A virome-wide clonal integration analysis platform for discovering cancer viral etiology

Xun Chen, Jason Kost, Arvis Sulovari, et al.

Genome Res. published online March 14, 2019
Access the most recent version at doi:10.1101/gr.242529.118

A virome-wide clonal integration analysis platform for discovering cancer viral etiology

Xun Chen[1], Jason Kost[1], Arvis Sulovari[1], Nathalie Wong[2], Winnie S. Liang[3], Jian Cao[4,5*], and Dawei Li[1,6,7*]

[1]*Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, USA*

[2]*Department of Anatomical and Cellular Pathology, Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, NT, Hong Kong, P.R. China*

[3]*Translational Genomics Research Institute, Phoenix, Arizona 85004, USA*

[4]*Division of Medical Oncology, Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, 08903, USA*

[5]*Department of Medicine, Rutgers Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, 08903, USA*

[6]*Neuroscience, Behavior, and Health Initiative, University of Vermont, Burlington, Vermont 05405, USA*

[7]*Department of Computer Science, University of Vermont, Burlington, Vermont 05405, USA*

[*]To whom correspondence should be addressed:

Dawei Li, Ph.D., Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, USA. E-mail: dawei.li@uvm.edu or Jian Cao, Ph.D., Division of Medical Oncology, Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, 08903, USA. E-mail: jian.cao@rutgers.edu

Abstract

Oncoviral infection is responsible for 12-15% of cancer in humans. Convergent evidence from epidemiology, pathology, and oncology suggest that new viral etiologies for cancers remain to be discovered. Oncoviral profiles can be obtained from cancer genome sequencing data; however, wide-spread viral sequence contamination and non-causal viruses complicate the process of identifying genuine oncoviruses. Here, we propose a novel strategy to address these challenges by performing virome-wide screening of early stage clonal viral integrations. To implement this strategy, we developed VIcaller, a novel platform for identifying viral integrations that are derived from any characterized viruses and shared by a large proportion of tumor cells using whole-genome sequencing (WGS) data. The sensitivity and precision were confirmed with simulated and benchmark cancer datasets. By applying this platform to cancer WGS datasets with proven or speculated viral etiology, we newly identified or confirmed clonal integrations of hepatitis B virus (HBV), human papillomavirus (HPV), Epstein-Barr virus (EBV), and BK Virus (BKV), suggesting the involvement of these viruses in early stages of tumorigenesis in affected tumors, such as HBV in *TERT* and *MLL4* gene loci in liver cancer, HPV and BKV in bladder cancer, and EBV in non-Hodgkin's lymphoma. We also showed the capacity of VIcaller to identify integrations from some uncharacterized viruses. This is the first study to systematically investigate the strategy and method of virome-wide screening of clonal integrations to identify oncoviruses. Searching clonal viral integrations with our platform has the capacity to identify virus-caused cancers and discover cancer viral etiologies.

Keywords: Cancer viral etiology; Oncovirus; Clonal viral integration; Integration allele fraction; High-throughput sequencing

Introduction

Oncoviral infections are responsible for some human cancers (zur Hausen 1991). To date, there are seven well-accepted human oncoviruses, including Epstein-Barr virus (EBV), human T-lymphotropic virus type 1 (HTLV-1), hepatitis B virus (HBV), human papillomavirus (HPV), hepatitis C virus (HCV), Kaposi's sarcoma-associated herpesvirus (KSHV, also known as human herpesvirus 8 or HHV8), and Merkel cell polyomavirus (MCV or MCPyV) (Moore and Chang 2010). These are collectively responsible for 12-15% of human cancers worldwide (zur Hausen 1991; Bouvard et al. 2009). Additional oncoviruses and virus-caused cancer types have been suggested by convergent evidence from epidemiological, pathological, and oncological studies (Javier and Butel 2008; zur Hausen 2009b; Moore and Chang 2010). An increasing number of cancer types have been newly found to have etiologies partially attributable to known oncoviruses such as HPV and EBV (Young and Rickinson 2004; Woodman et al. 2007) or potentially new oncoviruses. In the past, the discovery of new oncoviruses has led to novel preventative approaches, such as vaccination for HBV and HPV. These approaches have reduced the burden of related cancers, probably more than any single therapeutic treatment (Andre et al. 2008). Thus, it is crucial to uncover new cancer viral etiologies.

The list of known oncoviruses is short, partially due to challenges in identifying them. Virus-caused cancers usually have a long latency after the initial infection, from years to decades, and do not follow the principles of causality (Moore and Chang 2010). Among the seven well-established oncoviruses, three of them (HPV, KSHV, and MCV) were identified by directed searches for viral DNA sequences in tumor tissues (Cao and Li 2018), highlighting the historic successes of nucleotide sequence analysis strategies. In the era of high-throughput sequencing (HTS), the genomes of thousands of tumors have been sequenced, providing exceptional opportunities for discovering new oncoviruses and new cancer types associated with known oncoviruses. Theoretically, one or more reads containing a portion of nucleic acid sequence uniquely mapping to a viral genome can indicate the presence of a virus in a sample. However, two major challenges have prevented previous efforts from successfully identifying new cancer-associated viruses using HTS data: First, contaminating viral sequences are common in HTS data (Moustafa et al. 2017). Viral contaminations here refer to viral sequences that are not a result of infection, but are introduced during sample collection, preparation, and/or sequencing procedures.

They frequently arise from the introduction of environmental microbes, human resources, synthetic DNA (e.g., vectors) or common cell lines (Jun et al. 2012; Laurence et al. 2014; Salter et al. 2014; Strong et al. 2014; Cantalupo et al. 2015; Friis-Nielsen et al. 2016), as indicated by batch patterns (Laurence et al. 2014), such as sequencing facility-associated virus patterns (Tae et al. 2014). Second, viral infection also may have arisen from an infection occurring after tumor formation, and thus have no contribution to tumorigenesis. The presence alone of viral sequence in tumor genome sequencing data is insufficient to prove an oncogenic role, especially for viruses widely present in the population such as EBV and human herpesvirus 6.

Viral integration is the process by which a virus inserts its DNA or cDNA into its host cell's genome. Integration is a required stage in the life cycle of retroviruses (Goff 1992). Integration may also occur with non-retroviruses via mechanisms such as homologous recombination. The majority of virus-caused human tumors, including most of the tumors caused by HBV, HPV, HTLV-1, and MCV, carry multiple viral integration events in their genomes (Feng et al. 2008; Sung et al. 2012; Hu et al. 2015; Kataoka et al. 2015; Xiao et al. 2016). Thus, identifying viral integrations provides a unique opportunity to overcome the two major challenges in discovering new oncovirus candidates from HTS data: First, identifying integrated viral sequences eliminates the misidentification of contaminating viral sequences since human-virus chimeric sequences flanking integration sites, rather than virus-only sequences, are unlikely to be derived from random viral sequence contamination. Viral integrations derived from cell line contaminations can also be eliminated based on the identical viral integrations already found in cell lines, e.g., known HPV-18 integrations in HeLa cells (Cantalupo et al. 2015); Second, by further analyzing the percentage of cells with identical integrations (cellular proportion), we can determine clonal integrations that occurred during the early stages of tumorigenesis. As with somatic mutations, viruses that lead to high cellular proportion "early stage" clonal integrations are potential oncogenic drivers. Some of the integrations, e.g., those disrupting tumor suppressor genes, may have direct oncogenic roles. Thus, as a functional consequence, these integrations may confer selective growth advantages to the cells, leading to increased cellular proportion of these integrations. The presence of high cellular proportion clonal integrations suggests an oncogenic role of the identified virus. Indeed, clonal integration was considered "the strongest evidence" when determining the oncogenic role of MCV in Merkel cell carcinoma (Moore and

Chang 2010). We recently compared different methods, including genetic, molecular, and epidemiologic methods, and concluded that identifying clonal integrations using HTS was capable of providing strong genetic evidence for discovering viral etiologies (Cao and Li 2018; Chen et al. 2018; Sulovari and Li 2019). By restricting further analysis exclusively to viruses with clonal integrations, most non-causal viruses can be eliminated in a systematic search for oncoviruses. Therefore, analyzing clonal viral integrations in HTS data is an effective strategy for identifying new oncoviruses and new cancer types associated with known oncoviruses.

Results

VIcaller: a novel bioinformatics tool for virome-wide integration calling

To implement our strategy, we developed a novel bioinformatics platform, Viral Integration caller (VIcaller), for detecting virome-wide viral sequences, integration events, and fusion transcripts from HTS data. Specifically, VIcaller determines the cellular proportion for each integration to identify viruses with clonal integrations in the tumor genome as oncovirus candidates. It overcomes the two major challenges for identifying oncoviruses from tumor HTS data because (1) contaminating viral sequences are unlikely to integrate as they are not present in the live cells, and (2) non-causal viruses may integrate; however, integration events derived from non-causal viruses should be independent and sporadic, and thus are unlikely to be clonal. Only viruses that are involved in the early stages of tumorigenesis have the capacity to form clonal viral integrations, and VIcaller is capable of distinguishing them (**Figure 1A**). We first generated a comprehensive viral genome reference library containing 411,195 unique whole and partial viral genomes, representing 10,662 distinct viruses (**Supplemental Table S1**) and covering all six virus taxonomic classes, including dsDNA, ssDNA, dsRNA, ssRNA (-), ssRNA (+), and retroviruses (**Figure 1B**). This library incorporates all currently characterized viral genomes and represents the largest existing viral reference genome library.

