The landscape of viral associations in human cancers

Marc Zapatka[1]*, Ivan Borozan[2]*, Daniel S. Brewer[3,4,5]*, Murat Iskar[1]*, Adam Grundhoff[6], Malik Alawi[7], Nikita Desai[8,9], Colin S. Cooper[3,4], Roland Eils[10,11], Vincent Ferretti[12], Peter Lichter[1,13] on behalf of the PCAWG Pathogens Working Group, and the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network.

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 The Institute of Cancer Research, London, UK.
4 Norwich Medical School, University of East Anglia, Norwich, UK
5 Earlham Institute, Norwich, UK.
6 Virus Genomics, Heinrich-Pette-Institute, Hamburg, Germany
7 Bioinformatics Core, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
8 Division of Cancer Studies, King's College London, London, UK
9 Cancer Systems Biology Laboratory, The Francis Crick Institute, London, UK
10 Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany.
11 Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, Heidelberg University and BioQuant Center, Heidelberg, Germany
12 Ontario Institute for Cancer Research, MaRS Centre, Toronto, Canada
13 German Cancer Consortium (DKTK), Heidelberg, Germany.

Abstract

Potential viral pathogens were systematically investigated in the whole-genome and transcriptome sequencing of 2384 donors across Pan-Cancer Analysis of Whole Genomes using a consensus approach integrating three independent pathogen detection pipelines. Viruses were detected in 485 genomic and 70 transcriptome data sets. Besides confirming the prevalence of known tumor associated viruses such as EBV, HBV and several HPV types, numerous novel features were discovered. A strong association was observed between HPV infection and the APOBEC mutational signatures, suggesting the role of impaired mechanism of antiviral cellular defense as a driving force in the development of cervical, bladder and head neck carcinoma. Viral integration into the host genome was observed for HBV, HPV16, HPV18 and AAV2 and associated with a local increase in copy number variations. The recurrent viral integrations at the *TERT* promoter were coupled to high telomerase expression uncovering a further mechanism to activate this tumor driving process. Most importantly, our systematic analysis revealed a novel association between mastadenovirus and several tumor entities. In renal cancer, mastadenovirus presence was significantly exclusive with well-known driver mutations in kidney cancer and defined a patient subgroup with better survival. Independent from mastadenovirus presence, high levels of endogenous retrovirus ERV1 expression is linked to worse survival outcome in kidney cancer.

Figure Outline

Figure 1 Overview, design and summary statistics

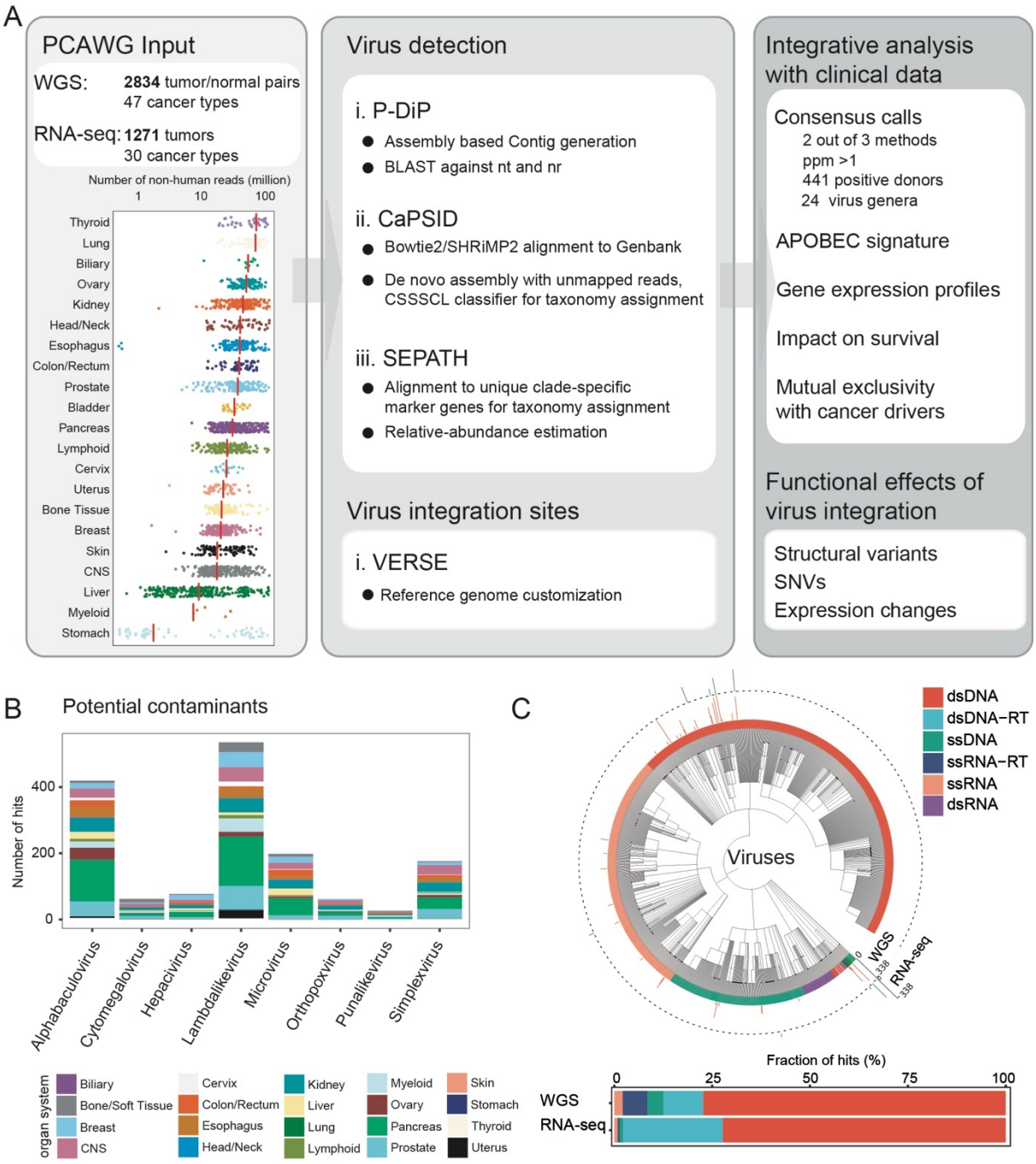

Figure 1: Overview, design and summary statistics. (A) Workflow to identify and characterize viral sequences from the whole-genome and RNA sequencing of tumor and non-malignant samples. Viral hits were characterized in detail using several clinical annotations and resources generated by PCAWG. (B) Identified 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 and normal samples are indicated in the outer ring (log scale). 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.

