The landscape of viral associations in human cancers

Marc Zapatka[1]*, Ivan Borozan[2]*, Daniel S. Brewer[3,4]*, Murat Iskar[1]*, Adam Grundhoff[5], Malik Alawi[5,6], Nikita Desai[7,8], Holger Sültmann[9,10], Holger Moch[11], PCAWG-Pathogens[12], Colin S. Cooper[4,13], Roland Eils[14,15,16], Vincent Ferretti[17,18], Peter Lichter[1,10], PCAWG Consortium[19]

1 Division of Molecular Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany.

2 Informatics and Bio-computing Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada

3 Norwich Medical School, University of East Anglia, Norwich, UK

4 Earlham Institute, Norwich, UK

5 Heinrich-Pette-Institute, Leibniz Institute for Experimental Virology, Hamburg, Germany

6 Bioinformatics Core, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

7 Bioinformatics Group, Department of Computer Science, University College London, London, UK

8 Biomedical Data Science Laboratory, Francis Crick Institute, London UK

9 Division of Cancer Genome Research, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany

10 German Cancer Consortium (DKTK), Heidelberg, Germany

11 Department of Pathology and Molecular Pathology, University and University Hospital Zürich, Zürich, Switzerland

12 A full list of authors can be found at the end of the article

13 The Institute of Cancer Research, London, UK

14 Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany

15 Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, Heidelberg University and BioQuant Center, Heidelberg, Germany

16 Center for Digital Health, Berlin Institute of Health and Charité Universitätsmedizin Berlin, Berlin, Germany

17 Ontario Institute for Cancer Research, MaRS Centre, Toronto, Canada

18 Department of Biochemistry and Molecular Medicine, University of Montreal, Montreal, Canada.

19 A full list of authors can be found in the Supplementary Note

Equal contributions statement

MZ, IB, DSB and MI contributed equally.

Corresponding author statement

Peter Lichter peter.lichter@dkfz-heidelberg.de

Abstract

Here, as part of the Pan-Cancer-Analysis-of-Whole-Genomes (PCAWG), which aggregated whole genome and, for a subset, transcriptome sequencing data from 2,658 cancers across 38 tumor types, we systematically investigated potential viral pathogens using a consensus approach integrating three independent pipelines. Viruses were detected in 382 genome and 168 transcriptome datasets. We showed the high prevalence of known tumor-associated-viruses such as EBV, HBV and HPV16/18. The study revealed significant exclusivity of HPV with driver mutations in head-and-neck cancer and associated HPV with APOBEC mutational signatures, suggesting a role of impaired antiviral defense as driving force in cervical, bladder and head-and-neck carcinoma. For HBV, HPV16/18 and AAV2 viral integration was associated with local variations in genomic copy number. Integrations at the *TERT* promoter were coupled to high telomerase expression evidently activating this tumor driving process. High levels of endogenous retrovirus ERV1 expression were linked to worse survival outcome in kidney cancer.

Introduction

The World Health Organization estimates that 15.4% of all cancers are attributable to infections and 9.9% are linked to viruses[1,2]. Cancers attributable to infections have a greater incidence than any individual type of cancer worldwide. Eleven pathogens have been classified as carcinogenic agents in humans by the International Agency for Research on Cancer (IARC)[3]. After *Helicobacter pylori* (associated with 770,000 cases), the four most prominent infection related causes of cancer are estimated to be viral[2]: human papilloma virus (HPV)[4,5] (associated with 640,000 cancers), hepatitis B virus (HBV)[5] (420,000), hepatitis C virus (HCV)[6] (170,000) and Epstein-Barr Virus (EBV)[7] (120,000). It has been shown that viruses can contribute to the biology of multistep oncogenesis and are implicated in many of the hallmarks of cancer[8]. Most importantly, the discovery of links between infection and cancer types has provided actionable opportunities, such as HPV vaccines as preventive measure, to reduce the global impact of cancer. The following characteristics were proposed to define human viruses causing cancer through direct or indirect carcinogenesis[9]: i) Presence and persistence of viral DNA in tumor biopsies; ii) Growth promoting activity of viral genes in model systems; iii) Dependence of malignant phenotype on continuous viral oncogene expression or modification of host genes; iv) Epidemiological evidence that a virus infection represents a major risk for development of cancer.

The worldwide efforts of comprehensive genome and transcriptome analyses of tissue samples from cancer patients generate appropriate facilities for capturing information not only from human cells, but also from other - potentially pathogenic - organisms or viruses present in the tissue. A comprehensive collection of whole genome and transcriptome data from cancer tissues has been generated within the ICGC (International Cancer Genome Consortium) project PCAWG (Pan-Cancer Analysis of Whole Genomes)[10], providing a unique opportunity for a systematic search for tumor-associated viruses.

The PCAWG Consortium aggregated whole genome sequencing data from 2,658 cancers across 38 tumor types generated by the ICGC and TCGA projects. These sequencing data were re-analyzed with standardized, high-accuracy pipelines to align to the human genome (build hs37d5) and identify germline variants and somatically acquired mutations[10]. The PCAWG working group "Exploratory Pathogens" analyzed the whole genome sequencing (WGS) and whole transcriptome sequencing (RNA-seq) data of the PCAWG consensus cohort (2,656 donors). Focusing on viral pathogens, we applied three independently developed pathogen detection pipelines 'Computational Pathogen Sequence Identification' (CaPSID)[11], 'Pathogen Discovery Pipeline' (P-DiP), and 'SEarching for PATHogens' (SEPATH) to generate a large compendium of viral associations across 38 cancer types. We extensively characterized the known and novel viral associations by integrating driver mutations, mutational signatures, gene expression profiles and patient survival data of the same set of tumors analyzed in PCAWG.

Results

Identification of tumor-associated viruses

To identify the presence of viral sequences, we explored the WGS data of 5,354 tumor/normal samples across 38 cancer types, and 1,057 tumor RNA-seq data across 25 cancer types (Supplementary Tables 1,2,20). 195.8 billion reads were considered for analysis as they were not sufficiently aligned to the human reference genome in the PCAWG-generated alignment. Remaining reads ranged from 28,036 to 800 million per WGS and up to 120 million per RNA-seq tumor sample (Fig. 1a, Extended Data Figure 1a-c). Viral sequences were detected and quantified independently by three recently developed pathogen discovery pipelines CaPSID, P-DiP and SEPATH. The estimated relative abundance of a virus was calculated as viral reads per million extracted reads (PMER) at the genus level to improve consistency between pipelines. To minimize the rate of false positives in virus detection, we applied a strict threshold of PMER>1 supported by at least three viral reads as similarly suggested by previous studies[11,12]. Virus detection in a sample by at least two pipelines was considered as a consensus hit. In total, 532 genera were considered for the extensive virus search in at least two of the pipelines (Extended Data Figure 1d, Supplementary Table 18). Filtering of suspected viral laboratory contaminants was achieved through P-DiP, by examining each assembled contig of viral sequence segments for artificial, non-viral vector sequences and inspecting virus genome coverage across all positive samples (Extended Data Figure 2a). The most frequent hits prone to suspected contamination were lambdavirus, alphabaculovirus, microvirus, simplexvirus, hepacivirus, cytomegalovirus, orthopoxvirus and punalikevirus; these were observed across many tumor types (Fig. 1b). For example, mastadenovirus showed an uneven genome coverage which could result from contaminating vector sequences. Therefore, we analyzed the virus detections across sequencing dates (Extended Data Figure 2b) to assess any batch effect indicative of a contaminant; in mastadenovirus, we identified an association with sequencing date in early-onset prostate cancer regardless of tumor/normal state. We conclude that our mastadenovirus detections are due to a contamination occurring across projects worldwide where similar patterns could be identified.

We generally observed a strong overlap of the genera identified across pipelines (Extended Data Figure 1e, Supplementary Tables 6,7,11). From the whole genome dataset, we identified 321, 598 and 206 virus-tumor pairs for P-DiP, CaPSID and SEPATH, respectively (Fig. 2a, overlap after random permutation of detections, Extended Data Figure 3a, Supplementary Tables 3-5). The number of hits derived from the RNA-seq dataset differed between the pipelines (virus-tumor pairs: 101 for P-DiP, 83 for CaPSID, 41 for SEPATH; Fig. 2b, Supplementary Tables 8-10). SEPATH, using a k-mer approach, detected the lowest number of virus hits and was the least sensitive. Despite this, the identified viruses matched well with the consensus (DNA 90%, RNA 95%). P-DiP, based on an assembly and BLAST approach, detected more hits with 59% of the DNA and 54% of the RNA hits in the consensus set, while CaPSID, being most sensitive, implementing a two-step alignment process complemented by an assembly step, identified 60% (DNA) and 80% (RNA) hits within the consensus set. While the majority of the virus hits from RNA-seq (n=61/68) were overlapping with the WGS data, a lower fraction of detections in WGS data were present in the RNA-seq data (n=61/168 of 382 virus detections with RNA-seq data), emphasizing the importance of DNA sequencing for generating an unbiased catalogue of tumor-associated viruses. This difference can also be attributed to the viral life cycle as during incubation or latent phases, viral gene expression can be minimal[13]. Contrasting virus positive and negative samples within each organ type shows that the organ system, as expected, has a significant influence, but not virus positivity ($P < 2 \times 10^{-16}$, ANOVA modeling candidate reads dependent on organ system and virus positivity, Extended Data Figure 1c). This indicates that virus-positive tumors were not detected due to a higher number of candidate reads and is in line with the fact that the viral reads in most cases do not substantially contribute to the reads analyzed. 86% of the sequence hits detected from WGS and RNA-seq data were found to be from double-stranded DNA viruses (dsDNA) and dsDNA viruses with reverse transcriptase (Fig. 1c, Supplementary Table 19). This could be attributed to i) a higher frequency of tumor-associated viruses from these genome types[15], ii) a larger sequence dataset for WGS in comparison to RNA-seq, iii) a potential limitation of our analysis due to DNA and RNA extraction protocols that are less likely to include single-stranded (ss)DNA or RNA viruses or iv) the selection bias of tumor entities included in the PCAWG study (Fig. 1c).

The virome landscape across 38 distinct tumor types

We employed a consensus approach that resulted in a reliable set of 389 distinct virus-tumor pairs from WGS and RNA-seq data (Fig. 2a-d). Overall, 23 virus genera were detected across 356 tumor patients (13%). The top five most prevalent viruses (lymphocryptovirus, orthohepadnavirus, roseolovirus, alphapapillomavirus, cytomegalovirus) account for 85% of the consensus virus hits in tumors (n=329 out of 389). Among these five prevalent virus genera, three have been well described in the literature as drivers of tumor initiation and progression[9]: i) lymphocryptovirus (n=145 samples, 5.5%, e.g. Epstein-Barr Virus, EBV) is the most common viral infection across a variety of tumor entities mainly from gastrointestinal tract, and showed a much lower prevalence in the matched non-malignant control samples (n=82, 3%) (Fig. 2c); ii) orthohepadnavirus (n=67, 2.5%, e.g. hepatitis B, HBV) are as expected the most frequent among liver cancer with HBV present in 62 of 330 donors (18.9%); and iii) alphapapillomavirus (discussed below). Lymphocryptovirus (n=11), orthohepadnavirus (n=18) and alphapapillomavirus (n=32) were detected both in RNA and DNA sequencing data (Fig. 2c, left panel), with alphapapillomavirus being the most frequent (32 out of 39 consensus hits). This is in line with the constitutive expression of viral oncogenes in cancers associated with these viruses, a parameter supporting a direct role in carcinogenesis[9]. An in-depth analysis of the virus genome equivalents per human tumor genome equivalent considering genome sizes, coverage and tumor purity showed overall low viral genome equivalents even for established tumor viruses (Extended Data Figure 3c, Supplementary Table 12). Evidence for mouse mammary tumor virus (MMTV, PMER = 3.4) was detected in one renal carcinoma sample and in none of the 214 analyzed breast cancer samples. Previous work has suggested that MMTV may play a role in breast cancer but our comprehensive search of viral sequences could not identify any MMTV-positive case in breast cancer that would support this claim.