The VIcaller pipeline consists of three phases: i) obtaining supporting reads, ii) detecting viral integrations, and iii) calculating integration allele fractions (**Figure 1C** and **Supplemental Table S2**). Three types of supporting reads were identified and used for determining integrated versus non-integrated viral sequences: 1) viral reads that can be paired-end mapped to a viral genome, 2) chimeric reads, and 3) split reads, the latter two of which can be mapped to both viral

and human genomes (see **Supplemental Table S3** for terms and abbreviations). Briefly, we first obtained the reads which could not be fully mapped to the human reference genome (unmapped reads), which included paired-end unmapped reads, one-end unmapped reads (one-end mapped reads are also extracted), and soft-clipped sequences. We then aligned these reads to the viral reference genome library to obtain all potential viral, chimeric, and split reads. After excluding reads which aligned to homologs of the human genome or contained primarily repeat sequences, we used the chimeric and split reads to determine viral species, integration status, and the upstream and downstream breakpoints on both the viral and human reference genomes (**Supplemental Fig. S1**). The integration breakpoints were fine-mapped at nucleotide resolution where possible. Each identified integration was then compared to a list of viral integrations or fusion transcripts present in commonly-used cell lines (Peter et al. 2006; Klijn et al. 2014; Cao et al. 2015; Liu et al. 2016) (**Supplemental Table S4**), such as HPV-18 integrations in HeLa cells and HPV-16 integrations in SiHa cells. Any events identical to (or within 10,000 bp of) the known integrations identified in these cell lines were annotated for further evaluation of potential cell line contaminations. The final viral integration events were summarized and reported in a uniform format (**Supplemental Table S5**). Lastly, the cellular proportion of each viral integration event was calculated based on the number of reads that supported integration and non-integration, i.e., the integration allele fraction (**Supplemental Fig. S2**). A 50% integration allele fraction represents 100% cellular proportion in diploid cells. Our comprehensive strategy considers all 12 possible scenarios of supporting reads (**Supplemental Fig. S3**) and includes all eight possible types of integration events (**Supplemental Fig. S4**), which increases the chances of discovering integrations.

Identifying viral integrations in simulated data

To evaluate the accuracy of our platform, we applied VIcaller to a series of simulated datasets. We randomly selected and fragmented over 5,000 viral sequences of various lengths from the virome-wide reference genome library and inserted them randomly into the human genome (**Supplemental Fig. S5**). An average of 3,700 substitution errors per megabase were introduced in all simulated sequence datasets. A series of viral integration-carrying HTS datasets were then simulated with read depths ranging from 1× to 150× (**Supplemental Table S6**). VIcaller was applied to each dataset and the sensitivity was calculated based on the ratio of correctly

identified versus total simulated integration events. The average numbers of chimeric and split reads for these viral integrations are shown in **Supplemental Table S**7. We found that 5× sequencing depth was sufficient to detect more than 90% of the simulated viral integrations (**Figure 2A**). When the sequencing depth was increased to 20× and 100×, VIcaller was able to capture 95% and 98% of the integrations, respectively (**Figure 2A**).

To mimic tumor purity and cancer cell heterogeneity, we varied the integration allele fractions by mixing the human genomes carrying integrations and those with no integrations at different ratios, e.g., 5%, 25%, and 50% for integration-bearing genomes. A series of HTS datasets carrying somatic viral integrations with different allele fractions were then simulated with read depths from 5× to 60× (**Supplemental Table S6**). At 30× depth, VIcaller was able to correctly detect 80% of the simulated viral integrations when the integration allele fraction was as low as 5%. Integration events with even lower abundance in tumor tissues likely do not support viral involvement in tumorigenesis. When the depth was increased to 60×, nearly all (> 98%) integrations with allele fractions > 25%, or more than 90% of integrations with allele fractions as low as 5% could be identified (**Figure 2B** and **Supplemental Table S8**). Precision was defined as the ratio of correctly identified to total identified integrations. In most analyses of the simulated datasets, the precision remained nearly 100%, i.e., only four false-positives were present in the 1,921 identified integrations, suggesting 99.8% precision (**Supplemental Table S8**). We further evaluated our approach for calculating integration allele fractions (**Supplemental Fig. S2**) using the simulated datasets. We found that regardless of sequencing depth, our calculated allele fractions were strongly correlated with the simulated values (**Figure 2C**).

We also evaluated other factors that may influence the detection power and precision for identifying integration events using HTS data. Overall, a longer insert size for paired-end reads significantly increased the detection power, particularly when it was longer than 500 bp (**Figure 2D**). Longer integrated viral sequences also increased the detection power, particularly when the integrated viral sequences were shorter than the insert size (**Figure 2E**). When the integrated viral sequences were longer than 1,000 bp, the detection power was consistently higher than 98% (**Figure 2E**). We further divided the simulated datasets into sub-groups by the size of the viral

reference genomes and then compared the detection power. VIcaller showed no bias regarding the length of the viral reference genomes (**Supplemental Fig. S6)**. In all cases, the detection power can be enhanced by increasing sequencing depth. Concurrent with the above analyses, we also tested simulated negative controls, i.e., a WGS dataset with no integrations (**Supplemental Table S6**), and VIcaller was then applied to detect integrations. As expected, we found zero evidence of integration (**Supplemental Table S9**).

We further compared VIcaller to three recently developed tools for detecting viral integrations (Chen et al. 2013; Ho et al. 2015; Wang et al. 2015). We simulated a series of datasets harboring 10 HPV-16 integrations and a series of datasets harboring 90 virome-wide viral integrations, all having varying sequencing depths and integration allele fractions (**Supplemental Table S6**). VIcaller achieved the highest detection power in identifying the 10 HPV-16 integrations under all conditions (**Figure 2F** and **Supplemental Table S10**). For detecting the virome-wide viral integrations, none of the tools except VIcaller could be used (**Figure 2G** and **Supplemental Table S10**).

Identifying HPV integrations in a cervical cancer dataset

To evaluate our VIcaller approach in identifying viral integrations from real HTS data, we first applied it to WGS data of tumor and paired normal tissues from one cervical carcinoma patient. We identified both HPV-18 integration events detected by the HPV-specific approach used in our previous study (Liang et al. 2014). Additionally, we detected three new HPV-18 integration events that were not originally detected (Liang et al. 2014) (**Figure 3A** and **Supplemental Tables S11A** and **S12**). All five integrations are present in the tumor, but not the paired normal tissue. All the newly identified integrations were validated via PCR and subsequent Sanger sequencing (**Figure 3B**).

Identifying HBV integrations in liver cancer datasets

We further applied VIcaller to an RNA-seq dataset of hepatocellular carcinoma (HCC) cell lines (Lau et al. 2014). Human HBV commonly integrates into the genomes of liver tumor cells (Brechot et al. 2000). When inserted into genic regions, the HBV genes may be expressed as host-virus fusion transcripts and contribute to tumorigenesis. By performing virome-wide

analyses of three cell lines, we detected eight out of the nine fusion transcripts reported by the HBV-specific approach used in the original study (Lau et al. 2014) (**Figure 3C**). In addition, we identified six new fusion transcripts that were not detected by the previous method (Lau et al. 2014) (**Supplemental Table S11B**). All six newly identified fusion transcripts were validated via reverse transcription PCR (RT-PCR) (**Figure 3D**) followed by Sanger sequencing (**Figure 3E**).

We then applied VIcaller to a larger WGS dataset derived from 99 HBV-associated HCC patients (Sung et al. 2012). We detected a total of 474 integration breakpoints (derived from 388 unique HBV integration events) in the 99 pairs of tumor and matched normal tissues (**Supplemental Table S11C**). Most of these integrations were derived from partial HBV genomes. For example, the average length of these integrated HBV sequences was 1,645 bp, with the minimum and maximum lengths being 59 bp and 3,201bp, respectively (**Supplemental Fig. S7** and **Supplemental Table S11C**). We randomly selected a total of 425 chimeric and split reads that supported these integrations and aligned them to the NCBI Nucleotide database using BLASTN (Camacho et al. 2009). All reads (100%) were verified (**Supplemental Table S13**). Among the 99 samples, 88 were also previously analyzed with an HBV-specific approach (Sung et al. 2012) (**Supplemental Table S14**). By comparison, among the 88 samples our virome-wide approach detected the majority (305 out of 399, 76.4%) of the 399 HBV integration breakpoints reported previously (Sung et al. 2012) (**Figure 3F** and **Supplemental Fig. S8**). A total of 94 of these previously identified breakpoints were not detected by VIcaller. However, these integration breakpoints were not fully validated, i.e., only 32 of the 399 breakpoints were selected for PCR/Sanger sequencing validation in the original study (Sung et al. 2012). Additionally, VIcaller identified 169 new HBV integration breakpoints that were not detected in the original study (Sung et al. 2012) (**Figure 3F**). For example, we found strong evidence of 19 read pairs supporting a new HBV integration event of at least 808 bp in the *HCG2032978* gene (**Figure 3G**). In support of the newly identified HBV integration events, we compared our findings to the results of an HBV sequence enrichment-based experimental analysis named HIVID (Li et al. 2013b) which was applied to 28 of the 88 tumors. Among the 28 tumors, VIcaller found 15 new integration events not detected in the original analysis, 13 of which (87%) were detected by HIVID. Seven of the 13 integrations identified by both VIcaller and HIVID, but missed in the

original study, were analyzed by PCR and Sanger sequencing (Li et al. 2013b), and all were successfully validated (**Supplemental Table S15**). On the other hand, VIcaller found none of the five false-positive HBV integrations which were detected by HIVID but not successfully validated by PCR and Sanger sequencing (Li et al. 2013b) (**Supplemental Table S16**).

Because the VIcaller approach is virome-wide and not limited to specific virus types, we also identified integrations of adeno associated virus (AAV) in two of the 99 tumor samples (**Supplemental Table S17**). One of the non-HBV integrations was derived from AAV-6 (AF028704) and was supported by 13 chimeric and split reads. It was 212 bp in length with a two bp deletion on the human genome at the breakpoint (**Figure 3H**). Our analysis revealed that the other non-HBV integration had very low abundance (**Supplemental Table S17**). These results collectively demonstrate that VIcaller can detect new viral integrations and fusion transcripts from WGS and RNA-seq data, respectively. In all, our virome-wide approach obtained significantly higher sensitivity and accuracy than the currently available virus-specific approaches.