Figure 2 Detected viruses: Consensus for detected viruses in whole genome and transcriptome sequences

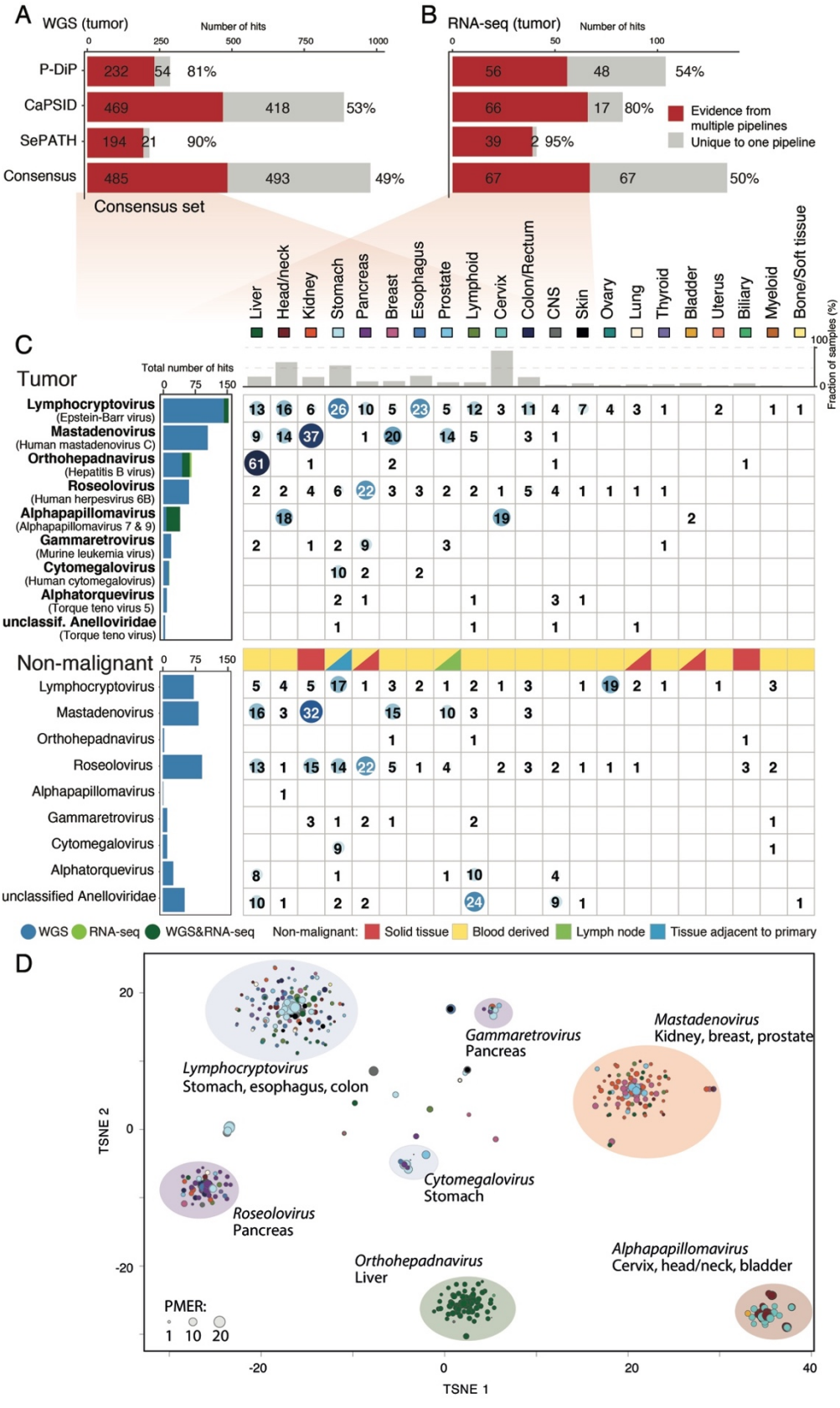

Figure 2: Detected viruses: Consensus for detected viruses in whole genome and transcriptome sequences. Number of genus hits among tumor samples for the three independent pipelines and the consensus set defined by evidence from multiple pipelines. (A) based on whole genome sequencing, (B) and based on 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).

