G0/G1 cell cycle arrest in MCV-positive MCC tumor cells and non-apoptotic cell death as measured by caspase cleavage (Houben et al., 2010). To further assess the importance of individual T antigens in MCC, we exploited an shRNA that knocks down sT antigen alone. Knockdown of MCV sT expression abolished tumor cell growth capacity (Shuda et al., 2011). On the other hand, Houben et al. (2015) demonstrated that shRNA knockdown of MCV LT alone also inhibits MCV-positive MCC cell proliferation. These results indicate that both MCV sT and LT play critical roles in the maintenance of a tumorigenic phenotype in MCC cells.

Our *in vitro* transformation studies demonstrated that MCV sT expression alone is sufficient to transform immortalized rodent fibroblast cell lines in a soft agar assay, while MCV LT expression was not (Shuda et al., 2011). MCV LT expression, however, promoted cell proliferation in human fibroblasts through its Rb binding domain (Cheng et al., 2013; Richards et al., 2015). Further, multiple groups demonstrated the oncogenic activities of MCV T antigen in transgenic mouse models. The MCC-derived MCV T antigen expression in mice, driven by the keratin 14 promoter induces papillomatosis (Spurgeon et al., 2015). MCV sT expression using a keratin 5 promoter also induces hyperproliferative lesions (Verhaegen et al., 2015). We have also demonstrated that MCV sT expression by the ubiquitin promoter in p53 null mice induces high-grade tumors in spleen and liver tissues (Shuda et al., 2015). These results are consistent with MCV being a human tumor virus encoding T antigen viral oncoproteins.

Among 13 known human polyomaviruses, eleven have been identified over the past 9 years by molecular technologies and next generation DNA sequencing (DeCaprio and Garcea, 2013; Rinaldo and Hirsch, 2013; Mishra et al., 2014). While JC virus and SV40 are suspected to play a role in human cancers, the viral copy numbers in the associated tumors are very low (less than one copy per cell; Gordon et al., 2002; Rollison et al., 2005). Recently, to survey for the presence of human polyomaviruses in cancer, Toptan et al. (2016) developed the immunohistochemistry-based pan-human polyomavirus screening that can detect T antigen protein expression of all known human polyomaviruses including SV40. However, the study did not find any evidence of the polyomavirus T antigen expression in the 1,184 various tumor cases except the MCV T antigen expression in MCC tumors (Toptan et al., 2016). Thus MCV is the only polyomavirus strongly linked to a human cancer at present.

CONCLUDING REMARKS

Identification of 'non-human' sequences from a whole transcriptome does not necessarily indicate the presence of other genomes in human disease, as contamination from sample preparation may introduce some 'non-human' sequences. The existence of such sequences is only meaningful in the context of a strong association with human disease. Koch's postulates and their revisions provide a valid standard for judging disease causation (Fredricks and Relman, 1996). In spite of acute infection, tumorigenesis usually occurs after a long latency period. To move from association to causation for putatively carcinogenic viral agents, zur Hausen (1999) has also proposed new criteria for defining a causal role of an infection in human cancers: (1) strong epidemiological plausibility and evidence that a viral infection represents a risk factor for the development of a specific tumor; (2) the consistent presence and persistence of the 'non-human' sequences in cells of the tumor; (3) the stimulation of cell proliferation by the 'non-human' genome (or part of it) in corresponding cell culture systems; (4) the demonstration that the 'non-human' genome induces proliferation and the malignant phenotype of the tumor. To

address whether MCV is causally relevant to MCC, additional evidence is needed to elucidate tumorigenic potential of the MCV T antigen.

In addition to the cancers known to be caused by human viruses, there are many tumors in which a viral etiology is suspected, including some hematopoietic malignancies, breast cancer, colorectal cancer, basal cell carcinoma of the skin, and lung cancers in non-smokers (Zur Hausen, 2009). Undoubtedly, DTS analysis on whole transcriptomes of human tumors, in conjunction with other approaches of the sort described above, will help us to identify new clues or eliminate potential viral etiologies. Despite the power of DTS in searching for viral etiologies, there are some caveats that should be kept in mind: (1) DTS relies on high database quality and unique features of non-human sequences. Viral sequences that have been accidentally deposited into human databases or viral sequences that are indistinguishable from human transcripts will cause analyses to fail. A specific subgroup of human endogenous retrovirus (HERV-K) is suspected to be associated with multiple cancers (Moyes et al., 2007). DTS against the human genome would subtract out endogenous retroviral sequences as human and thus fail to detect pathogenic sequences. (2) Viral transcripts that lack polyadenylation sites, such as found in *Flaviviridae* and *Reoviridae,* cannot be detected. (3) Viruses causing tumors without gene expression, such by insertional mutagenesis, will be missed. Thus, extra efforts should be made to overcome the caveats of DTS analysis by combining with other methods for virus discovery.