Determining clonal HBV integrations in the *TERT* and *MLL4* genes

Because of the high sensitivity and precision of VIcaller, we have identified many additional HBV integrations not detected in the original analysis, allowing us to perform a comprehensive characterization of HBV integrations in the HCC samples. Consistent with the previous HBV-specific study (Sung et al. 2012), the HBV integrations identified by VIcaller support the presence of integration hotspots, e.g., clusters of breakpoint locations in the telomerase reverse transcriptase (*TERT*) and mixed lineage leukemia 4 (*MLL4*) genes. In the 99 liver cancer samples, we detected most of the reported *TERT* integrations (20 of 24), plus nine new *TERT* integrations (**Figure 4A**). Among the 24 integrations, nine were experimentally validated (Sung et al. 2012), all of which were detected by VIcaller. We also identified all nine reported *MLL4* integrations (100%), plus two new *MLL4* integrations (**Figure 4B**). We found that most of the *TERT* integrations (26 out of 29) were located in the *TERT* promoter region (**Figure 4A**), and almost all of the integrations contained at least one viral gene enhancer (*Enh1* or *Enh2*) or promoter (*XP*, *CP*, *SP1* or *SP2*), particularly *Enh2*, *CP* and *SP2* (**Figure 4C** and **Supplemental Table S18**). The *TERT* gene expression levels were enhanced in the patients with integrations in *TERT*

relative to those without integrations in *TERT* (Lau et al. 2014). We also found that most of the *MLL4* integrations were located between exons 3 and 6 (**Figure 4B**). The HBV integrations enhanced the *MLL4* expression levels of the exons downstream of the integration breakpoint relative to the exons upstream of the breakpoint (Dong et al. 2015), indicating a potentially truncated MLL4, presumably serving as an oncoprotein. These results suggest oncogenic functions of up-regulation of *TERT* and truncation of *MLL4* in HCC. None of the samples had integrations in both *TERT* and *MLL4*, indicating that integrations in these two genes may be mutually exclusive (Fisher's $P = 0.01$). Such mutual exclusivity usually implies functional redundancy or synthetic lethality of the two oncogenic variants.

By comparing the integration allele fractions of different integration events (**Supplemental Fig. S2**), we found that integrations in *TERT* and *MLL4* gene loci had significantly higher allele fractions compared to integrations in other parts of the genome (**Figure 4D**). Even in the same individual, the *TERT* integration had higher allele fraction relative to other integrations (paired *t*-test $P = 0.0006$ and fraction difference = 21.86 (10.56-33.17)) (**Figure 4E**). A similar trend was found for *MLL4* (**Supplemental Fig. S9**). These observations indicate that the integrations in *TERT* and *MLL4* likely occurred in the early stages of liver tumorigenesis and/or conferred selective growth advantages to these cells. In both cases, the high cellular proportion implicates an oncogenic role for the virus.

Identifying HPV and BKV clonal integrations in bladder cancer

High-risk HPV and BKV have been linked to bladder cancer. The genomic sequences and fusion transcripts of the two viruses have been found in a small percentage of bladder tumors (Tang et al. 2013; The Cancer Genome Atlas Research Network 2014b). A recent paper reported genomic integrations of HPV and BKV in four bladder cancer samples (Cantalupo et al. 2018). To further evaluate our strategy and determine whether the integrations were clonal, we applied VIcaller to the WGS data of these four bladder cancer samples as well as additional 102 randomly selected bladder cancer samples (**Supplemental Table S19**). Using our virome-wide integration screening, we found ten integration events in the four samples, but no integration events in the additional samples. One sample carried five integrations of BKV (**Figure 5A** and **Supplemental Table S11D**), and the other three carried two HPV-16, two HPV-56, and one HPV-45

integration, respectively (**Figure 5A**). These integrations were only observed in tumor tissues, but not in paired normal tissues. The integration allele fractions for the most abundant integration event in each sample were 64.7%, 48.3%, 39.0%, and 30.0%, respectively (**Figure 5A**), indicating that these integrations were shared by most cancer cells. The fact that we observed clonal integrations suggests the involvement of these viruses in the early stages of bladder tumorigenesis. All three observed HPV strains are considered high-risk HPV, capable of inducing cancers in a wide range of tissues (zur Hausen 2002). The oncogenic role of BKV has also been suggested (Abend et al. 2009), but not yet established. **Figure 5B** and **Figure 5C** show a BKV integration and an HPV-45 integration, respectively. These results show that our novel strategy and platform can be used for the detection of oncoviruses present in additional cancer types using WGS data.

Identifying EBV integrations in non-Hodgkin's lymphoma and gastric adenocarcinoma

EBV was the first identified human oncovirus: associated with cancers such as non-Hodgkin's lymphoma (Heslop 2005) and gastric cancers (Iizasa et al. 2013). EBV genomes present largely as episomes, and to date, integrations of EBV into tumor genomes were mostly observed in established cell lines, such as Raji cells (Luo et al. 2004; Cao et al. 2015; Xiao et al. 2016). By screening a WGS dataset (**Supplemental Table S20**) of 12 diffuse large B-cell lymphoma tumors (Morin et al. 2011), we detected an EBV (AB828191.1) integration in a patient (09-33003/DLBCL-PatientM). This patient had the shortest overall survival (1.31 years) among the 12 patients (the average was 4.38 years), and the difference was statistically significant (Z-test $P = 0.008$). This EBV integration was only detected in the tumor, but not in paired blood cells. The integration was 20,941 bp in length and located in an intron of the *EHD1* gene (**Figure 5D**). The allele fraction of this integration was 18%, supported by 22 chimeric and split reads. Additionally, we applied VIcaller to gastric adenocarcinomas, of which approximately 10% are EBV positive (Iizasa et al. 2013; The Cancer Genome Atlas Research Network 2014a). In 29 stomach adenocarcinoma samples (**Supplemental Table S21**), we found an EBV integration in one sample which was detected in the tumor but not in the paired normal tissue (**Supplemental Fig. S10** and **Supplemental Table S11D**). These results demonstrate that our approach can identify integrations derived from a wide range of viruses in cancer WGS data.

Determining capacity to identify integrations from uncharacterized viruses

KSHV and MCV were first isolated from tumors when searching for oncoviruses. It is likely that additional unknown viruses capable of causing cancers remain undiscovered. Viral integration screenings should not be limited to characterized viruses. Thus, we evaluated whether our platform had the capacity to identify some integrations derived from uncharacterized viruses. We first simulated a dataset containing 97 random integrations of MCV (NC_010277.2) with 50% integration allele fraction in Chr22 at 30× sequencing depth. To mimic MCV as an uncharacterized virus, we removed the MCV genome from the VIcaller virome-wide reference library. By running VIcaller on the simulated data with the MCV-depleted library, we still found 90% of the simulated MCV integration events (87 out of 97). As expected, these events were annotated as arising from viruses having the highest sequence similarity with MCV, including Gorilla gorilla gorilla polyomavirus and Pan troglodytes verus polyomavirus (**Figure 6A** and **Supplemental Table S22**). The former is known to be closely related to MCV (Leendertz et al. 2011). Similarly, we ran VIcaller on a simulated dataset containing 97 HPV-18 integrations using an HPV-18-depleted library. We detected 94% of the simulated HPV-18 integration events (91 out of 97). As expected, they were annotated as arising from other HPV strains (**Figure 6A** and **Supplemental Table S22**). Most of the detected integration allele fractions were as high as 20-30% (**Figure 6B**), allowing for most integrations to be identified as clonal or sub-clonal.

We applied a similar strategy to the four bladder cancer samples carrying HPV or BKV integrations. We aligned all supporting reads of the 10 integration events to the target virus-depleted library and still found integrations in two of the four samples (three of the 10 integration events) using the default parameters. As expected, they were annotated as from closely-related viruses. For example, the two BKV integrations were annotated as from Vervet monkey polyomavirus 2 and JC polyomavirus, respectively. However, when we applied a less stringent threshold (two or more supporting reads with a minimum alignment score of 25), we found integrations in all four samples (seven of the 10 integrations). The allele fractions of all seven identified integrations remained clonal or sub-clonal (**Supplemental Table S23**).

These results suggest that our strategy and platform have the potential capacity to capture integrations from some uncharacterized viruses, especially clonal and sub-clonal integrations,

benefiting from sequence similarity with related viruses. In future oncovirus discovery analyses, if certain non-human viruses are repeatedly identified with clonal or sub-clonal integrations, the possibility of the presence of an uncharacterized human virus should be considered because some novel oncoviruses may be missed if analyses are limited to characterized viruses.

Discussion

In this study, we first developed a novel strategy to discover oncoviruses and virus-caused tumors. Analyzing clonal integrations overcomes two major challenges: non-causal viruses and viral contaminations. Identification of clonal viral integrations was considered "the strongest evidence" in the discovery of MCV's role in Merkel cell carcinoma (Moore and Chang 2010). However, the strategy of high-throughput screening for clonal integrations has not been previously proposed or used to identify new oncovirus candidates. Second, we developed the first virome-wide clonal integration detection platform. Virome-wide analysis is necessary for identifying novel cancer viral etiologies. Our platform is also innovative because it includes integration allele fraction calculation, which provides additional supportive information to study the roles of viruses in tumorigenesis, yet, has not been utilized in previous bioinformatics analyses. Third, we have demonstrated that the strategy and platform are capable of identifying virus-caused tumors for a broad range of cancer types. By performing virome-wide screens using WGS data in several cancer types, we identified early stage clonal integrations of HBV, HPV, BKV, and EBV, which were known or speculated to be oncoviruses in related tumor types. Fourth, we demonstrated the potential capability of our platform to identify integrations derived from uncharacterized viruses based on their sequence similarity with characterized viruses, allowing for further investigation. For example, if the MCV references were removed from our virome-wide reference library, most of the simulated MCV integration events were still found as clonal or sub-clonal integrations derived from other viruses having a high degree of sequence similarity to MCV. It is worth noting that MCV's oncoviral role was discovered because a piece of its genome that had high sequence similarity with African green monkey lymphotropic polyomavirus was "fished out" in Merkel cell carcinoma (Feng et al. 2008). Thus, for future cancer genomic analyses, if there are any new integration-capable oncoviruses or new cancer types associated with known oncoviruses, application of our strategy and platform to large collections of WGS datasets will be able to identify them.

EBV infects 90% of the world population (Cohen 2000); however, only a small proportion develop EBV-associated cancers (Young et al. 2016). The oncogenic role of EBV was established based on epidemiological, pathological, and oncological evidence, such as increased risks of EBV-associated cancers in EBV-infected individuals; EBV's capacity to cause infectious mononucleosis, a self-limiting lymphoproliferative disease; and its capacity to transform human normal lymphocytes (Cao and Li 2018). Monoclonal amplification of the EBV genome in cancer cells, indicated by loss of tandem repeat polymorphism, was also used to prove EBV's oncogenic role (Raab-Traub and Flynn 1986; Weiss et al. 1989). Our clonal integration analysis has a similar concept, and it can be applied directly to HTS data. EBV was largely considered incapable of integrating into the human genome. Our findings of clonal EBV integrations in human cancer further support ERV's role as an oncovirus (at a lower frequency compared to HBV and HPV).