Figure 3 Virus specific findings

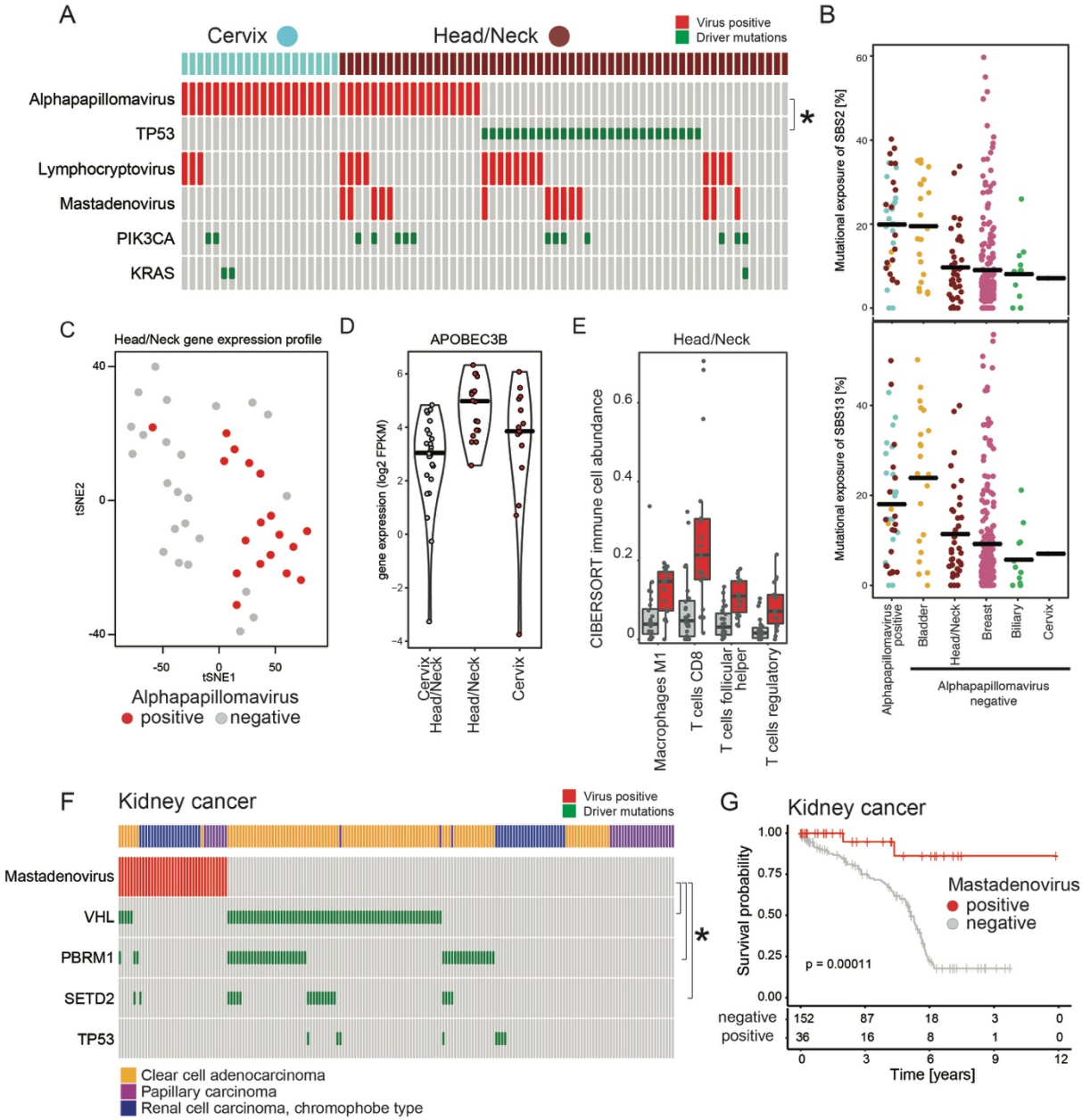

Figure 3: Virus specific findings. (A) Virus detections and driver mutations in cervix and head and neck cancer as visualized with Oncoprint graph. (B) Alphapapillomavirus detection and exposures of mutational APOBEC signatures SBS2 and SBS13. (C) Gene expression based tSNE map of head and neck cancer samples show a distinct gene expression profile for virus positive samples. (D) The boxplot of APOBEC3B gene expression for alphapapillomavirus positive and negative samples in cervix and head/neck cancer. (E) Tumor-infiltrating immune cells as quantified by CIBERSORT linked to alphapapillomavirus infections in head and neck cancer. (F) An oncoprint summarizing the known driver mutations and mastadenovirus infection in kidney cancer. Three subtypes of the kidney cancer were

colored: yellow for clear cell adenocarcinoma, purple for papillary carcinoma and blue for renal cell carcinoma, chromophobe type. (G) Survival analysis in kidney cancer samples stratifying patients by mastadenovirus status.