Roseolovirus and Alphatorquevirus show a higher number of hits in the non-malignant control samples, which were mainly derived from blood cells (Fig. 2c). For example, we identified 59 patients as Roseolovirus-positive (HHV-6A, HHV-6B, HHV-7) in their tumor (pancreas 6%, stomach 8%, colon/rectum 8.3%) and 90 patients positive in the non-malignant control samples. Considering the known cell tropism of roseolovirus for B- and T-cells[14], we asked whether immune infiltration would be higher in roseolovirus-positive tumors. However, we could not identify a stronger contribution of immune cells in virus positive tumor cases as estimated using CIBERSORT[15] (false discovery rate (FDR) corrected $P > 0.05$ for pancreas; Extended Data Figure 4a). Therefore, in line with current knowledge (reviewed in[16]), we cannot confirm a link between roseolovirus and immune cell content or tumor development. Furthermore, we could not identify actively transcribed viral genes for Roseolovirus and Alphatorquevirus at the transcriptome level. This is in agreement with the latent state of these viruses reported for blood mononuclear cells[14], and their transmission through blood transfusions[17]. Cytomegalovirus (CMV) was found, as expected[18], after identifying and removing contaminations both in stomach tumors (n=13) and the adjacent non-malignant tissue (n=11). In line with a recent publication[19], we could not detect CMV in the analyzed 294 CNS-tumors (146 medulloblastomas, 89 pilocytic astrocytoma, 41 glioblastomas, 18 oligodendrogliomas). Therefore, a previously debated role of this virus is not supported. Notably, we did not identify a significant enrichment of co-infection of multiple viruses in any tumor type (Extended Data Figure 3d).

Hepatitis B virus

Hepatitis B virus was most frequently detected among liver cancers (n=62). Comparing to the histopathological gold standard HBV PCR test[20,21] (n=228), we found the WGS-based consensus detections had the same high specificity (96.1%) and a high sensitivity (84.0%), indicating that the HBV detections by WGS are reliable (Fig. 3a, Extended Data Figure 4b, Supplementary Table 13). Furthermore, five out of seven cases positive in WGS and negative for HBV PCR showed positivity for HBAg indicating a high sensitivity of the WGS analysis. In summary, the precision (85.7%) and recall (84%) for the detection of HBV based on ~30x WGS is comparable to targeted PCR. We confirmed a significant exclusivity between HBV infection and *CTNNB1*, *TP53* and *ARID1A* mutations that was found in a larger liver cancer cohort analyzed by high throughput sequencing (FDR corrected $P = 5.35 \times 10^{-6}$, 0.0023 and 0.0023, DISCOVER[22])[23].

Epstein-Barr virus

Epstein-Barr virus was detected in many different tumor entities and normal samples (Fig. 2c). Comparing EBV PMERs in tumors and matched normals we see a stronger contribution in matched normals from matched solid tissue or tissue adjacent to the tumor (Extended Data Figure 4c). For samples showing reads for EBV in WGS and with available RNA sequencing data, the absolute score for immune cells based on CIBERSORT[15] was not significantly different between virus positive and negative samples (FDR corrected $P > 0.05$ for colon/rectum, head/neck, lymphoid, stomach; Extended Data Figure 4a). In summary, there is no evidence for a detection of EBV due to infiltrating immune cells. This indicates EBV presence in the respective organs. Based on the expression data available for the tumor samples we identified viral transcripts of the latent as well as lytic phase of the viral lifecycle (Fig. 3b, Extended Data Figure 4d, Supplementary Table 13). Eight of the nine tumors expressing lytic EBV transcripts are from stomach, confirming the active contribution of EBV to gastric cancer[24].

Alphapapillomavirus

Alphapapillomaviruses were mainly detected in head-and-neck cancer (n=18 of 57), cervical cancer (n=19 of 20) and in two bladder cancer cases out of 23, in agreement with previous studies[4,25,26]. There is also supporting evidence for 32 out of 39 alphapapillomavirus hits in the transcriptome data (Fig. 2c). We observed only one HPV subtype per tumor according to the P-DiP results with HPV16 being the dominant type in cervix (n=11) and head-and-neck (n=15) tumors, followed by HPV18 only present in cervical cancer (n=6). As reported previously[27], HPV33 was identified in head-and-neck (n=3) and cervix (n=1) tumors. Different HPV variants, type 6 and 45, were detected in bladder cancer.

In head-and-neck cancer, HPV-positive tumors exhibit an almost complete mutual exclusivity with mutations in known drivers like *TP53*, *CDKN2A* and *TERT* (FDR corrected $P = 1.73 \times 10^{-5}$, 1.73×10^{-5}, 0.012; multiple testing corrected for presented mutations in EBV and HPV, DISCOVER[22]) (Fig. 3c, Supplementary Table 13), as reported previously[25], which could be explained by a mutation independent inactivation of TP53 through the human papillomaviruses[28–30]. Furthermore, we identified mutational signature 2 as enriched for alphapapillomavirus positive cases in head-and-neck cancers (FDR corrected P=0.02; Fig. 3d, Supplementary Table 12,22)[31]. In addition, the expression of APOBEC3B is significantly higher in the virus positive head-and-neck cancers compared to their negative counterparts (P=1.6 × 10^{-4}, Fig. 3f)[32]. However, we did not observe enrichment of APOBEC signatures and expression changes for EBV-positive samples either in cervix or in other tissues.

Distinct expression profiles between virus positive and negative tumors in head-and-neck cancer are observed (Fig. 3e, Supplementary Table 23)[33]. Analyzing the immune cells estimated by CIBERSORT, we identified a significant increase in macrophages and T-cell signals in alphapapillomavirus positive head-and-neck cancers (P=0.004, 0.012 and 0.012 for follicular helper, CD8 and regulatory T-cells and P=0.018 for M1-Macrophages; FDR corrected for all viruses and cell types tested; Fig. 3g, Supplementary Table 24). Our integrative analysis on HPV reconfirms many of the findings related to HPV infection, illustrating the potential of our systematic approach in identifying and characterizing tumor-associated viruses.

Activation of endogenous retroviruses linked to outcome

Human endogenous retroviruses (HERV) are integrated in the human DNA originating from infection of germline cells by retroviruses over millions of years[34] and contribute over 500,000 individual sites, or 2.7% of the overall sequence the human genome[35,36]. The endogenous retroviruses were identified by the three pathogen detection pipelines but filtered by CaPSID and SEPATH. In addition, an alignment-based approach was used to detect HERV sequences embedded in the human reference genome that could be missed by the pipelines focusing only on non-human reads. In this study, we quantified the expression of HERV-like LTR (long terminal repeat) retrotransposons categorized into several clades by Repbase[37] as ERVL, ERVL-MaLR, ERV1, ERVK and ERV (Supplementary Table 14). In comparison to the other HERV families, ERV1 shows the strongest expression on average (Fig. 4a) and ERVK the highest fraction of active loci (Fig. 4b). Analyzing the expression of HERVs we could identify a strong expression for ERV1 in chronic lymphocytic leukemia compared to all other tumor tissues and adjacent normal tissues (Fig. 4c). However, we could not identify a link between transcriptionally active stemness markers (OCT3/4, SOX2, KLF4) and increased HERV expression, in contrast to what was reported in Ohnuki et al.[38] (Spearman Rank correlation < 0.35, Extended Data Figure 5). New data suggest that expression of HERVs is associated with prognosis in clear cell renal cell carcinoma (ccRCC)[39]. Analyzing the HERV expression in relation to patient survival, we identified a high ERV1 expression in kidney cancer linked to worse survival outcome (P=0.0081; Log-rank test; Fig. 4d, Extended Data Figure 6, Supplementary Table 15).

Genomic integration of viral sequences

Viral integration into the host genome has been shown to be a causal mechanism that can lead to cancer development[40]. This process is well-established for human papilloma viruses (HPVs) in cervical, head-and-neck and several other carcinomas, and for hepatitis B virus (HBV) in liver cancer[41,42].

Low confidence integration events were detected for the HHV4 (gastric cancer and malignant lymphoma) and HPV6b (head-and-neck and bladder carcinoma), while integration events with high confidence were demonstrated for HBV (liver cancer), Adeno-associated virus-2 (AAV2) (liver), HPV16/18 (both in cervical and head-and-neck carcinoma). Most of these integration events were found to be distributed across chromosomes and a significant number of viral integrations occur in the intronic (40%) regions while only 3.4% were detected in gene coding regions (Extended Data Figure 7a-d).

HBV was found to be integrated in 36 liver cancer specimens out of 61 patients identified as HBV-positive. Notably, genomic clusters of viral integrations were identified in *TERT* (ngc=6, number of integration sites within a genomic cluster), *KMT2B* (ngc=4), recently identified to be a likely cancer driver gene[43,44] and *RGS12* (ngc=3)(Extended Data Figure 7e).

Furthermore, two or more integration events in individual samples were observed in the gene (or gene promoter) regions of *CCCNE1*, *CDK15*, *FSIP2*, *HEATR6*, *LINC01158*, *MARS2* and *SLC1A7* (Fig. 5a). Additional events with two integration sites were also detected within a 50 kb distance away from *CLMP*, *CNTNAP2* and *LINC00359* genes. Integration events at *TERT* were found to recur in five different liver cancer samples. One sample had a genomic cluster of three viral integration events within *TERT* and four samples contained a single integration event in the *TERT* promoter, (3') or 5' UTR regions (Supplementary Table 17). When comparing gene expression in samples with virus integration to those without, only TERT was over-expressed (fold change 2.0) in two liver cancer samples (Fig. 5e). Additional genes with increased expression impacted by integration events include *TEKT3*, *CCNA2*, *CDK15* and *THRB* (Fig. 5a).

There was a significant association between HBV viral integrations and somatic copy number alterations (SCNAs, Fig. 5c). For samples with HBV integration events, the number of SCNAs was higher on average in the vicinity of viral integration sites (within 1 Mb) when compared to samples without HBV integration (mean: 4.2 vs 2.3, P=7.4 × 10^{-3}; two-sided paired t-test). No evidence for an SCNA association was seen for other integrated viruses like HPV16/18 (Extended Data Figure 8a-b).

HPV18 integration events were detected in seven tumors in total (Fig. 5b), with the most notable clusters of integration events in cervical cancer samples affecting *TALDO1* (ngc = 4) (Extended Data Figure 7g).

In 20 samples, HPV16 integration events were detected. Genomic clusters of viral integration sites were identified in cervical and head-and-neck cancer samples (Extended Data Figure 7f). None of these multiple integration events were observed to recur across patients (Fig. 5b). Integration events were also observed in two different lncRNAs, *LINC00111* and the plasmacytoma variant translocation 1 gene (*PVT1*), an oncogenic lncRNA[45,46]. Expression of both genes is strongly increased in the cases with HPV16 integration (Extended Data Figure 8f, Supplementary Table 17).