VIcaller, the platform that we developed for implementing this strategy, is capable of correctly detecting virome-wide integrations in WGS and fusion transcripts in RNA-seq data. VIcaller incorporates the majority of the useful functions and features from existing virus and viral integration detection software (Hawkins et al. 2011; Kostic et al. 2011; Bhaduri et al. 2012; Borozan et al. 2012; Chen et al. 2013; Li et al. 2013a; Naeem et al. 2013; Wang et al. 2013; Katz and Pipas 2014; Naccache et al. 2014; Forster et al. 2015; Ho et al. 2015; Wang et al. 2015; Liang et al. 2017; Tennakoon and Sung 2017) (**Supplemental Table S24**) as well as existing transposable element detection tools (Bourque et al. 2018; Goerner-Potvin and Bourque 2018). VIcaller implements many features, such as accurate estimation of integration allele fraction and fine-mapping of breakpoints (**Supplemental Table S25**). Because it considers all possible combinations of supporting read (**Supplemental Fig. S3**) and integration types (**Supplemental Fig. S4**), and applies stringent and comprehensive quality controls, VIcaller achieves high detection power and precision. We compared VIcaller with three recently developed software for candidate viral integration detection, and VIcaller consistently obtained the highest detection power and precision (**Supplemental Table S10** and **Supplemental Fig. S11**). VIcaller includes a step to annotate integration events that have been previously reported in commonly-used cell lines (such as the Hela and SiHa cell lines). VIcaller accepts either raw FASTQ reads or aligned

BAM files as its input and supports both single-end and paired-end sequences. VIcaller is open source and has the flexibility to be customized. All major software and utilities used by VIcaller can be easily replaced with alternative tools or updated to different versions. VIcaller is user-friendly with minimal bioinformatics skills required for users. Taken together, VIcaller can accurately detect virome-wide integrations in various HTS datasets, even for integrations with low cellular abundance and in cancer genomes with high mutation rates. The sensitivity and precision of our platform demonstrate the effectiveness of VIcaller for capturing clonal integrations present in tumor-derived WGS data. Therefore, VIcaller is a powerful tool for identifying new cancer viral etiologies. In the future, we will continue adding new viral reference genomes to our virome-wide library and will incorporate *de novo* assembly and viral sequence taxonomy prediction for the discovery of uncharacterized viruses.

In a proof-of-principle viral integration analysis, we found recurrent clonal integrations of high-risk HPV strains in bladder cancer samples. Although HPV transcripts, integrations, and fusion transcripts were previously detected in bladder cancer (Tang et al. 2013; The Cancer Genome Atlas Research Network 2014b; Cantalupo et al. 2018), the casual relationship between HPV exposure and bladder cancer has not been well-established (zur Hausen 2009a). As shown in this study, finding a high proportion of cells carrying identical HPV integrations provides strong genetic evidence to support the involvement of HPVs in the early stages of bladder tumorigenesis or the positive selection of the infected cells. Therefore, this study provides new genetic evidence in support of high-risk HPVs as oncovirus candidates in a sub-group of bladder carcinomas. Finding new oncoviruses is likely to be achieved by applying this strategy to cancer types with speculated, but undiscovered, viral etiologies. The most-recently discovered human oncovirus was identified 10 years ago from a rare skin cancer (Moore and Chang 2010). With decreasing sequencing costs, an increasing number of tumors, especially in under-studied cancer types, will be sequenced, providing unique opportunities to identify novel viral etiologies. Our method will provide one solution that can be easily adapted for future cancer genome sequencing studies. Adding clonal integration analysis to standard analysis pipelines will aid in maximizing gains from sequencing investments.

In addition to discovering oncoviruses and virus-induced tumors, VIcaller has other applications. Viral integration analysis can lead to a more in-depth cancer genome profiling. For instance, HBV integrations at certain locations, such as the promoter of *TERT* and the 5' coding region of *MLL4*, may contribute to liver tumorigenesis and play a role in clonal selection. Our analysis found that the allele fractions of integrations in five other genes: *GAS7*, *PRDM16*, *ARID1B*, *AFF1*, and *NGR1*, were also among the highest in the corresponding tumor. Three of these five genes have been reported to be associated with cancer (So et al. 2003; Huang et al. 2004; Helming et al. 2014), indicating that the other two genes might also be cancer related. Given the potential roles of viral integration in tumorigenesis, this study has shown that viral integration should be included in standard WGS data analyses along with point mutations, Indels, copy number alterations/aberrations, and chromosomal rearrangements. Furthermore, tumor growth follows a pattern of clonal evolution comprised of continuous clonal expansion, genetic diversification, and clonal selection (Greaves and Maley 2012). Viral integration sites are largely random, and most of the events are likely neutral. Continuous integrations at random locations contribute to the genomic diversity of tumor cells, and the integration sites are inherited during clonal expansion. Therefore, these integrations can also be used to study cancer heterogeneity and to track cell lineages during metastasis and drug resistance, especially when combined with single cell sequencing techniques. Cell line contamination is relatively common in HTS studies (Jun et al. 2012; Laurence et al. 2014; Salter et al. 2014; Strong et al. 2014; Cantalupo et al. 2015; Friis-Nielsen et al. 2016). Identifying viral integrations identical to the events in an established cell line at nucleotide resolution suggests contamination by this cell line. This strategy was used by VIcaller to discover and filter out this type of contamination. Currently, we include a total of 25 cell lines (Peter et al. 2006; Klijn et al. 2014; Cao et al. 2015; Liu et al. 2016) (**Supplemental Table S4**). By incorporating a comprehensive database of fine-mapped viral integrations and fusion transcripts derived from all commonly-used cell lines in the future, our platform can also be used to detect cell line contaminations. Therefore, VIcaller can be easily adapted for use in a broad range of new applications.

Methods

VIcaller pipeline. VIcaller aligns FASTQ reads to the human reference genome, using BWA-MEM (Li 2013) for WGS data and TopHat2 (Kim et al. 2013) for RNA-seq data, with default

parameters, or directly uses aligned BAM/SAM files. SAMtools (Li et al. 2009) is then used to extract paired-end unmapped reads with the parameters '-f 12 -F 256', reads having one end mapped with the parameters '-f 8 -F 260', and reads having one end unmapped with the parameters '-f 4 -F 264'. HYDRA's bamToFastq (Quinlan et al. 2010) and in-house Perl scripts are then used to convert these reads to paired-end files (**Supplemental Code**). SE-MEI (https://github.com/dpryan79/SE-MEI) is used to extract the soft-clipped sequences, i.e., unmapped sequences in mapped reads. All soft-clipped sequences that are 20 bp or longer are extracted with the parameter "-l 20", and only the obtained soft-clipped sequences from paired-end reads mapped in proper pairs are kept for subsequent analyses (**Supplemental Methods**). To eliminate low-quality data, VIcaller uses NGS QC Toolkit (Patel and Jain 2012) and Perl scripts (**Supplemental Code**) to remove low-quality nucleotides (Q20) from the end of each read, or the entire read based on its quality score (i.e., the read is shorter than 20 bp or high-quality nucleotides account for less than 80% of the read). The clean reads are then aligned to our virome-wide reference genome library using BWA-MEM with the parameters '-k 19 -c 100000 -m 50 -T 20 -h 10000 -Y -M', which are less stringent than the defaults. The putative supporting reads, including viral, chimeric, and split reads, as well as their physical locations on the human and viral genomes are obtained using Perl scripts (**Figure 1C**, **Supplemental Figs. S1** and **S3,** and **Supplemental Code**). VIcaller further re-aligns all resulting reads which are fully or partially mappable to viruses back to the human reference genome using BWA-MEM with the same (less stringent) parameters used for viral genome alignment (**Supplemental Fig. S12**). As human repetitive sequences, such as tandem and complex repeats, and other low-complexity regions, influence viral integration detection (**Supplemental Fig. S13**), as an optional step, VIcaller can use three different tools, including RepeatMasker (Smit et al. 2015), TRF (Benson 1999), and DUST (Morgulis et al. 2006) to screen all supporting reads for repeat sequences. VIcaller removes the reads that map to the human genome, which are potential virus-human homologs, with these less stringent parameters, and the reads containing shorter than 20 bp non-repetitive (uniquely mapped) sequence if there are no other supporting reads surrounding the breakpoint (i.e., within the insert size length). Lastly, the chimeric and split reads that do not follow the insert size distribution are also removed (**Supplemental Fig. S14**). The remaining reads constitute the clean supporting reads.

Viral integration status is determined primarily by the presence of chimeric and split reads (**Supplemental Fig. S15**). For each viral integration event, VIcaller identifies the physical locations of the upstream and downstream breakpoints on both the human and viral reference genomes. For integration events which are matched to multiple viruses or chromosomal locations, we determine the most likely candidate based on the number of supporting reads and their alignment scores. If the chimeric and split reads are mapped to more than one virus, this usually reflects homologous sequences present among these viruses, not random sequence matches. Considering that multiple levels of evidence are used to detect an integration, including 1) at least 20 bp mapped sequence required for each read, 2) multiple chimeric/split reads required for each breakpoint, and 3) at least one breakpoint (rather than viral sequences) required for each integration, it is extremely unlikely that random sequence matches make a significant contribution to the ranking of viral candidates. Viruses with larger genome may have increased random sequence matches; however, except for extreme situations, such ranking metrics should not be biased for larger viruses.

The details are described in **Supplemental Fig. S1**. When only one integration event exists in a genome, the viral read depth distribution will also be used as a confirmation of the detected event (**Supplemental Fig. S16**). To report a viral integration event, VIcaller requires a minimum of two chimeric and/or split read pairs with at least one uniquely mapped chimeric read pair (i.e., alignment score $\geq$ 50 if two supporting reads and alignment score $\geq$ 30 if three or more supporting reads). These criteria were proven reliable based on Sanger sequencing experiments from this study and others (Sung et al. 2012).

Quality control. To eliminate potential false-positives and low-quality viral integrations, stringent quality control metrics were developed and implemented in VIcaller, as described in detail in **Supplemental Table S26**. Furthermore, to achieve highly confident results, we used four different bioinformatics methods or settings to verify whether each of the supporting chimeric and split reads was uniquely mapped to expected human and viral genomic locations, and whether the mapped locations were consistent across different mapping tools (*in silico* validation). Only the integrations having reads uniquely mapped to both human and viral genomes, and that were consistently aligned by different alignment tools, were kept for further

analyses. Lastly, we annotated the results with a list of integrations detected in commonly-used cell lines. Similarly, we also annotated the results for potential vector contaminations with the VecScreen database.