The invention of high throughput sequencing technology developed massive transcriptome database generated from various human cancers in the Cancer Genome Atlas (TCGA), which could be a direct resource for DTS. Several studies interrogated transcriptome sequencing data from nearly 4,000 different tumors generated within the TCGA consortium for the presence of viral sequences by using the strategy similar to DTS (Khoury et al., 2013; Tang et al., 2013). One of these studies revealed a rare bladder cancer with BK polyomavirus integration that expresses full length LT antigen (Tang et al., 2013) and also identified a novel enterovirus in colon adenocarcinoma which is unlikely to be oncogenic. The use of DTS to the expanding transcriptome database in the TCGA may allow us to discover the eighth human tumor virus in the near future.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

REFERENCES

Alter, H. J., and Houghton, M. (2000). Clinical medical research award. Hepatitis C virus and eliminating post-transfusion hepatitis. *Nat. Med.* 6, 1082–1086. doi: 10.1038/80394

Beral, V., Bull, D., Darby, S., Weller, I., Carne, C., Beecham, M., et al. (1992). Risk of Kaposi's sarcoma and sexual practices associated with faecal contact in homosexual or bisexual men with AIDS. *Lancet* 339, 632–635. doi: 10.1016/0140-6736(92)90793-3

Beral, V., Peterman, T. A., Berkelman, R. L., and Jaffe, H. W. (1990). Kaposi's sarcoma among persons with AIDS: a sexually transmitted infection? *Lancet* 335, 123–128. doi: 10.1016/0140-6736(90)90001-L

Bishop, J. O., Morton, J. G., Rosbash, M., and Richardson, M. (1974). Three abundance classes in HeLa cell messenger RNA. *Nature* 250, 199–204. doi: 10.1038/250199a0

Blumberg, B. S. (1977). Australia antigen and the biology of hepatitis B. *Science* 197, 17–25. doi: 10.1126/science.325649

Blumberg, B. S., Alter, H. J., and Visnich, S. (1965). A "New" antigen in leukemia sera. *JAMA* 191, 541–546. doi: 10.1001/jama.1965.03080070025007

Blumberg, B. S., Gerstley, B. J., Hungerford, D. A., London, W. T., and Sutnick, A. I. (1967). A serum antigen (Australia antigen) in Down's syndrome, leukemia, and hepatitis. *Ann. Int. Med.* 66, 924–931. doi: 10.7326/0003-4819-66-5-924

Boshart, M., Gissmann, L., Ikenberg, H., Kleinheinz, A., Scheurlen, W., and Zur Hausen, H. (1984). A new type of papillomavirus DNA, its presence in genital cancer biopsies and in cell lines derived from cervical cancer. *EMBO J.* 3, 1151–1157.

Burkitt, D. (1958). A sarcoma involving the jaws in African children. *Br. J. Surg.* 46, 218–223. doi: 10.1002/bjs.18004619704

Chang, Y., Cesarman, E., Pessin, M. S., Lee, F., Culpepper, J., Knowles, D. M., et al. (1994). Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* 266, 1865–1869. doi: 10.1126/science.7997879

Cheng, J., Rozenblatt-Rosen, O., Paulson, K. G., Nghiem, P., and Decaprio, J. A. (2013). Merkel cell polyomavirus large T antigen has growth-promoting and inhibitory activities. *J. Virol.* 87, 6118–6126. doi: 10.1128/JVI.00385-13

Choo, Q. L., Kuo, G., Weiner, A. J., Overby, L. R., Bradley, D. W., and 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