Using the PCAWG single nucleotide variant (SNV) calls[10] we have found a significant increase in the number of mutations occurring within +/- 10,000 bp of high-confidence viral integration sites (average number of mutations per sample = 0.41 (HPV16+) vs 0.14 (HPV16-), P=0.02; paired t-test one-sided, alternative greater, Extended Data Figure 8cd). Interestingly the integration sites are, compared to a random genome background, enriched in close proximity (<1000 bp) to common fragile sites (P=0.0018, Kolmogorov–Smirnov test). These results suggest that HPV16 integration reflects either characteristics of chromatin features that favor viral integration, such as fragile sites or regions with limited access to DNA repair complexes, or the influence of integrated HPV16 on the host genome. Such a correlation was not seen for the integration sites of other viruses (Extended Data Figure 8e). Finally, a single AAV2 integration event located in the intronic region of the cancer driver gene *KMT2B*[47] was detected in one liver cancer sample.

Identification of novel viral species or strains

De novo analysis using the CaPSID-pipeline has generated 56 different contigs that have been classified into taxonomic groups at the genus level by CSSSCL[48]. After filtering de novo contigs for their homology to known reference sequences, we have identified 29 contigs in 28 different tumor samples showing low sequence similarity (in average 63%) to any nucleotide sequence contained in the BLAST database. In this respect, our analysis has shown that WGS and RNA-seq can be used to identify isolates from potentially new viral species. However, the total number of novel isolates were quite low in comparison to viral hits to well-defined genera (Fig. 2c). These *de novo* contigs were not enriched for a specific tumor entity but rather distributed across cancer types including bladder, head/neck and cervical cancers (Extended Data Figure 9).

Discussion

Searching large pan-cancer genome and transcriptome data sets allowed the identification of an unexpectedly high percentage of virus associated cases (16%). In particular, analysis of tumor genomes, which were sequenced on average to a depth of at least 30-fold coverage, identified considerably more virus positive cases than investigations of transcriptome data alone, which is the search space analyzed in most previous virome studies. This is probably mainly due to viruses with no or only weak transcriptional activity in the given tumor tissue. Co-infections, generally believed to indicate a weak immune system, were very rare (Extended Data Figure 3d). This could, however, also be the result of selection processes during tumorigenesis.

While universal criteria for a causality of viral pathogens are prone to errors, it is worthwhile to look at individual features that might support a potentially pathomechanistic contribution of a given pathogen. These include aspects that affect the expression of host factors, e.g. upon viral integration, or the mutual exclusivity of the presence of viral genomes and other host factors, which are already known to play a role in the etiology of a given tumor type. Such aspects need to be carefully considered when discussing of what strengthens a potentially pathogenic role of virus.

Not surprisingly, known tumor associated viruses, such as EBV, HBV and HPV16/18, were among the most frequently detected targets. Interestingly, viral detection based on whole genome sequencing showed similar performance with respect to precision and recall as a targeted PCR for HBV indicating the sensitivity of this approach to detect viruses. This is particularly true for the common integration verified for HBV and HPV 16/18 in our study. In addition, the common theme of potential pathomechanistic effects by the genomic integration of viruses, also supported by the observations of multiple nearby integration sites in a given tumor genome that we also report in the present study, has gained further momentum. Analyzing the effect of viral integrations on gene expression, we identified several links to genes nearby the integration site. In this regard, the frequently observed integration of HBV at the *TERT* promoter accompanied with the transcriptional upregulation of *TERT*, constitutes an intriguing mechanistic example, since an increased activity of TERT is a well-understood driver of carcinogenesis[49]. Furthermore, we also linked viral integrations to increased mutations (SNVs and SCNAs) nearby the integration site.

The known causal role of HPV16/18 in several tumor entities, that triggered one of the largest measures in cancer prevention, has been the motivation for extensive elucidation of the pathogenetic processes involved. Nevertheless, comprehensive analyses of WGS and RNA-seq data sets revealed additional novel findings. While we confirmed the exclusivity of HPV infection and *TP53*, *CDKN2A* and *TERT* mutations in head-and-neck tumors, we could also link virus presence to an increase in mutations attributed to the mutational signature 2[50]. These are explained by the activity of APOBEC, which – among other effects – changes viral genome sequences as a mechanism of cellular defense against viruses[51,52]. This activation could play an important role in introducing further host genome alterations and, thus, constitute an important mechanism driving tumorigenesis[32,52]. In liver cancer mutations in *CTNNB1*, *TP53* and *ARID1A*, major primary oncogenes in this cancer type and HBV infections were confirmed to occur significantly exclusive[23]. Furthermore, the virus positive head-and-neck cancer samples had a significantly higher abundance of T-cell and M1 macrophage expression signals, which matches with the recently described subtypes of HNSCC that differ – among others – in virus infection and inflammation features.

Acknowledgements

We thank the IT Core Facility at the DKFZ for technical assistance, as well as to Michael Hain and Rolf Kabbe for computational support. We thank Sabrina Gerhardt for technical assistance in validation experiments. We acknowledge the contributions of the many clinical networks across ICGC and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for collation, realignment and harmonized variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects.

V.F. and I.B. received support for their work from the Ontario Institute for Cancer Research (OICR) through funding provided by the government of Ontario. A.G. received support for his work from the Leibniz Association (Grant Number: SAW 2015 IPB 2) and the German Center for Infection Research (Grant Number: TTU 01.801). P.L. and A.G. received support for this work from the German Federal Ministry of Education and Research (BMBF BioTop Grant Number 01EK1502C, ICGC-DE-Mining Grant Number 01KU1505A-G). D.S.B. and C.S.C. received support from Cancer Research UK C5047/A14835/A22530/A17528, the Dallaglio Foundation, Bob Champion Cancer Trust, The Masonic Charitable Foundation successor to The Grand Charity, The King Family and the Stephen Hargrave Trust. H.M. was supported by a Swiss National Science Foundation grant (No. S-87701-03-01).

Author Contributions

MI, DB, IB, MZ contributed equally, MZ, PL jointly supervised research. VF, RE, CC, MI, IB, MZ, PL conceived and designed the experiments. HS performed experiments. MI, DB, IB, MZ performed statistical analysis. ND, MI, AG, DB, IB, MZ analysed the data. VF, RE, CC, HM, MA, AG, DB, IB, MZ contributed reagents/materials/analysis tools. MI, DB, IB, MZ, PL wrote the paper. VF, AG, CC, DB, MI, IB, MZ and PL critiqued manuscript for intellectual content.

Competing Interests Statement

The authors have declared that they have no competing interests.

References

1. Parkin, D. M. The global health burden of infection-associated cancers in the year 2002. Int. J. cancer 118, 3030–44 (2006).

2. Plummer, M. et al. Global burden of cancers attributable to infections in 2012: a synthetic analysis. Lancet. Glob. Heal. 4, e609-16 (2016).

3. Bouvard, V. et al. A review of human carcinogens—Part B: biological agents. Lancet Oncol. (2009). doi:10.1016/S1470-2045(09)70096-8

4. Muñoz, N., Castellsagué, X., de González, A. B. & Gissmann, L. Chapter 1: HPV in the etiology of human cancer. Vaccine 24 Suppl 3, S3/1-10 (2006).

5. Bialecki, E. S. & Di Bisceglie, A. M. Clinical presentation and natural course of hepatocellular carcinoma. Eur. J. Gastroenterol. Hepatol. 17, 485–9 (2005).

6. Hermine, O. et al. Regression of splenic lymphoma with villous lymphocytes after treatment of hepatitis C virus infection. N. Engl. J. Med. 347, 89–94 (2002).

7. Thompson, M. P. & Kurzrock, R. Epstein-Barr virus and cancer. Clin. Cancer Res. 10, 803–21 (2004).

8. Mesri, E. A., Feitelson, M. A. & Munger, K. Human viral oncogenesis: a cancer hallmarks analysis. Cell Host Microbe 15, 266–82 (2014).
9. zur Hausen, H. Oncogenic DNA viruses. Oncogene 20, 7820–3 (2001).
10. PCAWG Consortium. Pan-cancer analysis of whole genomes. Nature (2019). doi:10.1101/162784
11. Borozan, I. et al. CaPSID: A bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. BMC Bioinformatics 13, 1–11 (2012).
12. Borozan, I., Watt, S. N. & Ferretti, V. Evaluation of alignment algorithms for discovery and identification of pathogens using RNA-Seq. PLoS One 8, e76935 (2013).
13. Nicoll, M. P. et al. The HSV-1 Latency-Associated Transcript Functions to Repress Latent Phase Lytic Gene Expression and Suppress Virus Reactivation from Latently Infected Neurons. PLoS Pathog. 12, e1005539 (2016).
14. Krug, L. T. & Pellett, P. E. Roseolovirus molecular biology: recent advances. Curr. Opin. Virol. 9, 170–7 (2014).
15. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods 12, 453–457 (2015).
16. Eliassen, E. et al. Human Herpesvirus 6 and Malignancy: A Review. Front. Oncol. 8, 512 (2018).
17. Spandole, S., Cimponeriu, D., Berca, L. M. & Mih escu, G. Human anelloviruses: an update of molecular, epidemiological and clinical aspects. Arch. Virol. 160, 893–908 (2015).
18. van de Berg, P. J. et al. Human cytomegalovirus induces systemic immune activation characterized by a type 1 cytokine signature. J. Infect. Dis. 202, 690–9 (2010).
19. Garcia-Martinez, A. et al. Lack of cytomegalovirus detection in human glioma. Virol. J. 14, 216 (2017).
20. Fujimoto, A. et al. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat. Genet. 42, 931–6 (2010).
21. Furuta, M. et al. Characterization of HBV integration patterns and timing in liver cancer and HBV-infected livers. Oncotarget 9, 25075–25088 (2018).
22. Canisius, S., Martens, J. W. M. & Wessels, L. F. A. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. Genome Biol. 17, 261 (2016).
23. Kawai-Kitahata, F. et al. Comprehensive analyses of mutations and hepatitis B virus integration in hepatocellular carcinoma with clinicopathological features. J. Gastroenterol. 51, 473–486 (2016).
24. Borozan, I., Zapatka, M., Frappier, L. & Ferretti, V. Analysis of Epstein-Barr Virus Genomes and Expression Profiles in Gastric Adenocarcinoma. J. Virol. 92, e01239-17 (2018).
25. Mork, J. et al. Human Papillomavirus Infection as a Risk Factor for Squamous-Cell Carcinoma of the Head and Neck. N. Engl. J. Med. 344, 1125–1131 (2001).
26. Li, N. et al. Human papillomavirus infection and bladder cancer risk: A meta-analysis. J. Infect. Dis. 204, 217–223 (2011).
27. Cao, S. et al. Divergent viral presentation among human tumors and adjacent normal tissues. Sci. Rep. 6, 28294 (2016).
28. Travé, G. & Zanier, K. HPV-mediated inactivation of tumor suppressor p53. Cell Cycle 15, 2231–2 (2016).
29. Werness, B. A., Levine, A. J. & Howley, P. M. Association of human papillomavirus types 16 and 18 E6 proteins with p53. Science 248, 76–9 (1990).
30. Scheffner, M., Werness, B. A., Huibregtse, J. M., Levine, A. J. & Howley, P. M. The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. Cell 63, 1129–36 (1990).
31. Henderson, S., Chakravarthy, A., Su, X., Boshoff, C. & Fenton, T. R. APOBEC-Mediated Cytosine Deamination Links PIK3CA Helical Domain Mutations to Human Papillomavirus-Driven Tumor Development. Cell Rep. 7, 1833–1841 (2014).
32. Burns, M. B., Temiz, N. A. & Harris, R. S. Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat. Genet. 45, 977–983 (2013).
33. Schlecht, N. et al. Gene expression profiles in HPV-infected head and neck cancer. J. Pathol. 213, 283–293 (2007).