Fine-mapping chromosomal locations of breakpoints. Both chimeric and split reads are used to fine map the physical locations of each breakpoint on the human and viral genomes (**Supplemental Fig. S17** and **Supplemental Methods**). The length of the integrated viral sequence is calculated for events with both upstream (5') and downstream (3') breakpoints detected on the viral genome.

VIcaller output. The identified viral integration events are summarized in a uniform format. Each row of this output file contains a unique viral integration event. Each event contains 27 characteristics (**Supplemental Table S5**), such as genomic location, number of supporting reads, and alignment scores.

Integration allele fraction and cellular proportion. For each viral integration event, we calculated its integration allele fraction based on the combined number of chimeric and split reads versus the number of human reads crossing the two integration breakpoints (**Supplemental Fig. S2**). The human reads were extracted from the sorted BAM files using SAMtools. Specifically, we calculate the total number of chimeric and split reads from both upstream and downstream breakpoints (denoted as "a"), which support integration; and that of the human reads crossing the breakpoint (denoted as "b"), which support non-integration at this chromosomal location. The ratio a/(a+2b) is used as the integration allele fraction (**Supplemental Fig. S2**). When only one breakpoint is identified, the ratio a/(a+b) is applied instead. A 50% integration allele fraction represents 100% cellular proportion in diploid cells.

Early stage and growth selected clonal integrations. An integration is considered "*early stage*" clonal and growth selected when its cellular proportion is reasonably high (such as 25% of the cells, considering various adjacent normal cells in the tumor) in the same specimen. By comparing the cellular proportions of clonal integrations with those of known tumorigenic somatic mutations, we can identify integrations that occurred in the early stages of tumorigenesis

and conferred selective growth advantages to these cells. Viruses occurring in multiple samples with clonal integrations are considered oncovirus candidates for further evaluation.

Software availability

VIcaller is an open-source software. VIcaller v1.1 source code, documentation, and example data are available at www.uvm.edu/genomics/software/VIcaller.html, and https://github.com/daweili-lab/VIcaller. The source code is also available in Supplemental Code.

Author contributions

D.L. and J.C. initiated and designed the project. X.C., D.L., and J.K. analyzed the results. X.C., W.L. and N.W. performed the experimental work. X.C. developed the software code. D.L., J.C. and X.C. wrote the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Acknowledgments

This work was supported by The University of Vermont Start-up Fund (to D. L.), the Institutional Research Grant 14-196-01 (126773-IRG) from the American Cancer Society (to D. L.), and the Career Development Award from Melanoma Research Foundation (to J. C.). The authors thank all the specimen donors. The authors thank the research groups, including TCGA and CGCI, and many other investigators who have contributed samples and data. The sequence datasets were obtained from the database of Genotypes and Phenotypes, Sequence Read Archive, and European Nucleotide Archive. The authors acknowledge the Vermont Advanced Computing Core and the Massachusetts Green High-Performance Computer C3DDB for providing computing resources. The authors thank Dr. Stephen Everse for his help with the access to computing resources. The authors thank Dr. Chao Cheng, Dr. Guangchen Liu, and Michael Mariani for their discussions on the research. The authors also thank Drs. Daniel DiMaio, Evan Eichler, Seth Frietze, Jinshun Zhong, and many others for their careful reviews of the manuscript. The authors also thank all the anonymous reviewers for their constructive comments and suggestions.

References

Abend JR, Jiang M, Imperiale MJ. 2009. BK virus and human cancer: innocent until proven guilty. *Seminars in cancer biology* **19**: 252-260.

Andre FE, Booy R, Bock HL, Clemens J, Datta SK, John TJ, Lee BW, Lolekha S, Peltola H, Ruff TA et al. 2008. Vaccination greatly reduces disease, disability, death and inequity worldwide. *Bull World Health Organ* **86**: 140-146.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573-580.

Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA. 2012. Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics* **28**: 1174-1175.

Borozan I, Wilson S, Blanchette P, Laflamme P, Watt SN, Krzyzanowski PM, Sircoulomb F, Rottapel R, Branton PE, Ferretti V. 2012. CaPSID: a bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. *BMC bioinformatics* **13**: 206.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvak Z, Levin HL, Macfarlan TS et al. 2018. Ten things you should know about transposable elements. *Genome Biol* **19**: 199.

Bouvard V, Baan R, Straif K, Grosse Y, Secretan B, Ghissassi FE, Benbrahim-Tallaa L, Guha N, Freeman C, Galichet L et al. 2009. A review of human carcinogens—Part B: biological agents. *The Lancet Oncology* **10**: 321-322.

Brechot C, Gozuacik D, Murakami Y, Paterlini-Brechot P. 2000. Molecular bases for the development of hepatitis B virus (HBV)-related hepatocellular carcinoma (HCC). *Seminars in cancer biology* **10**: 211-231.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Cantalupo PG, Katz JP, Pipas JM. 2015. HeLa nucleic acid contamination in the cancer genome atlas leads to the misidentification of human papillomavirus 18. *J Virol* **89**: 4051-4057.

Cantalupo PG, Katz JP, Pipas JM. 2018. Viral sequences in human cancer. *Virology* **513**: 208-216.

Cao J, Li D. 2018. Searching for human oncoviruses: Histories, challenges, and opportunities. *J Cell Biochem* **119**: 4897-4906.

Cao S, Strong MJ, Wang X, Moss WN, Concha M, Lin Z, O'Grady T, Baddoo M, Fewell C, Renne R et al. 2015. High-throughput RNA sequencing-based virome analysis of 50 lymphoma cell lines from the Cancer Cell Line Encyclopedia project. *J Virol* **89**: 713-729.

Chen X, Kost J, Li D. 2019. Comprehensive comparative analysis of methods and software for identifying viral integrations. *Briefings in Bioinformatics*: doi: 10.1093/bib/bby1070. In press

Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X. 2013. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics* **29**: 266-267.

Cohen JI. 2000. Epstein-Barr virus infection. *N Engl J Med* **343**: 481-492.

Dong H, Zhang L, Qian Z, Zhu X, Zhu G, Chen Y, Xie X, Ye Q, Zang J, Ren Z et al. 2015. Identification of HBV-MLL4 Integration and Its Molecular Basis in Chinese Hepatocellular Carcinoma. *PLoS One* **10**: e0123175.

Feng H, Shuda M, Chang Y, Moore PS. 2008. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* **319**: 1096-1100.

Forster M, Szymczak S, Ellinghaus D, Hemmrich G, Ruhlemann M, Kraemer L, Mucha S, Wienbrandt L, Stanulla M, Group UFOSCwI-BS et al. 2015. Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. *Sci Rep* **5**: 11534.

Friis-Nielsen J, Kjartansdottir KR, Mollerup S, Asplund M, Mourier T, Jensen RH, Hansen TA, Rey-Iglesia A, Richter SR, Nielsen IB et al. 2016. Identification of Known and Novel Recurrent Viral Sequences in Data from Multiple Patients and Multiple Cancers. *Viruses* **8**.

Goerner-Potvin P, Bourque G. 2018. Computational tools to unmask transposable elements. *Nat Rev Genet* **19**: 688-704.

Goff SP. 1992. Genetics of retroviral integration. *Annu Rev Genet* **26**: 527-544.

Greaves M, Maley CC. 2012. Clonal evolution in cancer. *Nature* **481**: 306-313.

Hawkins TB, Dantzer J, Peters B, Dinauer M, Mockaitis K, Mooney S, Cornetta K. 2011. Identifying viral integration sites using SeqMap 2.0. *Bioinformatics* **27**: 720-722.

Helming KC, Wang X, Wilson BG, Vazquez F, Haswell JR, Manchester HE, Kim Y, Kryukov GV, Ghandi M, Aguirre AJ et al. 2014. ARID1B is a specific vulnerability in ARID1A-mutant cancers. *Nature medicine* **20**: 251-254.

Heslop HE. 2005. Biology and Treatment of Epstein-Barr Virus–Associated Non-Hodgkin Lymphomas. *ASH Education Program Book* **2005**: 260-266.

Ho DWH, Sze KMF, Ng IOL. 2015. Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. *Oncotarget* **6**: 20959-20963.

Hu Z, Zhu D, Wang W, Li W, Jia W, Zeng X, Ding W, Yu L, Wang X, Wang L et al. 2015. 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.

Huang HE, Chin SF, Ginestier C, Bardou VJ, Adelaide J, Iyer NG, Garcia MJ, Pole JC, Callagy GM, Hewitt SM et al. 2004. A recurrent chromosome breakpoint in breast cancer at the NRG1/neuregulin 1/heregulin gene. *Cancer Res* **64**: 6840-6844.

Iizasa H, Nanbo A, Nishikawa J, Jinushi M, Yoshiyama H. 2013. Epstein-Barr Virus (EBV)-associated gastric carcinoma. *Viruses* **4**: 3420-3439.

Javier RT, Butel JS. 2008. The History of Tumor Virology. *Cancer Research* **68**: 7693-7706.

Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM. 2012. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American journal of human genetics* **91**: 839-848.

Kataoka K, Nagata Y, Kitanaka A, Shiraishi Y, Shimamura T, Yasunaga J, Totoki Y, Chiba K, Sato-Otsubo A, Nagae G et al. 2015. Integrated molecular analysis of adult T cell leukemia/lymphoma. *Nat Genet* **47**: 1304-1315.

Katz JP, Pipas JM. 2014. SummonChimera infers integrated viral genomes with nucleotide precision from NGS data. *BMC bioinformatics* **15**: 348.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**: R36.

Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H, Degenhardt J, Mayba O, Gnad F, Liu J et al. 2014. A comprehensive transcriptional portrait of human cancer cell lines. *Nature biotechnology* **33**: 306.

Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M. 2011. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nature biotechnology* **29**: 393-396.

Lau CC, Sun T, Ching AK, He M, Li JW, Wong AM, Co NN, Chan AW, Li PS, Lung RW et al. 2014. Viral-human chimeric transcript predisposes risk to liver cancer development and progression. *Cancer Cell* **25**: 335-349.

Laurence M, Hatzis C, Brash DE. 2014. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One* **9**: e97876.

Leendertz FH, Scuda N, Cameron KN, Kidega T, Zuberbuhler K, Leendertz SA, Couacy-Hymann E, Boesch C, Calvignac S, Ehlers B. 2011. African great apes are naturally infected with polyomaviruses closely related to Merkel cell polyomavirus. *J Virol* **85**: 916-924.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997*.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li JW, Wan R, Yu CS, Co NN, Wong N, Chan TF. 2013a. ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics* **29**: 649-651.