Figure 4 Endogenous retroviruses

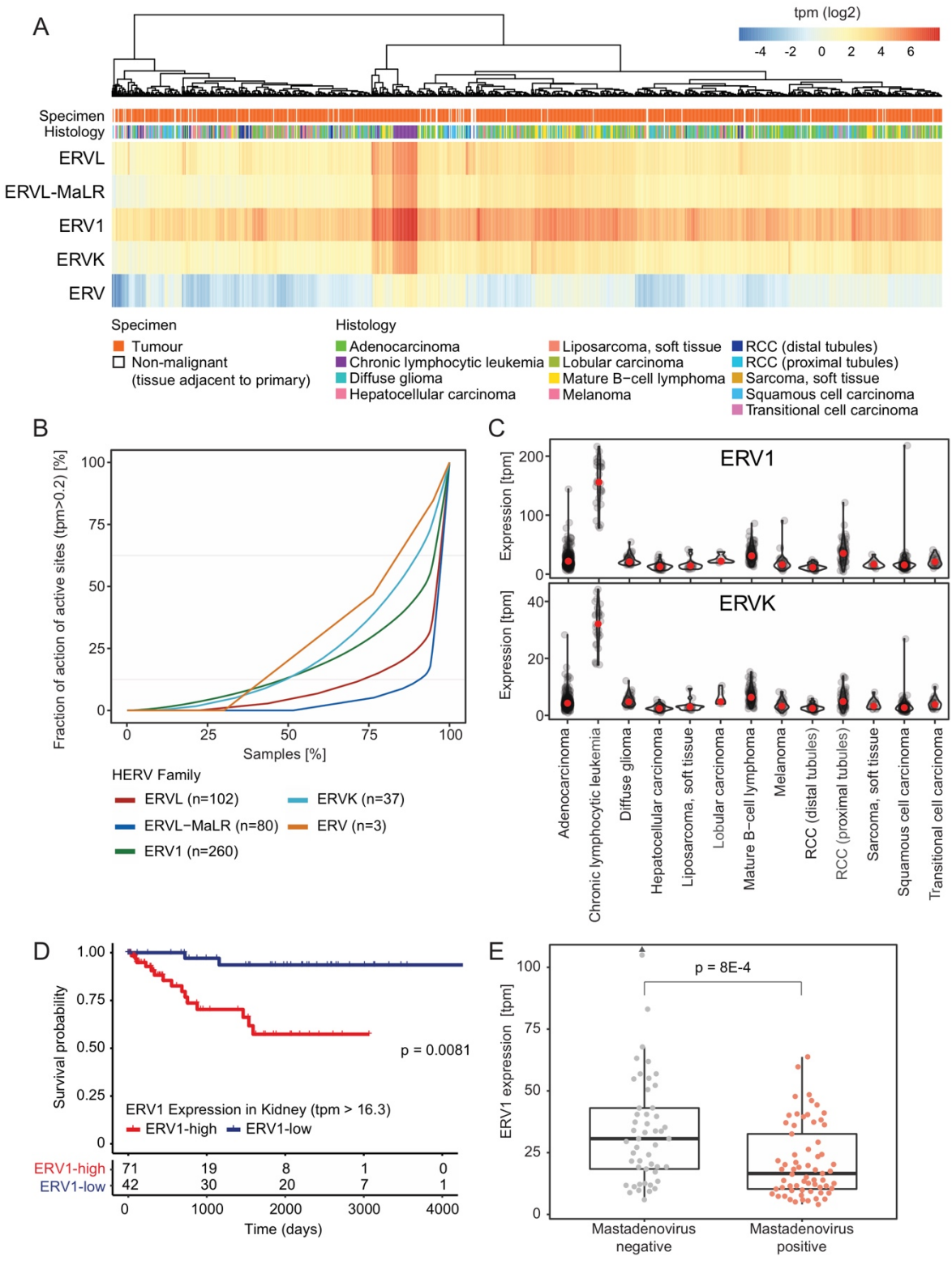

Figure 4: Endogenous retroviruses. (A) Heatmap showing the HERV expression across all tumor samples. HERV TPMs were grouped by family and summed up. Hierarchical clustering was performed by family based on Manhattan distance with complete linkage after log2

transformation of HERVs 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. (D) Survival difference between kidney cancer samples expressing high and low levels of ERV1. (E) The boxplot of ERV1 expression in kidney cancer in relation to mastadenovirus status.

Figure 5 **Impact of virus integration**

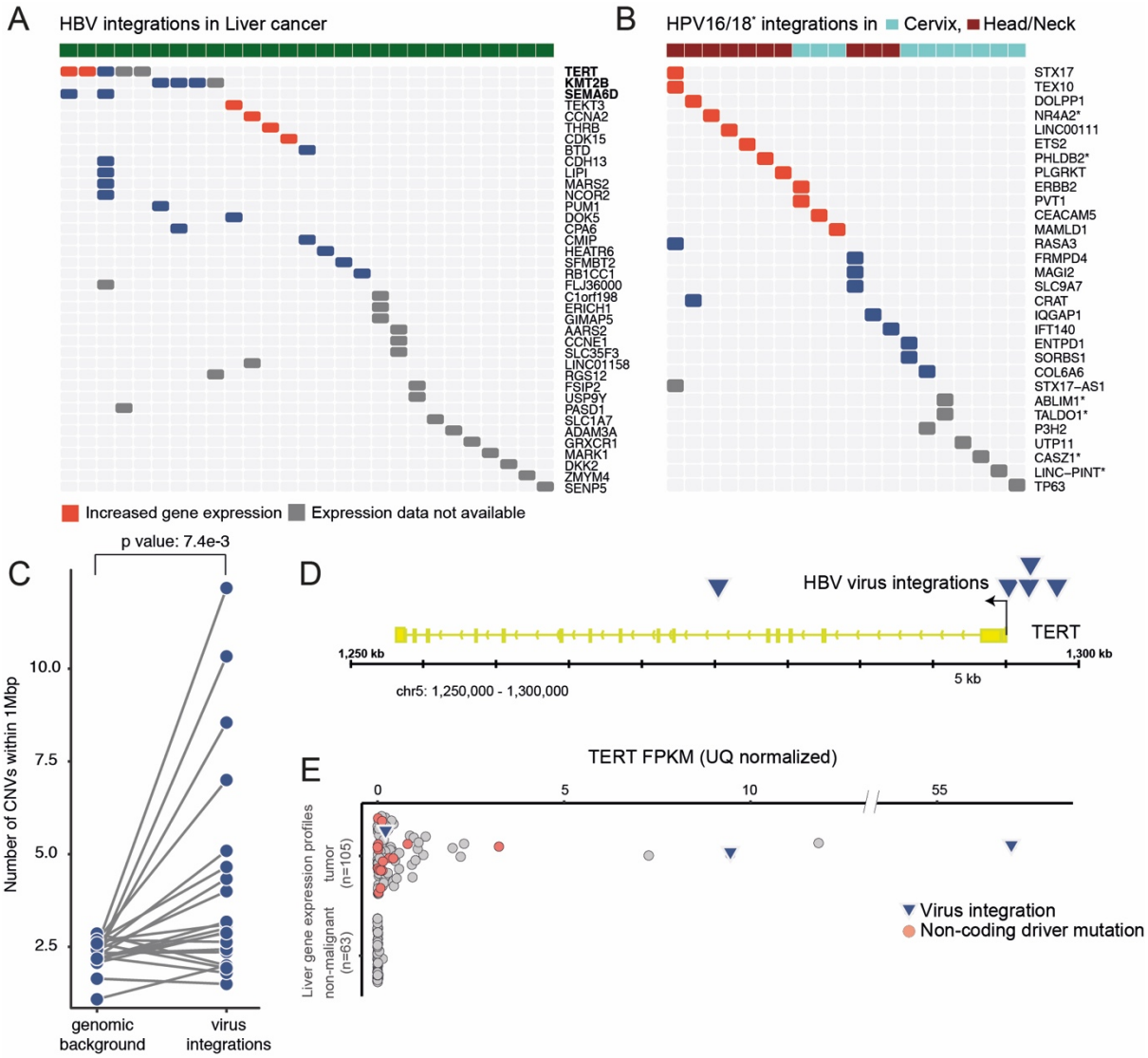

Figure 5: Impact of virus integration. (A) Integration sites detected in gene regions (including promoter, exon, intron and fiveprimeUTR regions) are labeled in red for increased gene expression and blue for expression measured. Rows of each heatmap designate nearest genes to the integration sites and columns represent individual ICGC donor and project ids. HBV integration sites detected in liver cancers (ICGC project codes: LIRI, LIHC and LINC), at the exception of TERT and SEMA6D (for which integration sites within their intergenic regions are shown as well) only genes with integration sites within their gene regions are displayed, frequency of HBV integration across samples is shown next to each gene label (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 CNVs was shown in the vicinity of HBV viral integrations (n=21). (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.