34. Nelson, P. N. et al. Demystified. Human endogenous retroviruses. Mol. Pathol. 56, 11–18 (2003).

35. Paces, J. et al. HERVd: the Human Endogenous RetroViruses Database: update. Nucleic Acids Res. 32, D50 (2004).

36. Pavlícek, A., Paces, J., Elleder, D. & Hejnar, J. Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution. Genome Res. 12, 391–9 (2002).

37. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6, 11 (2015).

38. Ohnuki, M. et al. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. Proc. Natl. Acad. Sci. 111, 12426–12431 (2014).

39. Smith, C. C. et al. Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma. J. Clin. Invest. 128, 4804–4820 (2018).

40. Tang, K.-W. & Larsson, E. Tumour virology in the era of high-throughput genomics. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 372, 20160265 (2017).

41. Jiang, Z. et al. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome Res. 22, 593–601 (2012).

42. Hu, Z. et al. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. Nat. Genet. 47, 158–163 (2015).

43. Zhao, L.-H. et al. Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. Nat. Commun. 7, 12992 (2016).

44. Li, X. et al. The function of targeted host genes determines the oncogenicity of HBV integration in hepatocellular carcinoma. J. Hepatol. 60, 975–84 (2014).

45. Shen, C.-J., Cheng, Y.-M. & Wang, C.-L. LncRNA PVT1 epigenetically silences miR-195 and modulates EMT and chemoresistance in cervical cancer cells. J. Drug Target. 25, 637–644 (2017).

46. Tang, K.-W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M. & Larsson, E. The landscape of viral expression and host gene fusion and adaptation in human cancer. Nat. Commun. 4, 1–9 (2013).

47. Nault, J.-C. et al. Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. Nat. Genet. 47, 1187–93 (2015).

48. Borozan, I. & Ferretti, V. CSSSCL: A python package that uses combined sequence similarity scores for accurate taxonomic classification of long and short sequence reads. Bioinformatics 32, 453–455 (2015).

49. Sung, W. K. et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat. Genet. 44, 765–769 (2012).

50. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Rep. 3, 246–259 (2013).

51. Wallace, N. A. & Münger, K. The curious case of APOBEC3 activation by cancer-associated human papillomaviruses. PLoS Pathog. 14, e1006717 (2018).

52. Roberts, S. A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat. Genet. 45, 970–976 (2013).

Figure Legends

Fig. 1: Overview, design and summary statistics. (a) Workflow to identify and characterize viral sequences from the whole-genome and RNA sequencing of tumor and non-malignant samples. Viral hits were characterized in detail using several clinical annotations and resources generated by PCAWG. The red line represents the median. (b) Identified viral hits in contigs showing higher PMER's (viral reads **p**er **m**illion **e**xtracted **r**eads) for artificial sequences like vectors than the virus. Displayed are all viruses that occur in at least 20 primary tumor samples in the same contig together with an artificial sequence. (c) Summary of the viral search space used in the analysis grouped by virus genome type. The number of

virus positive tumor samples are indicated in the outer rings (PMER log scale for WGS and RNA sequencing data) as detected by any of the pipelines. Taxonomic relations between the viruses are indicated by the phylogenetic tree. dsDNA: double stranded DNA virus, dsDNA-RT: double-stranded DNA reverse transcriptase virus, ssDNA: single-stranded DNA virus, ssRNA-RT: single-stranded RNA reverse transcriptase virus, ssRNA: single-stranded RNA virus, dsRNA: double-stranded RNA virus. Fraction of hits in WGS and RNA sequencing data are depicted as stacked barplot.

Fig. 2: Detected viruses: Consensus for detected viruses in whole genome and transcriptome sequences. Number of genus hits among tumor samples for the three independent pipelines and the consensus set defined by evidence from multiple pipelines. (a) based on whole genome sequencing, (b) and based on transcriptome sequencing. (c) Heatmap showing the total number of viruses detected across various cancer entities. The sequencing data used for detection is indicated among the total number of hits (WGS= blue, RNA-seq=green). The fraction of virus positive samples is shown on top and the type of non-malignant tissue used in the analysis is indicated if more than 15% of the analyzed samples are from a respective tissue type (solid tissue, lymph node, blood or adjacent to primary tumor). (d) t-SNE clustering of the tumor samples based on PMER of their consensus virome profiles, using Pearson correlation as the distance metric. Major clusters are highlighted by indicating the strongest viral genus and the dominant tissue types that are positive in that cluster. Dot size represents the viral reads per million extracted reads (PMER).

Fig. 3: Virus specific findings. (a) HBV detections, validations and driver mutations in liver cancer. Star indicating mutual exclusivity between HBV detections and somatic driver gene mutations. Red boxes represent virus-positive tumor samples, purple - viral genomic integrations, green – driver mutations, grey – missing data. (b) Virus detections in gastric cancer samples, indication of virus phase (lytic/latent, dark red) and driver mutations (green). Yellow color indicates donors with virus-positive non-malignant samples. Grey box refers to samples with available RNA-seq data. (c) Virus detections (red) and driver mutations (green) in cervix (blue) and head and neck cancer (brown). Star indicating mutual exclusivity between alphapapillomavirus detections and somatic driver gene mutations. (d) Alphapapillomavirus detection and exposures of mutational APOBEC signatures SBS2 and SBS13, with sample sizes shown below. Wilcoxon rank-sum test (two-sided) revealed a significant difference ($P = 0.02$) of mutational signature exposure between virus-positive and negative head/neck tumor samples. Black line indicates median in each group. (e) Gene expression based tSNE map of head and neck cancer samples show a distinct gene expression profile for virus positive samples. Virus-positive and negative samples were labeled as red and grey dots, respectively. (f) The violin plot of APOBEC3B gene expression for alphapapillomavirus positive and negative samples in cervix and head/neck cancer (FDR corrected Wilcoxon rank-sum test, two-sided, $P = 1.6 \times 10^{-4}$). The center line represents median, the upper and lower boundaries of the violin plot refer to the maximum and minimum values, respectively. (g) Tumor-infiltrating immune cells as quantified by CIBERSORT using RNA-seq samples from head and neck cancer patients. All four cell types showed significant enrichment of immune cells in virus positive samples (FDR corrected Wilcoxon rank-sum test two-sided, n=24 vs 18). Tukey boxplot indicates the median by the middle line and the 25-75th percentiles by the box. The whiskers were drawn up to the 1.5 interquartile range from the lower and upper quartile.

Fig. 4: Endogenous retroviruses. (a) Heatmap showing the HERV expression across all tumor samples. HERV TPMs were grouped by family and summed up. Hierarchical clustering was performed by family based on Manhattan distance with complete linkage after log2 transformation of HERVs transcripts per million (TPM) expression values. (b) Fraction of active loci in the genome with a TPM >0.2 plotted against the fraction of samples. (c) TPM based expression of the highly expressed HERVs ERV1 and ERVK across tumor types. n described number of tumor samples analyzed. Violin plots marked with the median as red dot. The upper and lower boundaries of the violin plot extend out to the maximum and minimum values. (d) Survival difference between kidney cancer samples expressing high (red) and low levels (blue) of ERV1. Kaplan-Meier curve shows the overall survival of

patients (n=113) with high and low levels of ERV1 using a cutoff of 16.3 tpm (Log-rank test *P*=0.0081). Patients at risk are provided below.

Fig. 5: Impact of virus integration. (a) Integration sites detected in gene regions (including promoter, exon, intron and 5' UTR regions) are labeled in red for increased gene expression and blue for expression measured. Rows of each heatmap designate nearest genes to the integration sites and columns represent individual ICGC donor and project IDs. Intragenic HBV integration sites detected in liver cancers (ICGC project codes: LIRI, LIHC and LINC). For TERT and SEMA6D intergenic integrations are shown as well. (b) Integration sites detected for HPV-16 and 18 in head/neck (samples color coded magenta) and cervical (samples color coded blue) cancers (ICGC project codes: HNSC and CESC) gene labels with star indicated HPV18 as opposed to HPV16 viral integrations. (c) A local increase in the number of SCNAs was shown in the vicinity of HBV viral integrations (n=21 viral integrations in individual patients, $P=7.4 \times 10^{-3}$; two-sided paired *t*-test). (d) Genomic visualization of the HBV virus integration sites relative to the *TERT* gene in five liver tumor patients. (e) The increased gene expression (FPKM) of *TERT* gene in two liver tumors with HBV viral integrations in comparison to the *TERT* expression in tumor and non-malignant adjacent tissue. Tumor samples with a non-coding driver mutation were labeled in orange.

Methods

Identifying potential pathogenic reads

To reduce the number of reads to be considered for the pathogen search, we identified potential pathogenic reads using script available at https://github.com/mzapatka/p-dip. Based on the reads aligned by BWA[53] or STAR[54] to hg19 using the standard PCAWG approach, we identified read pairs where at least one read did not show a good mapping to the human genome (longest stretch of mapped bases from 20 to 30 bases), were unmapped or mapped to NC_007605 (human herpesvirus 4, which is contained in the 1000 genomes version of the hg19 human reference genome) and extracted these for further processing. To speed up the extraction, we used bamcollate2 from Biobambam2[55] v2.08 as input stream to the python script.

Identification of endogenous retroviruses

The expression of the endogenous retroviruses was analyzed based on the RNA sequencing data and aligned STAR based on the setting developed within PCAWG (hg19 and Gencode 19). In contrast to the standard pipeline, the reference transcripts from Gencode 19 were enriched by adding HERV locations extracted from RepeatMasker (URL: http://www.repeatmasker.org, rmsk from UCSC, version 17/08/03) and Featurecounts (subread-1.5.3)[56] applied to identify reads mapping to the modified reference transcripts. Resulting reads counts were converted into transcripts per million (TPM) according to Wagner et al.[57].

Norwich SEarching for PATHogens (SEPATH) pipeline

Our starting point is to take reads that are not mapped to the human genome using the extracted potential pathogenic reads. Low quality bases (q<30) are trimmed from the read ends and the TruSeq indexed adapter and TruSeq universal adapter are removed using Cutadapt v1.8.1[58]. Reads less than 32 bp were discarded. Additional filtering is performed to remove reads containing more than 5% of Ns or those with low complexity (dust method with maximum score of 10) using Prinseq v0.20.3[59]. Metagenomic Phylogenic Analysis (MetaPhlAn)[60,61] is then applied to identify and quantify the presence of bacterial and viral populations. MetaPhlAn comes with a curated marker database of ~1M unique clade-specific marker genes identified from reference genomes (version 2.0 of the database was used).

Reads are aligned against the unique marker gene database using BowTie2 v2.2.1[62] with presets set to sensitive. Reads are then counted and normalized giving a relative-abundance estimation at each level of the phylogenetic tree.