Li W, Zeng X, Lee NP, Liu X, Chen S, Guo B, Yi S, Zhuang X, Chen F, Wang G et al. 2013b. HIVID: an efficient method to detect HBV integration using low coverage sequencing. *Genomics* **102**: 338-344.

Liang WS, Aldrich J, Nasser S, Kurdoglu A, Phillips L, Reiman R, McDonald J, Izatt T, Christoforides A, Baker A et al. 2014. Simultaneous characterization of somatic events and HPV-18 integration in a metastatic cervical carcinoma patient using DNA and RNA sequencing. *Int J Gynecol Cancer* **24**: 329-338.

Liang Y, Qiu K, Liao B, Zhu W, Huang X, Li L, Chen X, Li K. 2017. Seeksv: an accurate tool for somatic structural variation and virus integration detection. *Bioinformatics* **33**: 184-191.

Liu Y, Lu Z, Xu R, Ke Y. 2016. Comprehensive mapping of the human papillomavirus (HPV) DNA integration sites in cervical carcinomas by HPV capture technology. *Oncotarget* **7**: 5852-5864.

Luo W-J, Takakuwa T, Ham MF, Wada N, Liu A, Fujita S, Sakane-Ishikawa E, Aozasa K. 2004. Epstein-Barr virus is integrated between REL and BCL-11A in American Burkitt lymphoma cell line (NAB-2). *Lab Invest* **84**: 1193-1199.

Moore PS, Chang Y. 2010. Why do viruses cause cancer? Highlights of the first century of human tumour virology. *Nat Rev Cancer* **10**: 878-889.

Morgulis A, Gertz EM, Schaffer AA, Agarwala R. 2006. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol* **13**: 1028-1040.

Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, Corbett RD, Johnson NA, Severson TM, Chiu R, Field M et al. 2011. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**: 298-303.

Moustafa A, Xie C, Kirkness E, Biggs W, Wong E, Turpaz Y, Bloom K, Delwart E, Nelson KE, Venter JC et al. 2017. The blood DNA virome in 8,000 humans. *PLoS Pathog* **13**: e1006292.

Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, Bouquet J, Greninger AL, Luk KC, Enge B et al. 2014. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res* **24**: 1180-1192.

Naeem R, Rashid M, Pain A. 2013. READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. *Bioinformatics* **29**: 391-392.

Patel RK, Jain M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* **7**: e30619.

Peter M, Rosty C, Couturier J, Radvanyi F, Teshima H, Sastre-Garau X. 2006. MYC activation associated with the integration of HPV DNA at the MYC locus in genital tumors. *Oncogene* **25**: 5985.

Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, Hall IM. 2010. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* **20**: 623-635.

Raab-Traub N, Flynn K. 1986. The structure of the termini of the Epstein-Barr virus as a marker of clonal cellular proliferation. *Cell* **47**: 883-889.

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* **12**: 87.

Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015. *Institute for Systems Biology http://repeatmasker.org*.

So CW, Karsunky H, Passegue E, Cozzio A, Weissman IL, Cleary ML. 2003. MLL-GAS7 transforms multipotent hematopoietic progenitors and induces mixed lineage leukemias in mice. *Cancer Cell* **3**: 161-171.

Strong MJ, Xu G, Morici L, Splinter Bon-Durant S, Baddoo M, Lin Z, Fewell C, Taylor CM, Flemington EK. 2014. Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS Pathog* **10**: e1004437.

Sulovari A, Li D. 2019. VIpower: a simulation-based tool for estimating power of viral integration detection via high-throughput sequencing. *Genomics* **In press**.

Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y, Lee NP, Lee WH, Ariyaratne PN, Tennakoon C et al. 2012. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat Genet* **44**: 765-769.

Tae H, Karunasena E, Bavarva JH, McIver LJ, Garner HR. 2014. Large scale comparison of non-human sequences in human sequencing data. *Genomics* **104**: 453-458.

Tang KW, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E. 2013. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat Commun* **4**: 2513.

Tennakoon C, Sung WK. 2017. BATVI: Fast, sensitive and accurate detection of virus integrations. *BMC bioinformatics* **18**: 71.

The Cancer Genome Atlas Research Network. 2014a. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**: 202-209.

The Cancer Genome Atlas Research Network. 2014b. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**: 315-322.

Wang Q, Jia P, Zhao Z. 2013. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS One* **8**: e64465.

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

Weiss LM, Movahed LA, Warnke RA, Sklar J. 1989. Detection of Epstein-Barr viral genomes in Reed-Sternberg cells of Hodgkin's disease. *N Engl J Med* **320**: 502-506.

Woodman CBJ, Collins SI, Young LS. 2007. The natural history of cervical HPV infection: unresolved issues. *Nat Rev Cancer* **7**: 11-22.

Xiao K, Yu Z, Li X, Li X, Tang K, Tu C, Qi P, Liao Q, Chen P, Zeng Z et al. 2016. Genome-wide Analysis of Epstein-Barr Virus (EBV) Integration and Strain in C666-1 and Raji Cells. *J Cancer* **7**: 214-224.

Young LS, Rickinson AB. 2004. Epstein-Barr virus: 40 years on. *Nat Rev Cancer* **4**: 757-768.

Young LS, Yap LF, Murray PG. 2016. Epstein-Barr virus: more than 50 years old and still providing surprises. *Nat Rev Cancer* **16**: 789-802.

zur Hausen H. 1991. Viruses in human cancers. *Science* **254**: 1167-1173.

zur Hausen H. 2002. Papillomaviruses and cancer: from basic studies to clinical application. *Nat Rev Cancer* **2**: 342-350.

zur Hausen H. 2009a. Papillomaviruses in the causation of human cancers - a brief historical account. *Virology* **384**: 260-265.

zur Hausen H. 2009b. The search for infectious causes of human cancers: where and why (Nobel lecture). *Angew Chem Int Ed Engl* **48**: 5798-5808.

Figure Legends

Figure 1 Discovering oncovirus candidates through identification of clonal viral integrations. **(A)** Identifying clonal viral integrations to eliminate viral sequence contaminations and to prove the involvement of the identified virus in the early stages of tumorigenesis. **(B)** Composition of VIcaller virome-wide genome reference library. **(C)** The simplified analytic workflow of VIcaller.

Figure 2 Applying VIcaller to simulation datasets. **(A)** Detection power and precision were measured by simulated (germline) viral integrations with depths from 1× to 150×. **(B)** Detection power for integrations with 5%, 25% and 50% integration allele fractions. An average of 86 viral integrations were used for the calculation. **(C)** Accuracy of calculated integration allele fractions. **(D)** Relationship between detection power and insert sizes for paired-end sequence reads at different sequencing depths. **(E)** Relationship between detection power and lengths of integrated viral sequences. The viral integrations detected under different sequencing depths were combined for the calculation. Comparison of the detection power of VIcaller with existing tools for detecting 10 simulated HPV integrations **(F)** and 90 simulated virome-wide integrations **(G)**. VirusSeq was only capable of detecting less than 20 human viruses; thus, the detection power was extremely low. It also ran out of sever wall time at 60× sequencing depth. VirusFinder and Virus-Clip were not applicable for analyzing data containing the virome-wide integrations.

Figure 3 Virome-wide integrations detected in liver and cervical cancer genome datasets. **(A)** Comparison of the number of integration events identified in a metastatic cervical carcinoma sample by our VIcaller approach (light blue) and the HPV-specific approach of the original study (Liang et al. 2014) (light grey). **(B)** Sanger sequencing result for one example of the three HPV-18 integrations, newly detected by VIcaller, that existed in the tumor but not in the paired normal tissue. A 16 bp deletion on the human genome was found at the integration breakpoint. **(C)** Comparison of the number of HBV-human fusion transcripts identified in three HCC cell lines by the VIcaller virome-wide approach (light blue) and the HBV-specific approach described in the original study (Lau et al. 2014) (light grey). **(D)** Gel images from RT-PCR validation of the six fusion transcripts newly detected by VIcaller. **(E)** Sanger sequencing result of an example breakpoint of the six newly identified fusion transcripts. **(F)** Comparison of the number of HBV