Dane, D. S., Cameron, C. H., and Briggs, M. (1970). Virus-like particles in serum of patients with Australia-antigen-associated hepatitis. *Lancet* 1, 695–698. doi: 10.1016/S0140-6736(70)90926-8

de Sanjose, S., Quint, W. G., Alemany, L., Geraets, D. T., Klaustermeier, J. E., Lloveras, B., et al. (2010). Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. *Lancet Oncol.* 11, 1048–1056. doi: 10.1016/S1470-2045(10)70230-8

de Villiers, E. M., Fauquet, C., Broker, T. R., Bernard, H. U., and Zur Hausen, H. (2004). Classification of papillomaviruses. *Virology* 324, 17–27. doi: 10.1016/j.virol.2004.03.033

DeCaprio, J. A., and Garcea, R. L. (2013). A cornucopia of human polyomaviruses. *Nat. Rev. Microbiol.* 11, 264–276. doi: 10.1038/nrmicro2992

Della Torre, G., Pilotti, S., De Palo, G., and Rilke, F. (1978). Viral particles in cervical condylomatous lesions. *Tumori* 64, 549–553.

Durst, M., Gissmann, L., Ikenberg, H., and 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. U.S.A.* 80, 3812–3815. doi: 10.1073/pnas.80.12.3812

Epstein, M. A., Achong, B. G., and Barr, Y. M. (1964). Virus particles in cultured lymphoblasts from Burkitt's lymphoma. *Lancet* 1, 702–703. doi: 10.1016/S0140-6736(64)91524-7

Epstein, M. A., Henle, G., Achong, B. G., and Barr, Y. M. (1965). morphological and biological studies on a virus in cultured lymphoblasts from Burkitt's lymphoma. *J. Exp. Med.* 121, 761–770. doi: 10.1084/jem.121.5.761

Feinstone, S. M., Kapikian, A. Z., Purcell, R. H., Alter, H. J., and Holland, P. V. (1975). Transfusion-associated hepatitis not due to viral hepatitis type A or B. *N. Engl. J. Med.* 292, 767–770. doi: 10.1056/NEJM197504102921502

Feng, H., Shuda, M., Chang, Y., and Moore, P. S. (2008). Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* 319, 1096–1100. doi: 10.1126/science.1152586

Feng, H., Taylor, J. L., Benos, P. V., Newton, R., Waddell, K., Lucas, S. B., et al. (2007). Human transcriptome subtraction by using short sequence tags to search for tumor viruses in conjunctival carcinoma. *J. Virol.* 81, 11332–11340. doi: 10.1128/JVI.00875-07

Fischer, N., Hellwinkel, O., Schulz, C., Chun, F. K., Huland, H., Aepfelbacher, M., et al. (2008). Prevalence of human gammaretrovirus XMRV in sporadic prostate cancer. *J. Clin. Virol.* 43, 277–283. doi: 10.1016/j.jcv.2008.04.016

Fredricks, D. N., and Relman, D. A. (1996). Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin. Microbiol. Rev.* 9, 18–33.

Gao, S. J., and Moore, P. S. (1996). Molecular approaches to the identification of unculturable infectious agents. *Emerg. Infect. Dis.* 2, 159–167. doi: 10.3201/eid0203.960301

Gordon, G. J., Chen, C. J., Jaklitsch, M. T., Richards, W. G., Sugarbaker, D. J., and Bueno, R. (2002). Detection and quantification of SV40 large T-antigen DNA in mesothelioma tissues and cell lines. *Oncol. Rep.* 9, 631–634.

Groenewoud, M. J., Fagrouch, Z., Van Gessel, S., Niphuis, H., Bulavaite, A., Warren, K. S., et al. (2010). Characterization of novel polyomaviruses from Bornean and Sumatran orang-utans. *J. Gen. Virol.* 91, 653–658. doi: 10.1099/vir.0.017673-0

Grulich, A. E., Van Leeuwen, M. T., Falster, M. O., and Vajdic, C. M. (2007). Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis. *Lancet* 370, 59–67. doi: 10.1016/S0140-6736(07)61050-2

Guastafierro, A., Feng, H., Thant, M., Kirkwood, J. M., Chang, Y., Moore, P. S., et al. (2013). Characterization of an early passage merkel cell polyomavirus-positive merkel cell carcinoma cell line, MS-1, and its growth in NOD scid gamma mice. *J. Virol. Methods* 187, 6–14. doi: 10.1016/j.jviromet.2012.10.001