Detection and Analysis of Microbial Infectious Agents by NGS P-DiP - Pathogen discovery pipeline

The assembly based pipeline (P-DiP) was further developed based on a version implemented by Malik Alawi and Adam Grundhoff[63]. In summary, the pipeline runs preprocessing, assembly and BLAST searches and stores processing details and final results in a postgresql database. For the whole genome sequencing and RNA sequencing, we started with the potential pathogenic reads extracted from the BWA aligned whole genome sequencing bam files. As a first step, reads were trimmed based on quality using trimmomatic. Thereafter, host reads were subtracted by aligning to the human reference genome (WGS: hg19 excluding NC_007605 and hs37d5 and adding phiX, RNAseq: Homo_sapiens.GRCh37.dna.primary_assembly) using Bowtie2 (2.2.8)[62]. Trinity (v2.0.6)[64] was used for the read assembly of WGS reads which were not aligned by bowtie with sufficient quality (not aligned with --very-fast (-D 5 -R 1 -N 0 -L 22 -i S,0,2.50) to Homo_sapiens.GRCh37.ncrna, Homo_sapiens.GRCh37.cdna.all or PhiX) for the RNA sequencing data we applied idba assembler (V1.1.3)[65]. Assembled contigs were filtered by size (minimal length of 300 bp). Abundance was estimated by remapping all reads not aligning to the human reference to the assembled contigs using bowtie2 again. Putative PCR duplicates identified by mapping location were removed from the abundance count. The taxonomic classification of the size filtered contigs was performed using the BLAST+ package (2.2.30)[66] and nucleotide databases nt (2015-05-15) and nr (2015-04-20). For the extraction of pathogen hits R-scripts were used to filter the blast results (at https://github.com/mzapatka/p-dip). In summary, for each of the contig, the best BLAST hits for each segment of the contig were considered and the reads aligning to these segments identified. Potential contaminants were defined based on the taxonomy annotation in NCBI taxonomy. Any taxonomy id below plasmids (36549), transposons (2387), midivariant sequences (31896), insertion sequences (2673), artificial sequences (81077) and synthetic viruses (512285) was annotated as potential contamination. Segments with higher read counts of these sequences compared to pathogen hits were flagged as contaminants and not further considered.

Computational Pathogen Sequence Identification (CaPSID) description of the analysis workflow

CaPSID's[11] metagenomic analysis pipeline starts by first processing a BAM file containing reads sequenced from a tumor (or normal) sample aligned to the human reference sequence (GRCh37/hg19). Reads that did not map to the human reference are extracted and filtered for low complexity and quality using the SGA[67] preprocessing module and then aligned in single-end mode using the Bowtie2 aligner[62] to 5,652 NCBI[68] viral reference sequences (RefSeq) and a filter sequence reference database composed of 5,242 bacterial and 1,138 fungal reference sequences also downloaded from the NCBI. In order to improve the sensitivity and specificity with which viral sequences are detected, reads that did not map to any reference with Bowtie2 are realigned against the same viral RefSeq database, using a more sensitive SHRiMP2 aligner using its local alignment mode[69]. At the completion of this two-step alignment process, reads aligning to viral reference sequences are annotated using the information stored in the CaPSID's genome database containing full NCBI GenBank and taxa information. Using information from each aligned read CaPSID then calculates the following four metrics: (i) the total number of reads (or hits) aligning across any given viral genome, (ii) the total number of reads aligning only across gene regions within any given viral genome, (iii) the total coverage across each viral genome and (iv) the maximum coverage across any of the genes in a given viral genome.

Filtering of viral candidates with low significance

In a typical analysis of tumor whole genome or transcriptome sample, CaPSID reports candidate sequences from dozens of different viral genomes, some of which are not related to cancer phenotype. Some of these reported viral hits are also due to a series of experimental and computational artifacts. In order to reduce the number of potential false positives CaPSID pipeline flags viral genomes could be the result of artifacts present in the sequencing data or those with no obvious relation to cancer phenotype and that could be filtered later on. The following criteria are used to flag and filter for potential viral candidates: (i) flag viral candidates with low coverage, (ii) flag bacteriophage viral genome sequences, (iii) report only viral candidates with read composition different from the one expected when generated from the host's reference GRCh37/hg19 sequence, (iv) flag viral candidates that are typically not known to infect humans and those with low read abundance and/or low overall alignment read accuracy.

In the first step CaPSID flags viral genomes with low read count and/or coverage using its three metrics including: total number of uniquely aligned reads < 3, total genome coverage < 10% and maximum gene coverage < 50%. Viral genomes with low read count can arise as a result of i) low read/transcript abundance in the human sequenced sample, ii) non-specific alignment between sequenced short reads (for example low complexity reads) and viral reference sequences and iii) for RNA-seq library preparation where highly expressed transcripts generally dominate over low abundance targets. In order to limit reporting viral genomes with very low coverage, we chose to flag all those with maximum gene coverage < 50%. Since this lower bound on the maximum gene coverage applies to individual genes and not to the complete viral genome, it is unlikely that viruses with such low coverage are biologically significant. The second step in our filtering approach is to flag bacteriophage viral genomes that are most likely not related to any cancer phenotype. Bacteriophages are detected as a result of the presence of bacteria (or bacterial contamination) in human sequenced samples. The third step is used to determine whether the genome coverage observed for each viral candidate is different from the one expected to arise from reads originating exclusively from the human reference DNA GRCh37/hg19 sequence. To build the CaPSID background model we use the ART NGS read simulator. The entire GRCh37/hg19 sequence reference file is first fed to the ART[70] simulator (parameters: art_illumina [Illumina platform] -l [read length = 100 bp] -f [the fold of read coverage to be simulated = 100] with default values for indels and substitution rates), which then generates single-end (or paired-end) reads and base quality values.

Reads simulated by ART are then aligned to the viral reference sequence database using the same alignment approach for reads originating from tumor samples (see above). CaPSID then calculated the four metrics for the GRCh37/hg19 background model using the alignment information from simulated reads aligning to viral reference sequences. The fourth step consists of flagging viral candidates that are typically not known to infect humans using a dictionary of ~ 130 terms that we have compiled from a database of all viruses known to infect humans. In addition to the above filtering criteria CaPSID also considers the read abundance associated with each viral candidate sequence (abundance is expressed in terms of aligned reads in parts-per million of total number of unmapped reads) and the average read percent identity with which reads align to a given viral candidate reference sequence.

De novo assembly and taxonomic classification of contigs

The purpose of this analysis step is to attempt to characterize potential novel viral sequences at the species or subspecies level. Unaligned reads which could not be aligned to any of the filter/host or viral reference sequences are assembled into contigs using the IDBA algorithm[65]. Assembled contigs are then masked for repeat regions using RepeatMasker and then filtered for their size and read coverage (contig length >= 500 bp and coverage > 5x). Resulting contigs are then assigned into taxonomic groups at the genus level using the CSSSCL algorithm[48]. Contigs lacking sequence homology to reference sequences contained in the CaPSID or blast nucleotide databases with percent identity < 90% are then selected as suggestive of the presence of new viral strains/isolates or species.

Defining consensus hits

Identification of the consensus hits was achieved by optimizing two features of the individual genus hits: PMER 1 as cutoff (see analysis of the validation set) and percentage identity >90%. 90% percentage identity threshold was determined based on our benchmarking study[12] indicating that an alignment-based approach can still accurately characterize viral sequences with up to 10% mutation rate (when compared to sequences stored in a reference database). Lowering the threshold, with which short reads align to any given reference sequence below 90% identity on average, results in a drop of sequence coverage due to a high attrition rate of aligned reads, lowering the detection rate and thus providing more uncertain characterization of viral candidates. Notably, there was no difference in the PMER distribution of common hits across the three pipelines indicating that a common detection cut-off is reasonable (Extended Data Figure 3b).

The consensus set was restricted to genera that were covered in at least two detection pipelines (Extended Data Figure 1b). Notably, we could not detect any more hits with high PMERs using the unique search space of P-DiP, indicating that almost all of the viral hits from individual pipelines were also screened by another pipeline.

Virus integration detection analysis

A subset of viral candidates identified to be present in tumor samples by the CaPSID's analysis pipeline (parameters used: PMER >= 1.1 and genome coverage > simulated background model) was selected for the detection of viral integration events using the VERSE[71] algorithm. This subset of viruses included: Herpesviruses (HHV-1, 2, 4, 5, 6A/B), Simian virus 40 (SV40) and 12 (SV12), Human immunodeficiency virus (HIV1), Human and Simian T-cell lymphotropic virus type 1 (HTLV1 and STLV1), BK polyomavirus (BKP), Human parvovirus B19, Mouse mammary tumor virus, Murine type C retrovirus, Mason-Pfizer monkey virus, Hepatitis B (HBV), Papilloma viruses (HPV-16, 18 and 6a and Adeno-associated virus - 2 (AAV2). Below we describe the steps used for viral integration detection analysis.

Viral integration events in the host can be detected using paired-end NGS technologies that facilitate the detection of genomic rearrangements, as well as gene fusions and novel transcripts. VERSE is capable of determining virus integration sites within a single base resolution by requiring the presence of both chimeric and soft clipped reads. In addition, VERSE improves the detection through customizing reference genomes and was shown to substantially enhance the sensitivity of virus integration site detection[71]. VERSE categorizes its predictions into one of two classes: (a) a 'high' confidence hit with a single base resolution - if there is a sufficient number of soft-clipped reads to support an integration locus so that CREST is able to detect it; (b) a 'low' confidence hit with a 10 bp resolution where CREST has failed to detect an integration event for the lack of high quality soft-clipped reads.

In order to further limit the false positive rate associated with viral integration sites we compare results obtained with VERSE to those from Fujimoto et al[72]. Out of 64 whole genome liver cancer samples with HBV integration events reported in Fujimoto et al., 50 are part of the PCAWG dataset analyzed in this study. 45 out of 50 of these samples tested positive for HBV when analyzed by CaPSID (filtering criteria used; PMER >= 1.0, genome coverage > host background model and read % identity >= 89%). In addition, 50 of these WGS samples had 23 matching whole transcriptome (WT) samples and 22 of these were identified to be positive for HBV by CaPSID (filtering criteria used; maximum gene coverage >= 50%, read % identity >= 89% and PMER >= 1.0). By combining WGS and whole transcriptome tumor samples together, 47/50 in total tested positive for HBV when analyzed by CaPSID.

Using VERSE, virus integration sites were detected in 28/47 (60%) of these. This result indicates that for a subset of viral integration events, VERSE might be a more stringent approach when compared to the one used in Fujimoto et al. This can be explained by the fact that VERSE requires both the presence of paired-end chimeric and soft clipped reads while the method presented in Fujimoto et al. relies on paired-end reads only. In order to explore these results further we compared integration sites obtained with VERSE and Fujimoto et al. with an overlapping window of 10 bp. Our analysis indicates that among 23 integration sites identified by VERSE in whole transcriptome data and that overlap with the results from Fujimoto et al., 91% of these are classified with high confidence hits and only 9% with low (N total overlap = 23, high = 21 (91%) and low = 2 (9%)). However, a similar result is not observed for integration events found using WGS data (N total overlap = 14, high = 6 (43%), low = 8 (57%)) where the proportion of integration events classified as high and low is similar.

Thus, our analysis indicates that one important factor for improving the agreement between these two datasets is the confidence level assigned by VERSE to each candidate integration site - but only in the case when integration sites are detected using whole transcriptome data. In order to reduce the potential number of false positives we decided to use all integration sites predicted by VERSE when these are obtained using WGS data and only high confidence calls when using whole transcriptome data.