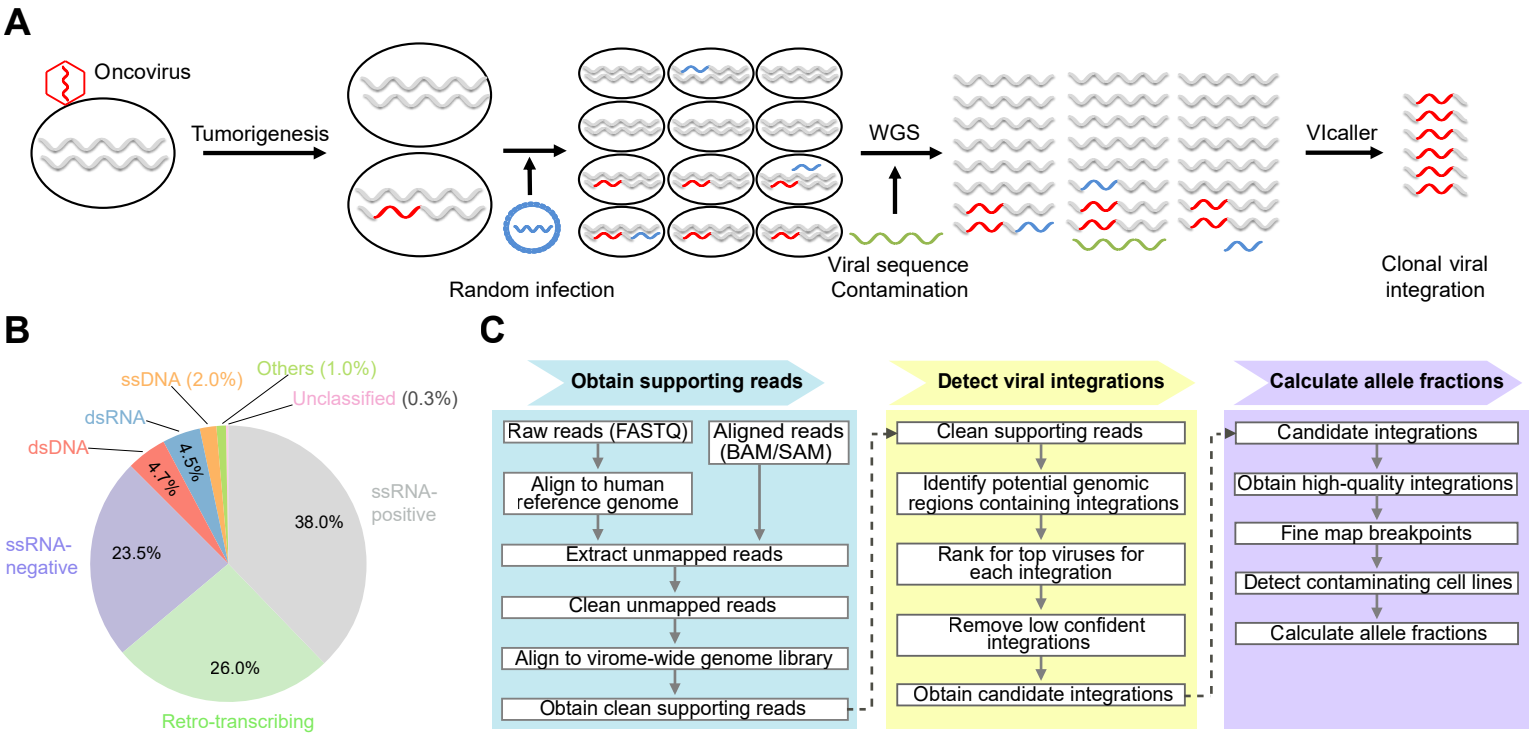
integration breakpoints identified in 88 HCC samples by our VIcaller virome-wide approach (light blue) and the HBV-specific approach described in the original study (Sung et al. 2012) (light grey). **(G)** Sequence read alignment of an HBV integration in the *HCG2032978* gene, newly identified by VIcaller, which existed in the tumor but not in the paired normal tissue. Seven chimeric and seven split reads at the upstream breakpoint and four chimeric reads at the downstream breakpoint were found for this integration event. The integrated HBV sequence is ~808 bp in length, starting from 3,170 bp to 3,182 bp, and then from one bp to ~796 bp on the circular HBV genome. Black and red represent reads mapped to the human (hg19) and HBV (NC_003977.2) reference genomes, respectively. **(H)** Sequence read alignment of an adeno associated virus 6 (AAV-6; AF028704.1) integration event detected by VIcaller (sample ID: 55T) that existed in the tumor but not in the paired normal tissue. Eight chimeric and five split reads were found across the two breakpoints. This integration is 212 bp in length, from 54 bp to 266 bp on the AAV-6 genome.

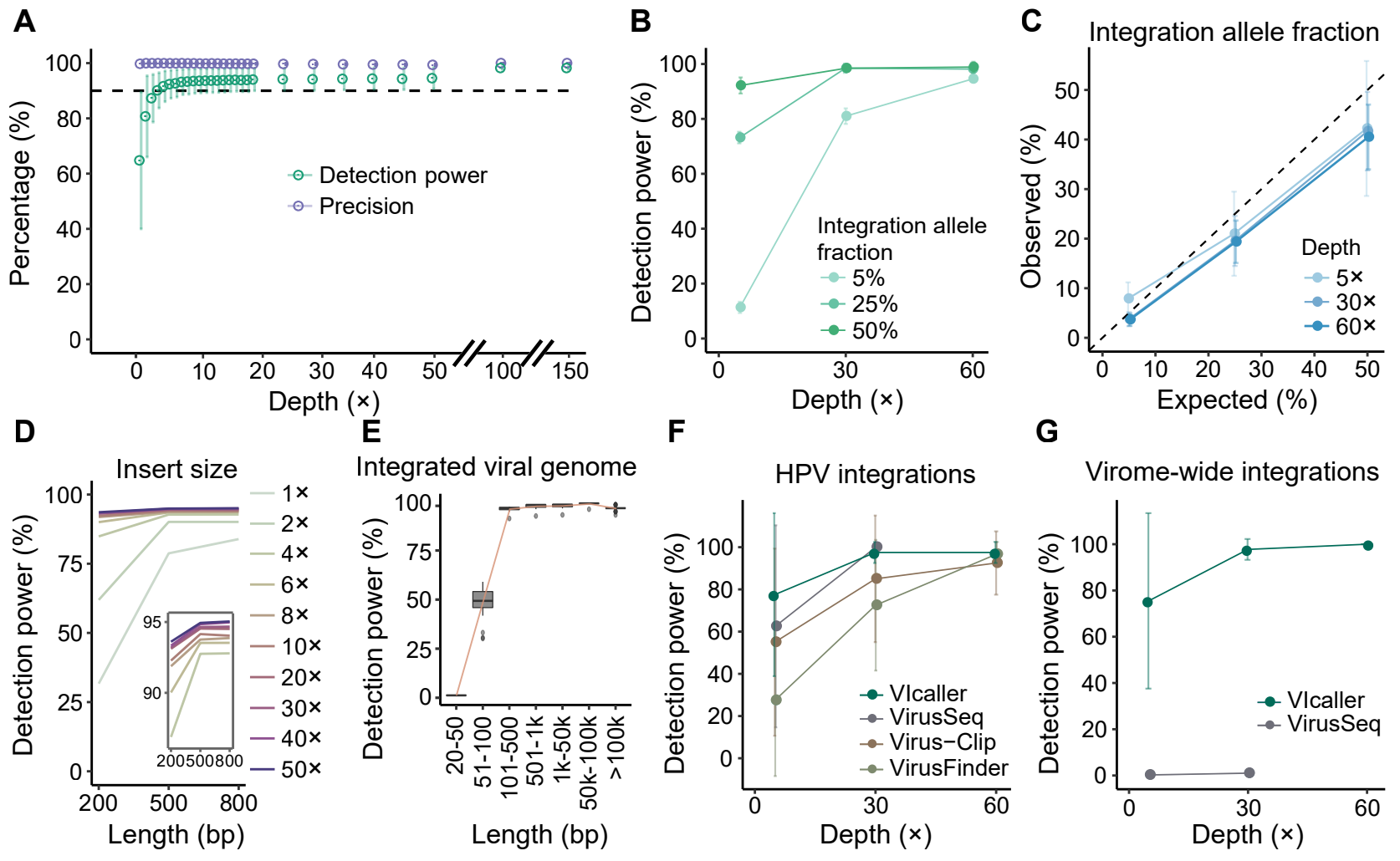
Figure 4 Characteristics of HBV integrations identified in tumors. **(A** and **B)** Sites of HBV integrations in two oncogenes **(A)** *TERT* and **(B)** *MLL4*. **(C)** The integrated HBV sequences in *TERT*. The solid red lines above the HBV genome represent the integrated sequences with both breakpoints identified while the dotted lines represent those with only one breakpoint identified. The HBV genes are in grey while the promoters and enhancers are in red. **(D)** Comparison of integration allele fractions among HBV integrations in *TERT*, *MLL4*, and other chromosomal regions. **(E)** Integration allele fraction comparison of all HBV integrations in the samples with integrations in *TERT*. The top shows the highest integration allele fraction in each sample. The bottom-left shows that all HBV integrations in each sample, including those in *TERT*, and other regions (except *MLL4*). The bottom-right shows the violin plot distributions of allele fractions of integrations in *TERT* compared to those in other regions (except *MLL4*). The result for *MLL4* is shown in **Supplemental Fig. S9**.

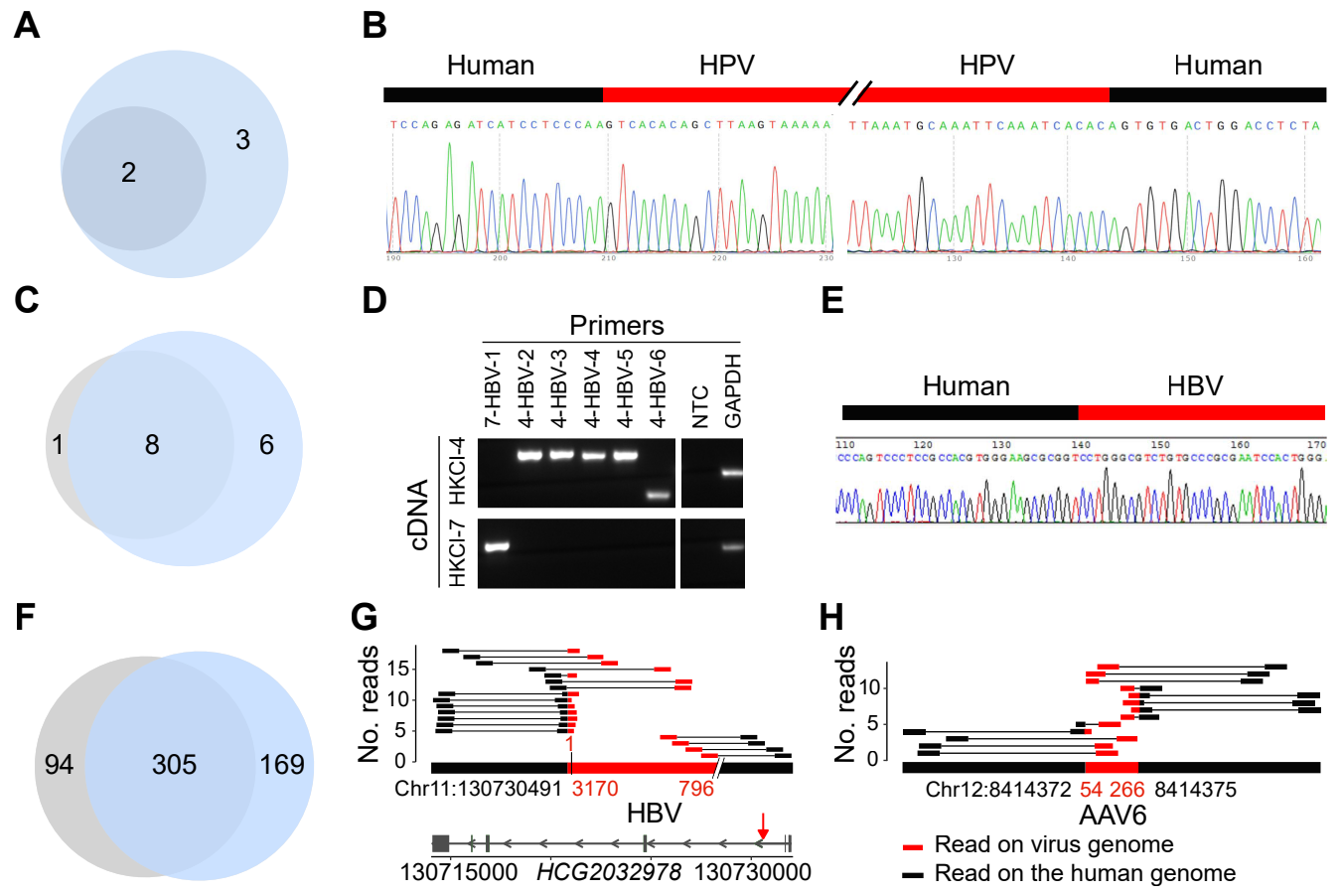
Figure 5 Identifying oncovirus candidates with integrations in bladder cancer, diffuse large B-cell lymphoma, and gastric adenocarcinoma samples. **(A)** Summary of identified integrations. BLCA: TCGA Urothelial Bladder Carcinoma; STAD: TCGA Stomach Adenocarcinoma; DLBCL: The Cancer Genome Characterization Initiative Diffuse Large B-Cell Lymphoma. **(B)**