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)[4] (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 congenial facilities for capturing information not only from human cells, but also from other - potentially pathogenic - organisms or viruses present in the tissue. The by far most 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 working group “Exploratory Pathogens” searched the whole genome sequencing (WGS) and whole transcriptome sequencing (RNA-seq) data of the PCAWG consensus cohort. 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 39 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 & Discussion

Identification of tumor-associated viruses using whole genome and transcriptome sequencing data

To identify the presence of viral sequences, we exploited the WGS data of 2,834 paired tumor/normal samples across 39 cancer types, and 1,271 tumor RNA-seq data across 27 cancer types (Supplementary Table, sheet "Candidate Reads"). 182,5 billion sequences were considered for further analysis as they were not sufficiently aligned to the human reference genome in the PCAWG-generated alignment (see Materials and Methods). Remaining reads ranged from 2,000 to 800 million per WGS tumor sample and 2.9 million to 120 million per RNA-seq tumor sample (Figure 1A, Supplementary Figure 1A, B for WGS non tumor and RNAseq reads). Viral sequences were detected and quantified independently by three recently developed pathogen discovery pipelines CaPSID, P-DiP and SEPATH (see supplementary methods). The estimated relative abundance of a virus was calculated as viral reads per million extracted reads (PMER) at the genus level to improve data consistency between pipelines. To minimize the rate of false positive scores in virus detection, we applied a strict threshold of PMER>1 supported by at least three viral reads as similarly suggested by previous studies[12,13]. If a viral genus was identified by at least two of the three pipelines, it was considered present in the sample. The overlapping search spaces of the three pipelines revealed a total of 532 genera present in at least two of the pipelines (Supplementary Figure 1C). Filtering of laboratory contaminants of viral sequences was achieved through P-DiP, by examining each assembled contig of viral sequence segments for artificial, non-viral vector sequences (see Materials and Methods). The most frequent hits prone to contamination were for lambdavirus, alphabaculovirus, microvirus, simplexvirus, hepacivirus, cytomegalovirus, orthopoxvirus and punalikevirus (Figure 1B), and these were observed across many tumor types.

We generally observed a strong overlap of the genera identified across pipelines (Supplementary Figure 1D). From the whole genome dataset, we identified 286, 887 and 215 virus-tumor pairs for P-DiP, CaPSID and SEPATH, respectively (Figure 2A). 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 (Supplementary Figure 2B). The

number of hits derived from the RNA-seq dataset decreases to 112, 83, 42 virus-tumor pairs for P-DiP, CaPSID and SEPATH, respectively (Figure 2B). 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 80% 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 53% (DNA) and 80% (RNA) hits within the consensus set. While the majority of the virus hits from RNA-seq (n=61/67) were overlapping with the WGS data, the reverse is not true, 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[14]. 89% of the sequence hits detected from WGS and RNA-seq data were found to be from the virus genome type of double-stranded DNA virus (dsDNA) and dsDNA with reverse transcriptase (Figure 1C). 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 ssDNA or RNA viruses (Figure 1C).

The virome landscape across 39 distinct tumor types

We employed a consensus approach that resulted in a reliable set of 491 distinct virus-tumor pairs from WGS and RNA-seq data (Figure 2A and B, see Materials and Methods). Overall, 24 virus genera were detected across 441 tumor patients (16%). The top five most prevalent viruses account for 86% of the consensus virus hits in tumors (n=422 out of 491). 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=153 samples, 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 control samples (n=71, 2.5%) (Figure 2C); ii) orthohepadnavirus (n=66, 2.3%, e.g. hepatitis B, HBV) are as expected the most frequent among liver cancer with Hepatitis B present in 61 of 329 donors (18.5%); and iii) alphapapillomavirus (findings discussed in detail below). Lymphocryptovirus (n=11), orthohepadnavirus (n=18) and alphapapillomavirus (n=32) were detected both in RNA and DNA sequencing data (Figure 2C, left panel), with Alphapapillomavirus being the most frequent one (32 out of 37 consensus hits). This is in line with the constant expression of viral oncogenes in these viruses, a parameter supporting a direct role in carcinogenesis (see

above) [9]. In contrast, our analysis did not find any support at the RNA-seq level for the remaining common genera, such as Mastadenovirus (see below for details) or Roseolovirus.

Roseolovirus, Alphatorquevirus and Torque teno viruses show a higher number of hits in the non-malignant control samples, which are mainly derived from blood cells (Figure 2C). For example, we identified 60 patients as Roseolovirus-positive for their tumor and 86 patients positive for the non-malignant control samples. The genus Roseolovirus is composed of human herpesvirus HHV-6A, HHV-6B and HHV-7. Infections occur typically early in life and result in chronic viral latency in several cell types, mainly umbilical cord blood lymphocytes and peripheral blood mononuclear cells[16]. In our systematic study, we detected Roseolovirus mainly colon/rectum, stomach, pancreas and thyroid tumors as positive (18.3%, 17.3%, 15.5% and 10.4%). However, we could not identify actively transcribed viruses for Roseolovirus, Alphatorquevirus and Torque teno virus at the transcriptome level. This is in agreement with the latent state of these viruses reported for blood mononuclear cells[16], and their transmission through blood transfusions (e.g. alphatorquevirus and unclassified anelloviridae[17]).

Cytomegalovirus (CMV) was found both in stomach tumors (n=10) and the adjacent non-malignant tissue (n=9). CMV is linked to inflammation of the stomach or intestine [18], as well as infections of the lung and the back of the eye. Thus, in these cases, CMV was not scored as a contamination.

Interestingly, we did not identify a significant enrichment of co-infection of multiple viruses in any tumor type (Suppl. Figure 2C).