Contaminations

Based on the presence of vector sequences in the contig assembled by the P-DiP and based on the background model from CaPSID we could identify which virus hits originate from common lab contaminants or due to sequence similarities to the human genome. In addition, we filtered known contaminants (see below). For P-DiP we filtered all hits not having more target reads than any artificial sequence (excluding artificial viruses) on an individual contig region. Hits caused by vector and other artificial sequences were identified analyzing the assembled contigs for combined hits to viral pathogens and artificial sequences. Checking viral hits occurring at least 40 times in a such contig we could clearly separate contaminants from viral pathogens.

The gammaretrovirus hits (NCBI taxonomy id: 153135, species: murine leukemia virus) were also marked as artifacts, based on the additional BLAST hits of the corresponding contigs to the *Mus musculus* genome by P-DiP, as well as on the background model of the CaPSID pipeline designed to limit the number of spurious hits. Most frequent virus hits prone to contamination by artificial sequences are Lambdalikevirus, Alphabaculovirus, Microvirus, Simplexvirus, Hepacivirus, Cytomegalovirus, Orthopoxvirus and Punalikevirus. But restricting to at least 1 PMER for the potential virus hit contaminants drop to one Cytomegalovirus case.

Filtering contaminants

We filtered all Microviridae (taxonomy ID: 10841) because of the phix174 spike-in used during sequencing. Caudovirales (taxonomy ID: 28883), tailed bacteriophages, were removed as they typically infect bacterial hosts. Baculoviridae were filtered because of infecting insect cells and commonly being used in the lab. The virus coverage was analyzed by aligning the potential pathogenic reads with BWA mem to the human hg19 reference genome after adding the respective virus reference sequence most frequently detected within the genus. Coverage was thereafter calculated base specific using BEDTools coverage. As we identified EBV in all 14 normal blood controls from ovarian cancer that were EBV immortalized these were removed from the virus hits.

Integration of external PCAWG datasets

We tested for mutual exclusivity e.g between virus detections and driver gene mutations by applying DISCOVER[22]. Based on the gene expression data, immune cell proportions were analyzed by CIBERSORT[15]. For survival analysis, Cox proportional hazards analysis was performed using R libraries ‘survival’ and ‘survminer’ for the figures. The optimal cutpoints were identified by maxstat using the method presented in Lausen and Schumacher[73] (library maxstat).

Virus load

The viral load in relation to the human genome equivalents was calculated based on the human bases sequenced (read length x number of reads mapped to the human genomes), tumor sample purity (if available 100% otherwise) assuming a ploidy of two and using a human genome size of 2,897,310,462 bases (mappable part of the human genome). This number of human genome equivalents was then related to the viral genome equivalents calculated based on viral reads identified, read length and virus genome size.

$$tumor\ genome\ equivalents = \frac{read\ length \times number\ of\ reads\ mapped\ to\ the\ human\ genomes}{mappable\ human\ genome\ size\ \times tumor\ ploidy} \times tumor\ purity$$

$$virus\ genome\ equivalents = \frac{readlength\ \times\ number\ of\ viral}{virus\ genome\ size}$$

$$virus\ load\ = \frac{virus\ genome\ equivalents}{tumor\ genome\ equivalents}$$

Human research participants

The Ethics oversight for the PCAWG protocol was undertaken by the TCGA Program Office and the Ethics and Governance Committee of the ICGC. Each individual ICGC and TCGA project that contributed data to PCAWG had their own local arrangements for ethics oversight and regulatory alignment.

Statistics

If not specified otherwise, we used two-sided Wilcoxon rank-sum test for groups with n >3. Further details can be accessed at the ' Life Sciences Reporting Summary'.

Data Availability Statement

Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium is described here[10] and available for download at https://dcc.icgc.org/releases/PCAWG. Additional information on accessing the data, including raw read files, can be found at https://docs.icgc.org/pcawg/data/. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identification information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; http://icgc.org/daco) for the ICGC portion. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization.

Data sets described specifically in this manuscript can be found in the Supplementary Tables.

Code availability Statement

The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at https://dockstore.org/search?search=pcawg under the GNU General Public License v3.0, which allows for reuse and distribution. The pathogen discovery pipeline P-DiP is available on github at https://github.com/mzapatka/p-dip. CaPSID is available from the github pages (
pipeline: https://github.com/capsid/capsid-pipeline,
webapp: https://github.com/capsid/capsid-webapp). The taxonomic classifier CSSSCL is available from https://github.com/oicr-ibc/cssscl.

Methods-only References

53. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr. arXiv 1303.3997 (2013).

54. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).

55. Tischler, G. & Leonard, S. Biobambam: Tools for read pair collation based algorithms on BAM files. Source Code Biol. Med. 9, 1–17 (2014).

56. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930 (2014).

57. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 131, 281–285 (2012).

58. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10 (2011).

59. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–4 (2011).

60. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods 12, 902–3 (2015).

61. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods 9, 811–4 (2012).

62. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–9 (2012).

63. Fischer, N. et al. Rapid Metagenomic Diagnostics for Suspected Outbreak of Severe Pneumonia. Emerg. Infect. Dis. 20, 1072–1075 (2014).

64. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652 (2011).

65. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–8 (2012).

66. Camacho, C. et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 421 (2009).

67. Simpson, J. T. & Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 22, 549–56 (2012).

68. Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res. 37, D32–D36 (2009).

69. David, M., Dzamba, M., Lister, D., Ilie, L. & Brudno, M. SHRiMP2: Sensitive yet Practical Short Read Mapping. Bioinformatics 27, 1011–1012 (2011).

70. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: A next-generation sequencing read simulator. Bioinformatics 28, 593–594 (2012).

71. Wang, Q., Jia, P. & Zhao, Z. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome Med. 7, 2 (2015).

72. Fujimoto, A. et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat. Genet. 48, 500–9 (2016).

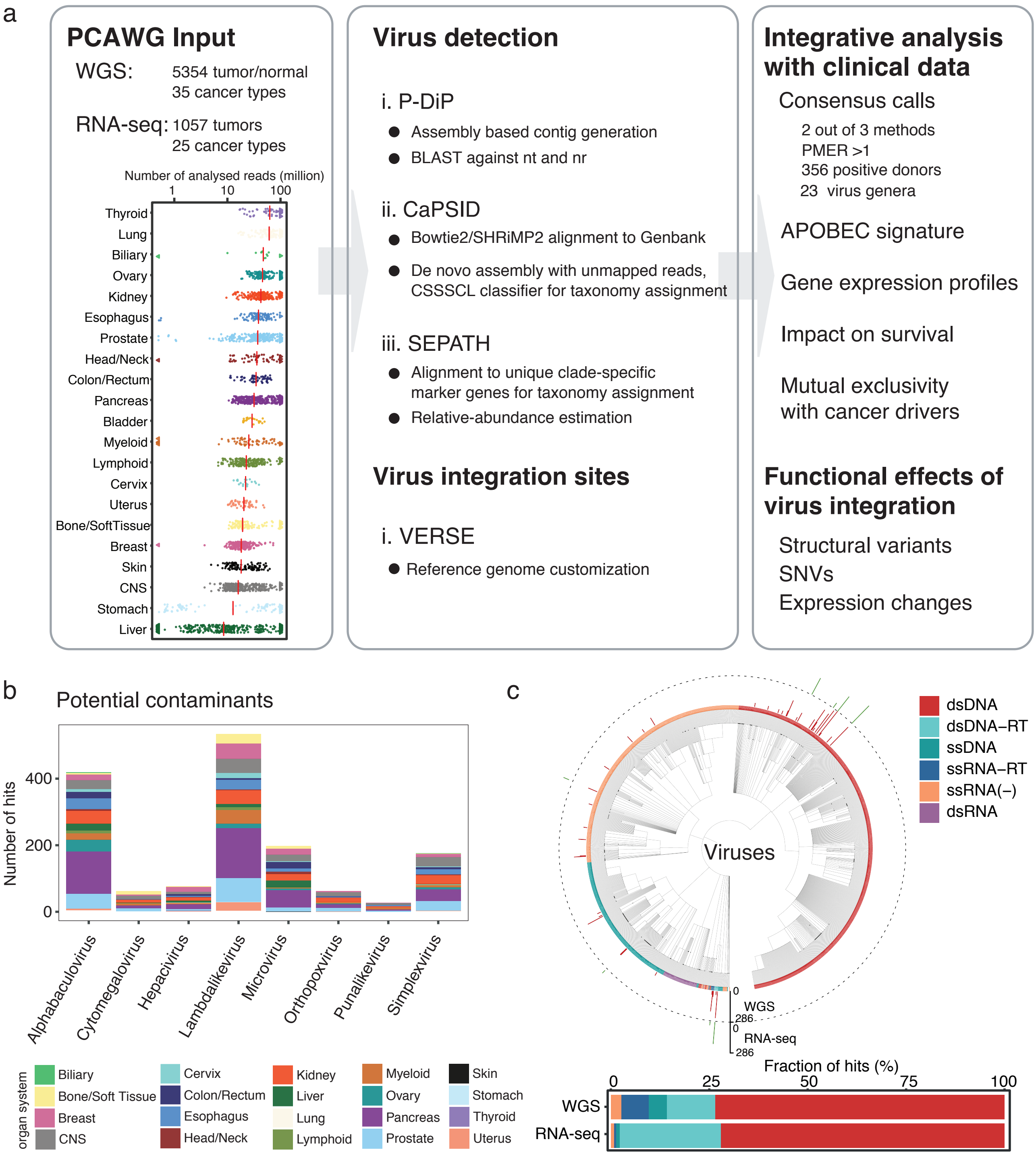
73. Lausen, B. & Schumacher, M. Maximally Selected Rank Statistics. Biometrics 48, 73–85 (1992).

PCAWG-Pathogens Members

Malik Alawi[5,6], Ivan Borozan[2], Daniel S Brewer[3,4], Colin S Cooper[4,13], Nikita Desai[7,8], Roland Eils[14,15,16], Vincent Ferretti[17,18], Adam Grundhoff[5], Murat Iskar[1], Kortine Kleinheinz[20,21], **Peter Lichter#**[1,10], Hidewaki Nakagawa[22], Akinyemi I Ojesina[23,24,25], Chandra Sekhar Pedamallu[26,27,28], Matthias Schlesner[20,29], Xiaoping Su[30] and Marc Zapatka[1]

# Corresponding author

20. Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany.
21. Institute of Pharmacy and Molecular Biotechnology and BioQuant, Heidelberg University, Heidelberg 69120, Germany.
22. RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan.
23. Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA.
24. HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA.
25. O'Neal Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL 35294, USA.
26. Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.
27. Harvard Medical School, Boston, MA 02115, USA.
28. Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA.
29. Bioinformatics and Omics Data Analytics, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany.
30. University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.