Hinuma, Y., Nagata, K., Hanaoka, M., Nakai, M., Matsumoto, T., Kinoshita, K. I., et al. (1981). Adult T-cell leukemia: antigen in an ATL cell line and detection of antibodies to the antigen in human sera. *Proc. Natl. Acad. Sci. U.S.A.* 78, 6476–6480. doi: 10.1073/pnas.78.10.6476

Hohn, O., Krause, H., Barbarotto, P., Niederstadt, L., Beimforde, N., Denner, J., et al. (2009). Lack of evidence for xenotropic murine leukemia virus-related virus(XMRV) in German prostate cancer patients. *Retrovirology* 6:92. doi: 10.1186/1742-4690-6-92

Houben, R., Angermeyer, S., Haferkamp, S., Aue, A., Goebeler, M., Schrama, D., et al. (2015). Characterization of functional domains in the Merkel cell polyoma virus Large T antigen. *Int. J. Cancer* 136, E290–E300. doi: 10.1002/ijc.29200

Houben, R., Shuda, M., Weinkam, R., Schrama, D., Feng, H., Chang, Y., et al. (2010). Merkel cell polyomavirus-infected Merkel cell carcinoma cells require expression of viral T antigens. *J. Virol.* 84, 7064–7072. doi: 10.1128/JVI.02400-09

Ikenberg, H., Gissmann, L., Gross, G., Grussendorf-Conen, E. I., and Zur Hausen, H. (1983). Human papillomavirus type-16-related DNA in genital Bowen's disease and in Bowenoid papulosis. *Int. J. Cancer* 32, 563–565. doi: 10.1002/ijc.2910320507

Javier, R. T., and Butel, J. S. (2008). The history of tumor virology. *Cancer Res.* 68, 7693–7706. doi: 10.1158/0008-5472.CAN-08-3301

Jokelainen, P. T., Krohn, K., Prince, A. M., and Finlayson, N. D. (1970). Electrn microscopic observations on virus-like particles associated with SH antigen. *J. Virol.* 6, 685–689.

Khoury, J. D., Tannir, N. M., Williams, M. D., Chen, Y., Yao, H., Zhang, J., 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

Kutok, J. L., and Wang, F. (2006). Spectrum of Epstein-Barr virus-associated diseases. *Annu. Rev. Pathol.* 1, 375–404. doi: 10.1146/annurev.pathol.1.110304.100209

Lassnig, C., Kolb, A., Strobl, B., Enjuanes, L., and Muller, M. (2005). Studying human pathogens in animal models: fine tuning the humanized mouse. *Transgenic Res.* 14, 803–806. doi: 10.1007/s11248-005-1676-y

Lee, D., Das Gupta, J., Gaughan, C., Steffen, I., Tang, N., Luk, K. C., 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

Lisitsyn, N., Lisitsyn, N., and Wigler, M. (1993). Cloning the differences between two complex genomes. *Science* 259, 946–951. doi: 10.1126/science.8438152

Lombardi, V. C., Ruscetti, F. W., Das Gupta, J., Pfost, M. A., Hagen, K. S., Peterson, D. L., et al. (2009). Detection of an infectious retrovirus, XMRV, in blood cells of patients with chronic fatigue syndrome. *Science* 326, 585–589. doi: 10.1126/science.1179052

Meisels, A., Roy, M., Fortier, M., Morin, C., Casas-Cordero, M., Shah, K. V., et al. (1981). Human papillomavirus infection of the cervix: the atypical condyloma. *Acta Cytol.* 25, 7–16.

Mishra, N., Pereira, M., Rhodes, R. H., An, P., Pipas, J. M., Jain, K., et al. (2014). Identification of a novel polyomavirus in a pancreatic transplant recipient with retinal blindness and vasculitic myopathy. *J. Infect. Dis.* 210, 1595–1599. doi: 10.1093/infdis/jiu250

Moore, P. S., Gao, S. J., Dominguez, G., Cesarman, E., Lungu, O., Knowles, D. M., et al. (1996). Primary characterization of a herpesvirus agent associated with Kaposi's sarcomae. *J. Virol.* 70, 549–558.