Alphapapillomavirus

Alphapapillomaviruses were mainly detected in head and neck cancer (n=18 out of 57), cervical cancer (n=19 of 20) and in two bladder cancer cases out of 23, in agreement with previous studies[5,19,20]. There is also supporting evidence for 32 out of 39 alphapapillomavirus hits in the transcriptome data (Figure 2C). We observed only one HPV subtype per tumor according to the P-DIP results. At the subtype level, HPV16 was found to be 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[13], HPV33 was identified both in head and neck (n=3) and cervix (n=1) tumor samples. On the other hand, different HPV variants, type 6 and type 45, were detected in bladder cancer.

We further characterized the functional effects of alphapapillomaviruses in tumors by integrating external PCAWG datasets such as driver mutations, mutational signatures, structural variations, gene expression profiles and patient survival. In head and neck cancer, HPV-positive tumors exhibit an almost complete mutual exclusivity with mutations in known drivers like TP53 (p=4.88e-06) (Figure 3A), as reported previously [19], which could be explained by a mutation independent inactivation of TP53 through the human papillomaviruses [21–23]. Analyzing the mutational signatures enriched in these cases, we identified mutational signatures 2 and 13 as enriched for alphapapillomavirus positive cases (Figure 3B) [24]. In addition, the expression of APOBEC3B is significantly higher in the virus positive cervix and head and neck cancers compared to their virus negative counterparts (Figure 3D) [25]. However, we did not observe the enrichment of APOBEC signatures and expression changes for EBV or mastadenovirus positive samples neither in cervix nor in other tissues.

Distinct expression profiles between virus positive and negative tumors in head and neck cancer are observed (Figure 3C) [26]. Analyzing the immune cells estimated by CiBERSORT [27], we could identify a significant increase in macrophages and T-cell signals in alphapapillomavirus positive head and neck cancers (Figure 3E). 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.

Mastadenovirus

Mastadenovirus infection is very common in humans and estimated to be responsible for between 2% and 5% of all respiratory infections. It is also responsible for mild respiratory, gastrointestinal and eye infections. Its protein E1B is known to interact and inactivate the p53 tumor suppressor protein [28,29]. Mastadenovirus sequences were detected in several tumor tissues with the highest rate in renal cell carcinoma, breast adenocarcinoma, prostate adenocarcinoma, head and neck carcinoma and hepatocellular carcinoma (Supplementary Table, sheet “Integration”). Mastadenovirus positive samples were found in both the tumor and the adjacent non-malignant tissues (Supplementary Figure 3B) across multiple TCGA sampling sites excluding a common lab-based contamination (Supplementary Figure 3A). Further supporting our finding, mastadenovirus hits can be clearly distinguished from contaminants due to higher PMER and higher sequence similarity to mastadenovirus reference sequences (Supplementary Figure 3C). Analysing the read distribution across all mastadenovirus positive samples, most of the reads were detected in the L1, IIa, L3 pV and pVI regions of the virus genomes (Supplementary Figure 3D). Notably, we identified a

considerable diversity of detected mastadenovirus contigs, supporting again the finding of independent viral infections (rather than common lab specific contaminations) (see Supplementary Figure 3F). In renal cell carcinoma of chromophore type, 21 of 45 samples were positive for mastadenovirus (Supplementary Table, sheet "Integration"). The presence of mastadenovirus was found to be significantly exclusive with mutations in common cancer drivers in Kidney cancer namely VHL, PBRM1 and SETD2 (q=1.48e-07, q=1.48e-05, q=1.84e-02, FDR=0.1). This finding hints towards a contribution of mastadenovirus infection to cancer development independent of common drivers in kidney cancer. There was a significantly better outcome of virus positive vs negative kidney cancer patients (Figure 3G, p value < 0.001, Cox proportional Hazards model). Virus positivity is a significant independent prognostic factor with tumor histological type as a clinical covariate in a multivariate Cox proportional hazard ratio model (p value = 0.00126) (Supplementary Material). The estimated number of viral copies per cell assuming diploid genomes and based on the genome fraction of mastadenovirus sequence detected ranged from 10% to 60% cells with at least one virus genome copy (Supplementary Figure 3E).

Endogenous retroviruses

Human endogenous retroviruses (HERV) are integrated in the human DNA originating from infection of germline cells by retroviruses over millions of years [30] and contribute 2.7% of the overall sequence to over 500,000 individual sites in the human genome [31,32]. 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[33] database as ERVL, ERVL-MaLR, ERV1, ERVK and ERV (Supplementary Table, sheet "HERV expression"). In comparison to the other HERV families, ERV1 shows the strongest expression on average (Figure 4A) and ERVK the highest fraction of active loci (Figure 4B). Analyzing the expression of HERVs based on the available RNA sequencing data, we could identify a strong expression for ERV1 in chronic lymphocytic leukemia compared to all other tumor tissues and adjacent normal tissues (Figure 4C). Analyzing the HERV expression in relation to patient survival, we could identify a high ERV1 expression in kidney cancer linked to worse survival outcome (Figure 4D, for other HERVs and tumor types see Supplementary Figure 4). Comparing the ERV1 expression with mastadenovirus expression in renal cancers we identified, in line with the better survival outcome of mastadenovirus positive cases, a

significantly stronger activation of ERV1 in mastadenovirus negative samples (p=0.00083, Figure 4E).

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 [34]. This process has been known best for human papilloma viruses (HPVs) in cervical, head-and-neck and several other carcinomas, and for hepatitis B virus (HBV) in liver cancer [35,36]. We searched the PCAWG genome and transcriptome cohorts for integration of those viruses that were detected by the CaPSID platform using the "Virus intEgration sites through iterative Reference SEquence customization" (VERSE) algorithm [37]. This algorithm utilizes chimeric paired-end as well as soft-clipped sequence reads to determine integration with single base-pair resolution. Detailed assessment of this algorithm (e.g. distinction from background noise) is presented in the methods section (see Materials and Methods).