PCAWG Input

WGS: 5354 tumor/normal
35 cancer types

RNA-seq: 1057 tumors
25 cancer types

Virus detection

i. P-DiP
- Assembly based contig generation
- BLAST against nt and nr

ii. CaPSID
- Bowtie2/SHRiMP2 alignment to Genbank
- De novo assembly with unmapped reads, CSSSCL classifier for taxonomy assignment

iii. SEPATH
- Alignment to unique clade-specific marker genes for taxonomy assignment
- Relative-abundance estimation

Virus integration sites

i. VERSE
- Reference genome customization

Integrative analysis with clinical data

Consensus calls
2 out of 3 methods
PMER >1
356 positive donors
23 virus genera

APOBEC signature

Gene expression profiles

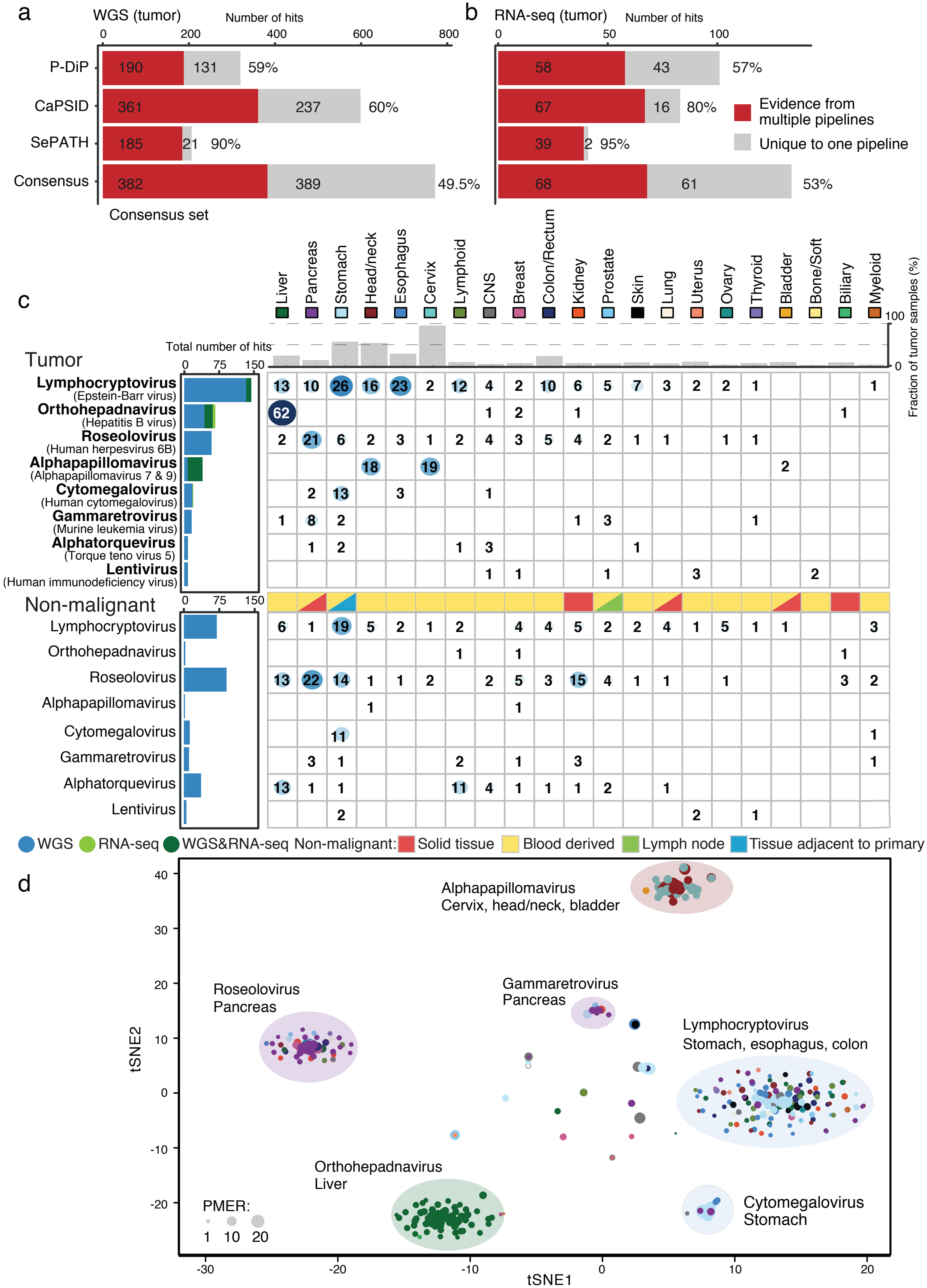
Impact on survival

Mutual exclusivity with cancer drivers

Functional effects of virus integration

Structural variants
SNVs
Expression changes

Potential contaminants

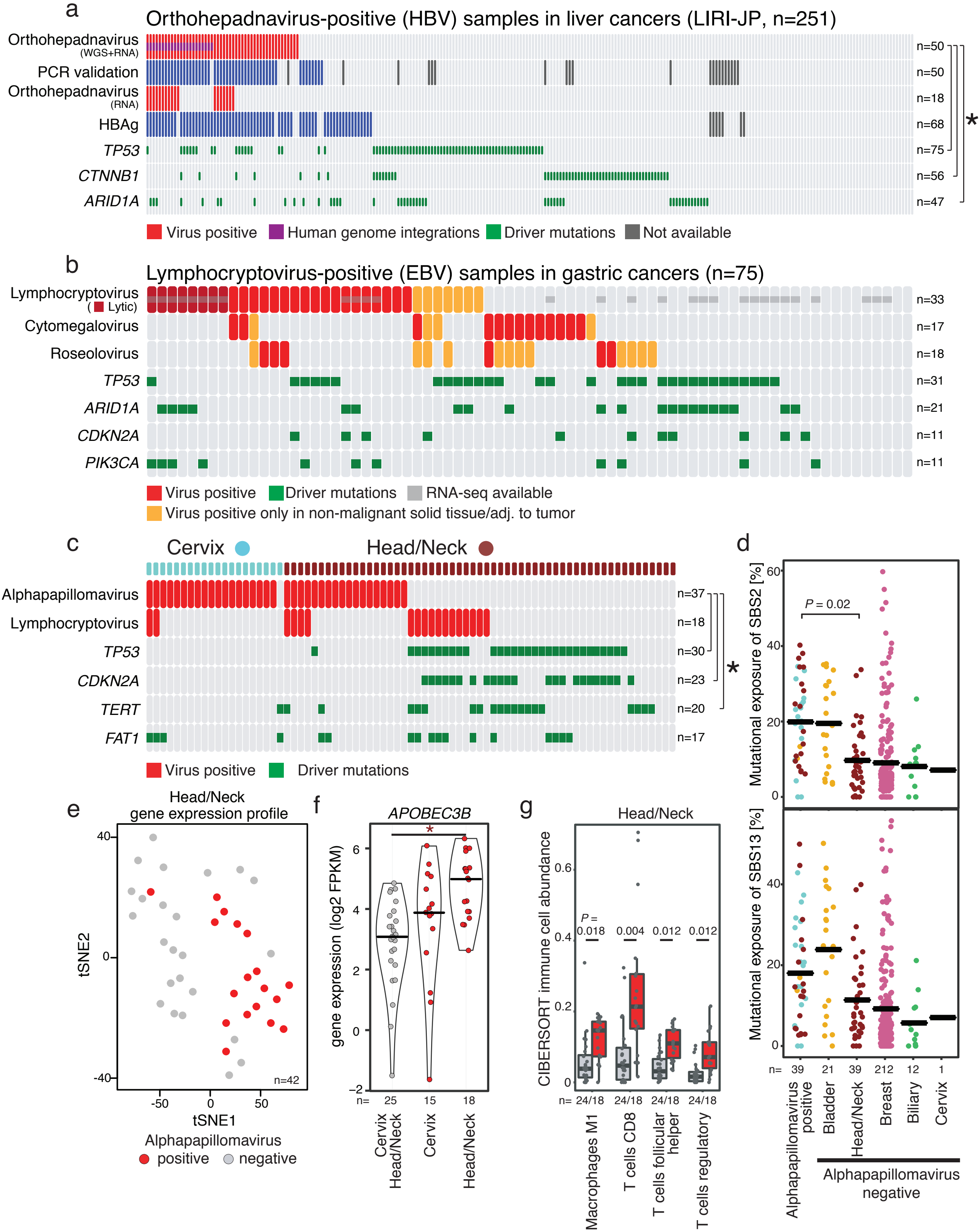


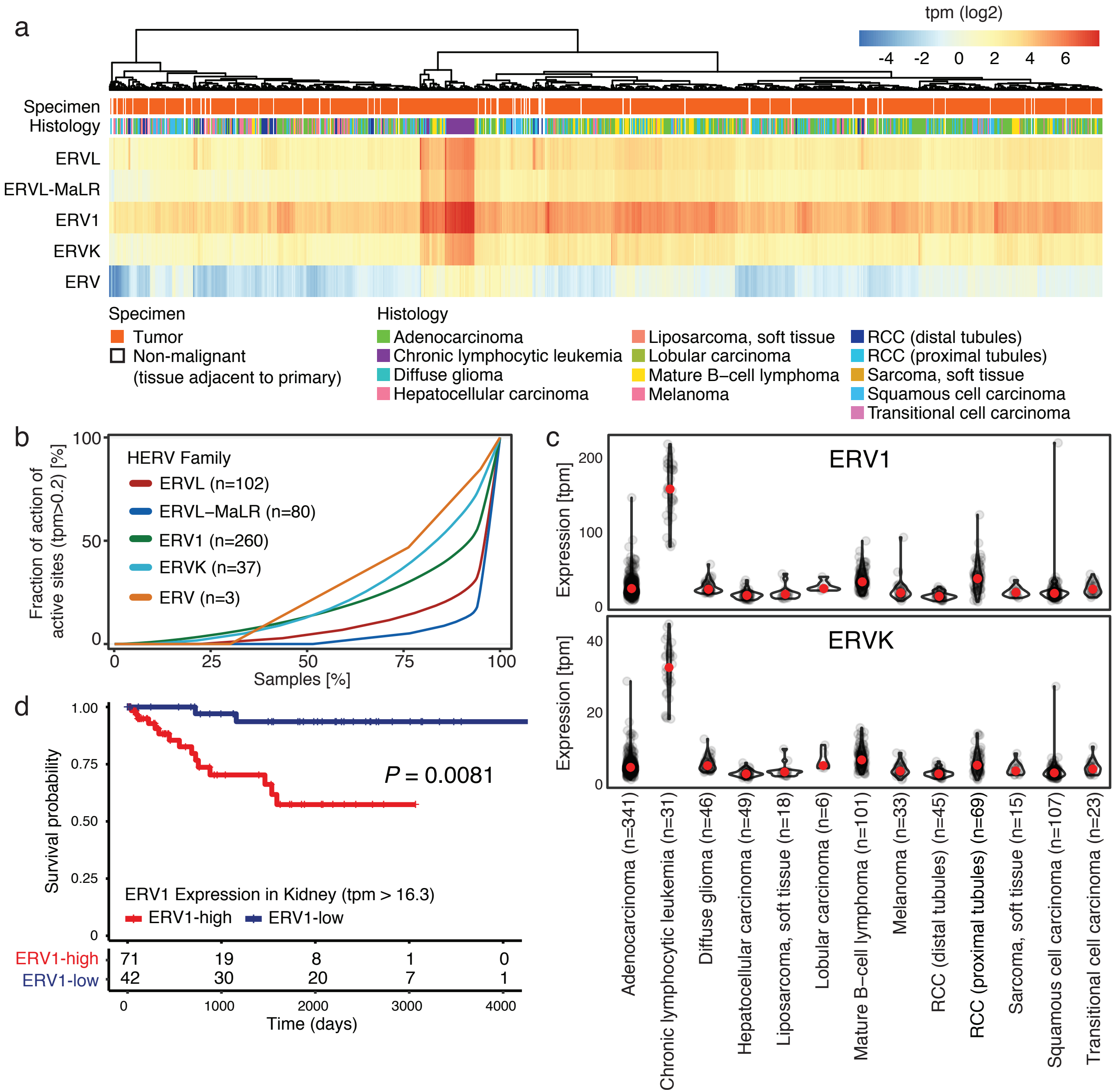
Tumor

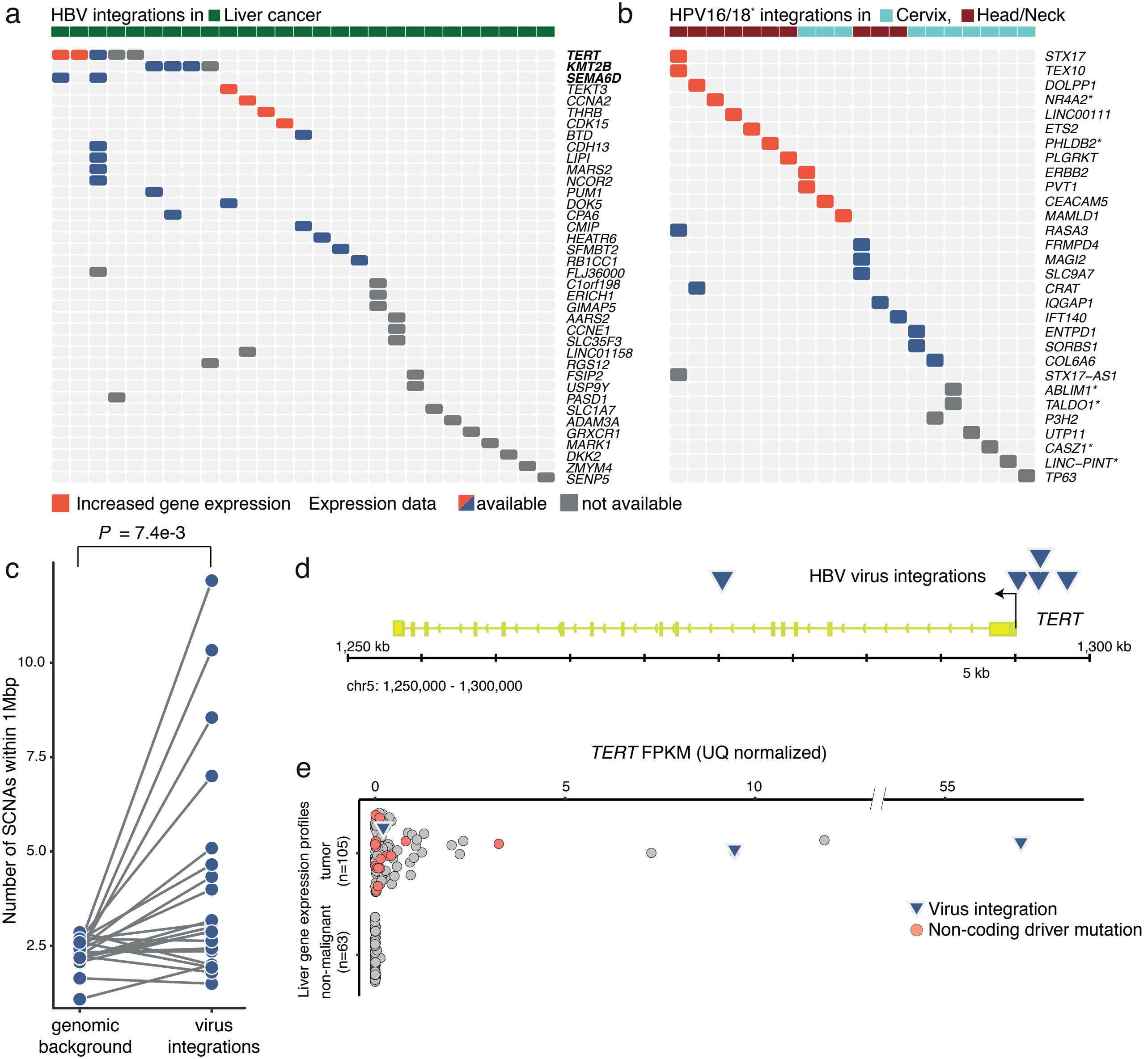
	Liver	Pancreas	Stomach	Head/neck	Esophagus	Cervix	Lymphoid	CNS	Breast	Colon/Rectum	Kidney	Prostate	Skin	Lung	Uterus	Ovary	Thyroid	Bladder	Bone/Soft	Biliary	Myeloid
Lymphocryptovirus (Epstein-Barr virus)	13	10	26	16	23	2	12	4	2	10	6	5	7	3	2	2	1				1
Orthohepadnavirus (Hepatitis B virus)	62							1	2		1									1	
Roseolovirus (Human herpesvirus 6B)	2	21	6	2	3	1	2	4	3	5	4	2	1	1		1	1				
Alphapapillomavirus (Alphapapillomavirus 7 & 9)				18		19												2			
Cytomegalovirus (Human cytomegalovirus)		2	13		3			1													
Gammaretrovirus (Murine leukemia virus)	1	8	2								1	3					1				
Alphatorquevirus (Torque teno virus 5)		1	2				1	3					1								
Lentivirus (Human immunodeficiency virus)								1	1			1			3				2		

Non-malignant

	Liver	Pancreas	Stomach	Head/neck	Esophagus	Cervix	Lymphoid	CNS	Breast	Colon/Rectum	Kidney	Prostate	Skin	Lung	Uterus	Ovary	Thyroid	Bladder	Bone/Soft	Biliary	Myeloid
Lymphocryptovirus	6	1	19	5	2	1	2		4	4	5	2	2	4	1	5	1	1			3
Orthohepadnavirus							1		1											1	
Roseolovirus	13	22	14	1	1	2		2	5	3	15	4	1	1		1				3	2
Alphapapillomavirus				1					1												
Cytomegalovirus			11																		1
Gammaretrovirus		3	1				2		1		3										1
Alphatorquevirus	13	1	1				11	4	1	1	1	2		1							
Lentivirus			2												2		1				


a
Orthohepadnavirus-positive (HBV) samples in liver cancers (LIRI-JP, n=251)
Orthohepadnavirus
(WGS+RNA)
PCR validation
Orthohepadnavirus
(RNA)
HBAg
TP53
CTNNB1
ARID1A
n=50
n=50
n=18
n=68
n=75
n=56
n=47
*
Virus positive
Human genome integrations
Driver mutations
Not available
b
Lymphocryptovirus-positive (EBV) samples in gastric cancers (n=75)
Lymphocryptovirus
(Lytic)
Cytomegalovirus
Roseolovirus
TP53
ARID1A
CDKN2A
PIK3CA
n=33
n=17
n=18
n=31
n=21
n=11
n=11
Virus positive
Driver mutations
RNA-seq available
Virus positive only in non-malignant solid tissue/adj. to tumor
c
Cervix