Moyes, D., Griffiths, D. J., and Venables, P. J. (2007). Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. *Trends Genet.* 23, 326–333. doi: 10.1016/j.tig.2007.05.004

Nishizawa, T., Okamoto, H., Konishi, K., Yoshizawa, H., Miyakawa, Y., and Mayumi, M. (1997). A novel DNA virus (TTV) associated with elevated transaminase levels in posttransfusion hepatitis of unknown etiology. *Biochem. Biophys. Res. Commun.* 241, 92–97. doi: 10.1006/bbrc.1997.7765

Paprotka, T., Delviks-Frankenberry, K. A., Cingoz, O., Martinez, A., Kung, H. J., Tepper, C. G., et al. (2011). Recombinant origin of the retrovirus XMRV. *Science* 333, 97–101. doi: 10.1126/science.1205292

Parkin, D. M. (2006). The global health burden of infection-associated cancers in the year 2002. *Int. J. Cancer* 118, 3030–3044. doi: 10.1002/ijc.21731

Poiesz, B. J., Ruscetti, F. W., Gazdar, A. F., Bunn, P. A., Minna, J. D., and Gallo, R. C. (1980). Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. *Proc. Natl. Acad. Sci. U.S.A.* 77, 7415–7419. doi: 10.1073/pnas.77.12.7415

Prince, A. M. (1968). Relation of Australia and SH antigens. *Lancet* 2, 462–463. doi: 10.1016/S0140-6736(68)90512-6

Richards, K. F., Guastafierro, A., Shuda, M., Toptan, T., Moore, P. S., and Chang, Y. (2015). Merkel cell polyomavirus T antigens promote cell proliferation and inflammatory cytokine gene expression. *J. Gen. Virol.* doi: 10.1099/jgv.0.000287 [Epub ahead of print].

Rinaldo, C. H., and Hirsch, H. H. (2013). The human polyomaviruses: from orphans and mutants to patchwork family. *APMIS* 121, 681–684. doi: 10.1111/apm.12125

Rollison, D. E., Utaipat, U., Ryschkewitsch, C., Hou, J., Goldthwaite, P., Daniel, R., et al. (2005). Investigation of human brain tumors for the presence of polyomavirus genome sequences by two independent laboratories. *Int. J. Cancer* 113, 769–774. doi: 10.1002/ijc.20641

Rota, P. A., Oberste, M. S., Monroe, S. S., Nix, W. A., Campagnoli, R., Icenogle, J. P., et al. (2003). Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 300, 1394–1399. doi: 10.1126/science.1085952

Saha, S., Sparks, A. B., Rago, C., Akmaev, V., Wang, C. J., Vogelstein, B., et al. (2002). Using the transcriptome to annotate the genome. *Nat. Biotechnol.* 20, 508–512. doi: 10.1038/nbt0502-508

Schlaberg, R., Choe, D. J., Brown, K. R., Thaker, H. M., and Singh, I. R. (2009). XMRV is present in malignant prostatic epithelium and is associated with prostate cancer, especially high-grade tumors. *Proc. Natl. Acad. Sci. U.S.A.* 106, 16351–16356. doi: 10.1073/pnas.0906922106

Schwarz, E., Freese, U. K., Gissmann, L., Mayer, W., Roggenbuck, B., Stremlau, A., et al. (1985). Structure and transcription of human papillomavirus sequences in cervical carcinoma cells. *Nature* 314, 111–114. doi: 10.1038/314111a0

Shuda, M., Arora, R., Kwun, H. J., Feng, H., Sarid, R., Fernandez-Figueras, M. T., et al. (2009). Human merkel cell polyomavirus infection I. MCV T antigen expression in merkel cell carcinoma, lymphoid tissues and lymphoid tumors. *Int. J. Cancer* 125, 1243–1249. doi: 10.1002/ijc.24510

Shuda, M., Feng, H., Kwun, H. J., Rosen, S. T., Gjoerup, O., Moore, P. S., et al. (2008). T antigen mutations are a human tumor-specific signature for Merkel cell polyomavirus. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16272–16277. doi: 10.1073/pnas.0806526105

Shuda, M., Guastafierro, A., Geng, X., Shuda, Y., Ostrowski, S. M., Lukianov, S., et al. (2015). Merkel cell polyomavirus small T antigen induces cancer and embryonic merkel cell proliferation in a transgenic mouse model. *PLoS ONE* 10:e0142329. doi: 10.1371/journal.pone.0142329