Low confidence integration events were detected for the two viruses HHV4 (in 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 and HPV18 (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 (Supplementary Figure 5A-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 (see Materials and Methods) were identified in the *TERT* (ngc = 6, where ngc indicates the number of integration sites within a genomic cluster), KMT2B (ngc = 4) and *RGS12 (ngc = 3)* genes (Supplementary Figure 5E). 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*. 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* was 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 "Integration"). Recurrent integration events across samples were also observed in the *KMT2B* gene (four events). *KMT2B* was recently identified to be a likely cancer driver gene[38,39]. When comparing gene expression in samples with virus integration to those without, only *TERT* was over-expressed (fold change $\geq$ 2.0) in two liver cancer samples (Figure 5E). Additional genes with

increased expression impacted by integration events include *TEKT3, CCNA2, CDK15 and THRB* (Figure 5A).

Novel genes, which are impacted by integration events and associated with cancer include: *CDK15* that was found to be over-expressed in our study and reported to mediate resistance to the tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) [40]; *SEMA6D* identified as a potential oncogene candidate in human osteosarcoma [41]; and *CDH13* that is commonly downregulated through promoter methylation in various cancers [42]. In addition, a novel integration site located in the promoter region of *ERICH1* was detected. For further details see Supplementary Table (sheet "Integration").

There was a significant association between HBV viral integrations and SCNAs (Figure 5C). For samples with HBV integration events, the number of SCNAs was higher on average in the vicinity of viral integration sites (within 1 Mbp) when compared to samples without HBV integration (4.2 vs 2.3, p = 7e-3; two sided paired *t*-test). No evidence for an SCNA association was seen for other integrated viruses, HPV16/18 (Supplementary Figure 5 A-D).

HPV18 integration events were detected in seven tumors in total, with the most notable clusters of integration events in cervical cancer samples affecting the *TALDO1* (gc = 4) gene (Supplementary Figure 5G). As shown in Figure 5B, single viral integration events were detected in the genes *CASZ1*, *LINC-PINT*, *NR4A2*, *ABLIM1* and *PHLDB2*.

In 20 samples, HPV16 integration events were detected. Genomic clusters of viral integration sites were identified in cervical and head and neck cancer samples affecting the genes *PVT1* (gc=9), *PLGRKT* (gc=4), *ETS2* (gc=3), *LINC00111* (gc=3) and *TEXT10* (gc=3) (Supplementary Figure 5F). Additional integration events with at least two sites were detected in *CRAT*, *ERBB2*, *FRMPD4*, *MAGI2*, *MAMLD1*, *SLC9A7*, *STX17* and *TP63* genes. None of these multiple integration events were observed to recur across multiple patients (Figure 5B). Integration events were also observed in two different lncRNAs, the plasmacytoma variant translocation 1 gene (PVT1), which is recognized as an oncogenic lncRNA observed in multiple cancers including cervical carcinoma [43,44], and LINC00111, the function of which is still to be determined. Expression of both genes is strongly increased in the cases with HPV16 integration (Supplementary Figure 6F). Individual HPV16 integration sites were also found in a number of other genes including known drivers of tumor pathogenesis (*TP63*, *P3H2*, *ETS2*, *CD274* (PD-L1), *ERBB2, IQGAP1*) (see Supplementary Tables, sheet "Integration") and genes that were previously not known to be strong candidates for playing a role in

tumorigenesis (*CEACAM5, CRAT, ENTPD1-AS1, FRMPD4, MAGI2, MAMLD1, UTP11, COL6A6, RASA3, SORBS1*, *STX17-AS1, IFT140 and DOLPP1*).

Using the merged single nucleotide variant (SNV) calls from the three mutation calling pipelines (DKFZ, Broad and Sanger)[10], and by comparing samples with viral integration events to those without, 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). Interestingly the integration sites are, compared to a random genome background, enriched in close proximity (<1000bp) to common fragile sites (p = 0.0018, statistical 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, both in close vicinity and a long distance away from the integration site. Such a correlation was not seen for the integration sites of other viruses (see Supplementary Figure 6 A-E).

Finally, a single AAV2 integration event located in the intronic region of the cancer driver gene KMT2B [45] was detected in one liver cancer sample.

Identification of novel viral species or strains

The CaPSID pipeline, combining both the reference based and de novo assembly approach, was used to search for potentially novel virus genera or species. De novo analysis has generated 56 different contigs that have been classified into taxonomic groups at the genus level by the CSSSCL algorithm[46]. After filtering de novo contigs for their homology to known reference sequences, we have identified 36 contigs in 35 different tumor samples showing low sequence similarity (in average 61%) to any nucleotide sequence contained in the Blast database (see Materials and Methods). In this respect, our analysis has shown that WGS and RNA-seq can be used to identify novel isolates potentially from new viral species. However, the total number of novel isolates were quite low in comparison to viral hits to well-defined genera (Figure 2B). In addition, the findings were not enriched for a specific tumor entity but rather distributed across various cancer types including bladder, head/neck and cervical cancers (alphapapillomavirus) and many more (Supplementary Figure 7).

Conclusions

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 30x coverage, revealed considerably more virus positive cases than investigations of transcriptome data alone, which is the search space looked at in most previous virome studies. This is probably mainly due to viruses that are transcriptionally inactive in the given tumor tissue. Co-infections, generally believed to indicate a weak immune system, were very rare (Supplementary Figure 2C). This could, however, also be the result of selection processes during tumorigenesis.

Not surprisingly, known tumor associated viruses, such as EBV, HBV and HPV16/18, were among the most frequently detected targets. This is in particular 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, nurtured also 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 example, since an increased activity of TERT is a well-understood driver of cancerogenesis [47]. 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 reason 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* mutations in head and neck tumors, we could also link virus presence to an increase in mutations attributed to the mutational signatures 2 and 13[48]. These are explained by the activity of APOBEC, which – among other effects – changes viral genome sequences as a mechanism of cellular defense against viruses [49,50]. This activation could play an important role in introducing further host genome alterations and, thus, constitute an important mechanism driving tumorigenesis [25,51]. 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.

The novel finding of a tumor association of mastadenovirus needs to be further emphasized. It was only detected in the whole genome sequencing data set, which might explain why it was not linked to cancer in previous large transcriptome-based studies. Virus positive cases mainly occurred in renal cancer especially of the chromophobe subtype. Interestingly, viral infections were highly enriched in the specimen without VHL driver mutations. The viral protein E1B is known to inactivate tumor suppressor p53 [28,29] and this could be an alternative route to malignant transformation in these carcinomas. Furthermore, virus positive cases showed a significantly improved overall survival. This observation further supports a different path of tumor development and progression of these virus positive cases. Future research should decipher the link between the mastadenovirus infections and cancer in detail.