Head/Neck
Alphapapillomavirus
Lymphocryptovirus
TP53
CDKN2A
TERT
FAT1
n=37
n=18
n=30
n=23
n=20
n=17
*
Virus positive
Driver mutations
d
Mutational exposure of SBS2 [%]
60
40
20
0
P = 0.02
Mutational exposure of SBS13 [%]
40
20
0
n= 39 21 39 212 12 1
Alphapapillomavirus positive
Bladder
Head/Neck
Breast
Biliary
Cervix
Alphapapillomavirus negative
e
Head/Neck
gene expression profile
tSNE2
40
0
-40
-50
0
50
tSNE1
n=42
Alphapapillomavirus
positive
negative
f
APOBEC3B
*
gene expression (log2 FPKM)
6
4
2
0
-2
n= 25 15 18
Cervix
Head/Neck
Cervix
Head/Neck
g
Head/Neck
CIBERSORT immune cell abundance
0.6
0.4
0.2
0
P =
0.018
0.004
0.012
0.012
n= 24/18 24/18 24/18 24/18
Macrophages M1
T cells CD8
T cells follicular helper
T cells regulatory


a
tpm (log2)
-4
-2
0
2
4
6
Specimen
Histology
ERVL
ERVL-MaLR
ERV1
ERVK
ERV
Specimen
Tumor
Non-malignant
(tissue adjacent to primary)
Histology
Adenocarcinoma
Chronic lymphocytic leukemia
Diffuse glioma
Hepatocellular carcinoma
Liposarcoma, soft tissue
Lobular carcinoma
Mature B–cell lymphoma
Melanoma
RCC (distal tubules)
RCC (proximal tubules)
Sarcoma, soft tissue
Squamous cell carcinoma
Transitional cell carcinoma
b
HERV Family
ERVL (n=102)
ERVL–MaLR (n=80)
ERV1 (n=260)
ERVK (n=37)
ERV (n=3)
Fraction of action of
active sites (tpm>0.2) [%]
0
50
100
0
25
50
75
100
Samples [%]
c
ERV1
Expression [tpm]
0
100
200
ERVK
Expression [tpm]
0
20
40
Adenocarcinoma (n=341)
Chronic lymphocytic leukemia (n=31)
Diffuse glioma (n=46)
Hepatocellular carcinoma (n=49)
Liposarcoma, soft tissue (n=18)
Lobular carcinoma (n=6)
Mature B–cell lymphoma (n=101)
Melanoma (n=33)
RCC (distal tubules) (n=45)
RCC (proximal tubules) (n=69)
Sarcoma, soft tissue (n=15)
Squamous cell carcinoma (n=107)
Transitional cell carcinoma (n=23)
d
Survival probability
1.00
0.75
0.50
0.25
0
P = 0.0081
ERV1 Expression in Kidney (tpm > 16.3)
ERV1-high
ERV1-low
ERV1-high 71 19 8 1 0
ERV1-low 42 30 20 7 1
0
1000
2000
3000
4000
Time (days)


a
HBV integrations in 🟩 Liver cancer
TERT
KMT2B
SEMA6D
TEKT3
CCNA2
THRB
CDK15
BTD
CDH13
LIPI
MARS2
NCOR2
PUM1
DOK5
CPA6
CMIP
HEATR6
SFMBT2
RB1CC1
FLJ36000
C1orf198
ERICH1
GIMAP5
AARS2
CCNE1
SLC35F3
LINC01158
RGS12
FSIP2
USP9Y
PASD1
SLC1A7
ADAM3A
GRXCR1
MARK1
DKK2
ZMYM4
SENP5
b
HPV16/18* integrations in Cervix, Head/Neck
STX17
TEX10
DOLPP1
NR4A2*
LINC00111
ETS2
PHLDB2*
PLGRKT
ERBB2
PVT1
CEACAM5
MAMLD1
RASA3
FRMPD4
MAGI2
SLC9A7
CRAT
IQGAP1
IFT140
ENTPD1
SORBS1
COL6A6
STX17–AS1
ABLIM1*
TALDO1*
P3H2
UTP11
CASZ1*
LINC–PINT*
TP63
🟥 Increased gene expression
Expression data available
not available
c
P = 7.4e-3
Number of SCNAs within 1Mbp
10.0
7.5
5.0
2.5
genomic background
virus integrations
d
HBV virus integrations
TERT
1,250 kb
1,300 kb
5 kb
chr5: 1,250,000 - 1,300,000
e
TERT FPKM (UQ normalized)
0
5
10
55
Liver gene expression profiles
tumor (n=105)
non-malignant (n=63)
▼ Virus integration
● Non-coding driver mutation