Shuda, M., Kwun, H. J., Feng, H., Chang, Y., and Moore, P. S. (2011). Human Merkel cell polyomavirus small T antigen is an oncoprotein targeting the 4E-BP1 translation regulator. *J. Clin. Invest.* 121, 3623–3634. doi: 10.1172/JCI46323

Simons, J. N., Leary, T. P., Dawson, G. J., Pilot-Matias, T. J., Muerhoff, A. S., Schlauder, G. G., et al. (1995a). Isolation of novel virus-like sequences associated with human hepatitis. *Nat. Med.* 1, 564–569. doi: 10.1038/nm0695-564

Simons, J. N., Pilot-Matias, T. J., Leary, T. P., Dawson, G. J., Desai, S. M., Schlauder, G. G., et al. (1995b). Identification of two flavivirus-like genomes in the GB hepatitis agent. *Proc. Natl. Acad. Sci. U.S.A.* 92, 3401–3405. doi: 10.1073/pnas.92.8.3401

Spurgeon, M. E., Cheng, J., Bronson, R. T., Lambert, P. F., and Decaprio, J. A. (2015). Tumorigenic activity of merkel cell polyomavirus T antigens expressed in the stratified epithelium of mice. *Cancer Res.* 75, 1068–1079. doi: 10.1158/0008-5472.CAN-14-2425

Tabor, E., Gerety, R. J., Drucker, J. A., Seeff, L. B., Hoofnagle, J. H., Jackson, D. R., et al. (1978). Transmission of non-A, non-B hepatitis from man to chimpanzee. *Lancet* 1, 463–466. doi: 10.1016/S0140-6736(78)90132-0

Tang, K. W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M., and 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

Toptan, T., Yousem, S. A., Ho, J., Matsushima, Y., Stabile, L. P., Fernandez-Figueras, M. T., et al. (2016). Survey for human polyomaviruses in cancer. *JCI Insight* 1:e85562. doi: 10.1172/jci.insight.85562

Urisman, A., Molinaro, R. J., Fischer, N., Plummer, S. J., Casey, G., Klein, E. A., et al. (2006). Identification of a novel Gammaretrovirus in prostate tumors of patients homozygous for R462Q RNASEL variant. *PLoS Pathog.* 2:e25. doi: 10.1371/journal.ppat.0020025

van der Hoek, L., Pyrc, K., Jebbink, M. F., Vermeulen-Oost, W., Berkhout, R. J., Wolthers, K. C., et al. (2004). Identification of a new human coronavirus. *Nat. Med.* 10, 368–373. doi: 10.1038/nm1024

Velculescu, V. E., Madden, S. L., Zhang, L., Lash, A. E., Yu, J., Rago, C., et al. (1999). Analysis of human transcriptomes. *Nat. Genet.* 23, 387–388. doi: 10.1038/70487

Verhaegen, M. E., Mangelberger, D., Harms, P. W., Vozheiko, T. D., Weick, J. W., Wilbert, D. M., et al. (2015). Merkel cell polyomavirus small T antigen is oncogenic in transgenic mice. *J. Invest. Dermatol.* 135, 1415–1424. doi: 10.1038/jid.2014.446

Wang, D., Coscoy, L., Zylberberg, M., Avila, P. C., Boushey, H. A., Ganem, D., et al. (2002). Microarray-based detection and genotyping of viral pathogens. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15687–15692. doi: 10.1073/pnas.242579699

Weber, G., Shendure, J., Tanenbaum, D. M., Church, G. M., and Meyerson, M. (2002). Identification of foreign gene sequences by transcript filtering against the human genome. *Nat. Genet.* 30, 141–142. doi: 10.1038/ng818

Xu, Y., Stange-Thomann, N., Weber, G., Bo, R., Dodge, S., David, R. G., et al. (2003). Pathogen discovery from human tissue by sequence-based computational subtraction. *Genomics* 81, 329–335. doi: 10.1016/S0888-7543(02)00043-5

zur Hausen, H. (1999). Viruses in human cancers. *Eur. J. Cancer* 35, 1878–1885. doi: 10.1016/S0959-8049(99)00113-6

Zur Hausen, H. (2009). The search for infectious causes of human cancers: where and why. *Virology* 392, 1–10. doi: 10.1016/j.virol.2009.06.001