References

1. Parkin, D. M. The global health burden of infection-associated cancers in the year 2002. *Int. J. Cancer* (2006). doi:10.1002/ijc.21731
2. Plummer, M. *et al.* Global burden of cancers attributable to infections in 2012: a synthetic analysis. *Lancet Glob. Heal.* (2016). doi:10.1016/S2214-109X(16)30143-7
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. Bialecki, E. S. & Di Bisceglie, A. M. Clinical presentation and natural course of hepatocellular carcinoma. *Eur. J. Gastroenterol. Hepatol.* (2005).
5. Muñoz, N., Castellsagué, X., de González, A. B. & Gissmann, L. Chapter 1: HPV in the etiology of human cancer. *Vaccine* **24,** S1–S10 (2006).
6. Hermine, O. *et al.* Regression of Splenic Lymphoma with Villous Lymphocytes after Treatment of Hepatitis C Virus Infection. *N. Engl. J. Med.* (2002). doi:10.1056/NEJMoa013376
7. Thompson, M. P. & Kurzrock, R. Epstein-Barr Virus and Cancer. *Clinical Cancer Research* (2004). doi:10.1158/1078-0432.CCR-0670-3
8. Mesri, E. A., Feitelson, M. A. & Munger, K. Human viral oncogenesis: A cancer hallmarks analysis. *Cell Host and Microbe* (2014). doi:10.1016/j.chom.2014.02.011
9. Zur Hausen, H. Oncogenic DNA viruses. *Oncogene* (2001). doi:10.1038/sj/onc/1204958
10. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *bioRxiv* (2017). 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,** (2013).
13. Cao, S. *et al.* Divergent viral presentation among human tumors and adjacent normal tissues. *Sci. Rep.* **6,** 28294 (2016).
14. 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.* (2016). doi:10.1371/journal.ppat.1005539
15. McLaughlin-Drubin, M. E. & Munger, K. Viruses associated with human cancer.

Biochim. Biophys. Acta (2008). doi:10.1016/j.bbadis.2007.12.005

16. Krug, L. T. & Pellett, P. E. Roseolovirus molecular biology: Recent advances. *Current Opinion in Virology* (2014). doi:10.1016/j.coviro.2014.10.004
17. Spandole, S., Cimponeriu, D., Berca, L. M. & Mihăescu, G. Human anelloviruses: an update of molecular, epidemiological and clinical aspects. *Archives of Virology* (2015). doi:10.1007/s00705-015-2363-9
18. van de Berg, P. J. *et al.* Human Cytomegalovirus Induces Systemic Immune Activation Characterized by a Type 1 Cytokine Signature. *J. Infect. Dis.* (2010). doi:10.1086/655472
19. 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).
20. Li, N. *et al.* Human Papillomavirus Infection and Bladder Cancer Risk: A Meta-analysis. *J. Infect. Dis.* **204,** 217–223 (2011).
21. Travé, G. & Zanier, K. HPV-mediated inactivation of tumor suppressor p53. *Cell Cycle* **15,** 2231–2232 (2016).
22. 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).
23. 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).
24. 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).
25. Burns, M. B., Temiz, N. A. & Harris, R. S. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.* **45,** 977–983 (2013).
26. Schlecht, N. *et al.* Gene expression profiles in HPV-infected head and neck cancer. *J. Pathol.* **213,** 283–293 (2007).
27. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).
28. Yew, P. R. & Berk, A. J. Inhibition of p53 transactivation required for transformation by adenovirus early 1B protein. *Nature* **357,** 82–85 (1992).
29. Blackford, A. N. & Grand, R. J. A. Adenovirus E1B 55-Kilodalton Protein: Multiple Roles in Viral Infection and Cell Transformation. *J. Virol.* **83,** 4000–4012 (2009).
30. Nelson, P. N. *et al.* Demystified. Human endogenous retroviruses. *Mol. Pathol.* **56,** 11–18 (2003).
31. Paces, J. *et al.* HERVd: the Human Endogenous RetroViruses Database: update.

Nucleic Acids Res. **32,** D50 (2004).

32. 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).
33. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6,** 11 (2015).
34. 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).
35. Jiang, Z. *et al.* The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res.* **22,** 593–601 (2012).
36. 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).
37. 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).
38. Zhao, L.-H. *et al.* Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. *Nat. Commun.* **7,** 12992 (2016).
39. 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).
40. Park, M. H., Kim, S. Y., Kim, Y. J. & Chung, Y.-H. ALS2CR7 (CDK15) attenuates TRAIL induced apoptosis by inducing phosphorylation of survivin Thr34. *Biochem. Biophys. Res. Commun.* **450,** 129–34 (2014).
41. Moriarity, B. S. *et al.* A Sleeping Beauty forward genetic screen identifies new genes and pathways driving osteosarcoma development and metastasis. *Nat. Genet.* **47,** 615–624 (2015).
42. Ye, M. *et al.* Role of CDH13 promoter methylation in the carcinogenesis, progression, and prognosis of colorectal cancer. *Medicine (Baltimore).* **96,** e5956 (2017).
43. 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).
44. 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).
45. Nault, J.-C. *et al.* Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. *Nat. Genet.* (2015). doi:10.1038/ng.3389
46. 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).

47. Sung, W. K. *et al.* Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat. Genet.* **44,** 765–769 (2012).
48. 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).
49. Wallace, N. A. & Münger, K. The curious case of APOBEC3 activation by cancer-associated human papillomaviruses. *PLoS Pathog.* (2018). doi:10.1371/journal.ppat.1006717
50. Warren, C., Westrich, J., Doorslaer, K. & Pyeon, D. Roles of APOBEC3A and APOBEC3B in Human Papillomavirus Infection and Disease Progression. *Viruses* (2017). doi:10.3390/v9080233
51. Roberts, S. A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45,** 970–976 (2013).