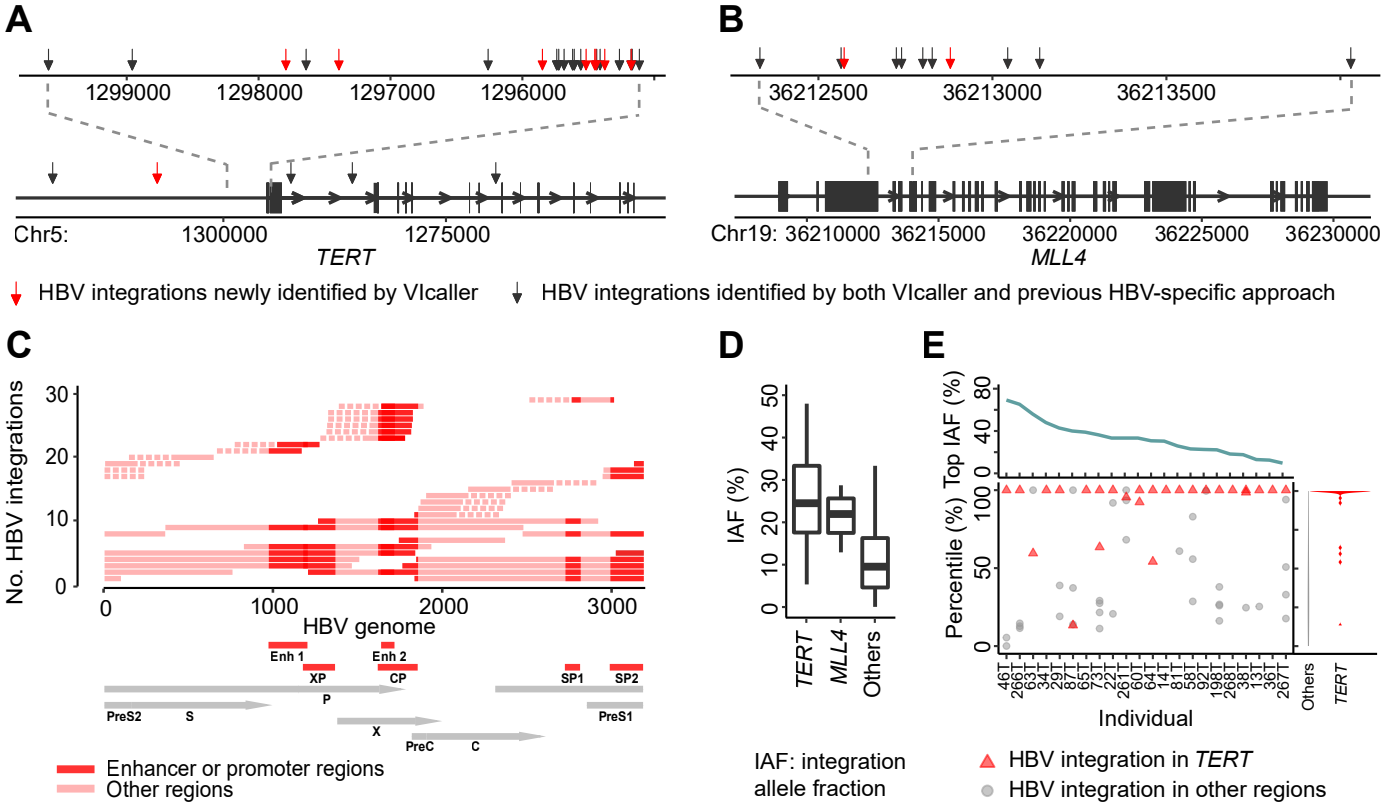
Sequence read alignment of a BKV (AB485698.1) integration event with 39% integration allele fraction detected in a bladder cancer sample TCGA-DK-A3IT. A total of 23 supporting read pairs, including 19 chimeric and four split reads, were found crossing the two breakpoints, supporting an integration, while 18 read pairs that support no integration were fully mapped to the human reference genome. **(C)** Sequence read alignment of an HPV-45 (EF202163.1) integration event with 30% integration allele fraction detected in a bladder cancer sample TCGA-BT-A20V. A total of 12 supporting read pairs, including 11 chimeric reads and one split read, were found crossing the two breakpoints, supporting an integration; while 14 read pairs were fully mapped to the human reference genome, supporting no integration. **(D)** Sequence read alignment of an EBV (AB828191.1) integration event with 18.6% integration allele fraction detected in a diffuse large B-cell lymphoma sample 09-33003. A total of 22 supporting read pairs, including 18 chimeric and four split reads, were found crossing the two breakpoints, supporting an integration, while 48 read pairs that support no integration were fully mapped to the human reference genome.

Figure 6 Viruses and integration events detected after removing the target viral genomes from our virome-wide database. **(A)** Percentage of simulated integrations detected after removing the HPV-18 (left) or MCV (right) references. **(B)** Integration allele fraction detected after removing the HPV-18 (left) or MCV (right) references. Three detected integration events that had > 50% fractions are not shown in the figure, including two among the 91 HPV-18 events, and one among the 87 MCV events.


A
Oncovirus
Tumorigenesis
Random infection
WGS
Viral sequence
Contamination
VIcaller
Clonal viral
integration
B
ssDNA (2.0%)
Others (1.0%)
Unclassified (0.3%)
dsRNA
dsDNA
4.5%
4.7%
38.0%
23.5%
26.0%
ssRNA-
positive
ssRNA-
negative
Retro-transcribing
C
Obtain supporting reads
Raw reads (FASTQ)
Aligned reads
(BAM/SAM)
Align to human
reference genome
Extract unmapped reads
Clean unmapped reads
Align to virome-wide genome library
Obtain clean supporting reads
Detect viral integrations
Clean supporting reads
Identify potential genomic
regions containing integrations
Rank for top viruses for
each integration
Remove low confident
integrations
Obtain candidate integrations
Calculate allele fractions
Candidate integrations
Obtain high-quality integrations
Fine map breakpoints
Detect contaminating cell lines
Calculate allele fractions


A
Percentage (%)
Detection power
Precision
Depth (×)
B
Detection power (%)
Integration allele fraction
5%
25%
50%
Depth (×)
C
Integration allele fraction
Observed (%)
Depth
5×
30×
60×
Expected (%)
D
Insert size
Detection power (%)
1×
2×
4×
6×
8×
10×
20×
30×
40×
50×
Length (bp)
E
Integrated viral genome
Detection power (%)
20-50
51-100
101-500
501-1k
1k-50k
50k-100k
>100k
Length (bp)
F
HPV integrations
Detection power (%)
VIcaller
VirusSeq
Virus−Clip
VirusFinder
Depth (×)
G
Virome-wide integrations
Detection power (%)
VIcaller
VirusSeq
Depth (×)

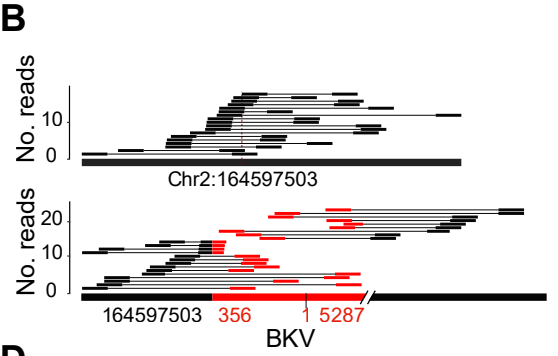

A
2
3
B
Human
HPV
HPV
Human
C
1
8
6
D
Primers
7-HBV-1
4-HBV-2
4-HBV-3
4-HBV-4
4-HBV-5
4-HBV-6
NTC
GAPDH
cDNA
HKCI-4
HKCI-7
E
Human
HBV
F
94
305
169
G
No. reads
Chr11:130730491
3170
796
HBV
130715000
HCG2032978
130730000
H
No. reads
Chr12:8414372
54 266
8414375
AAV6
Read on virus genome
Read on the human genome

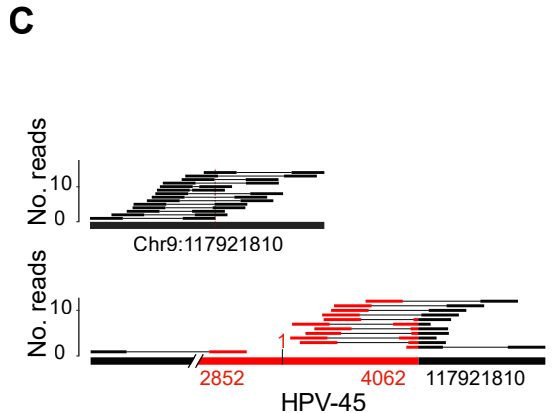

A
1299000
1298000
1297000
1296000
Chr5:
1300000
1275000
TERT
B
36212500
36213000
36213500
Chr19: 36210000
36215000
36220000
36225000
36230000
MLL4
HBV integrations newly identified by VIcaller
HBV integrations identified by both VIcaller and previous HBV-specific approach
C
No. HBV integrations
HBV genome
Enh 1
Enh 2
XP
CP
SP1
SP2
P
PreS2
S
X
PreS1
PreC
C
Enhancer or promoter regions
Other regions
D
IAF (%)
TERT
MLL4
Others
IAF: integration allele fraction
E
Percentile (%) Top IAF (%)
Individual
Others
TERT
HBV integration in TERT
HBV integration in other regions

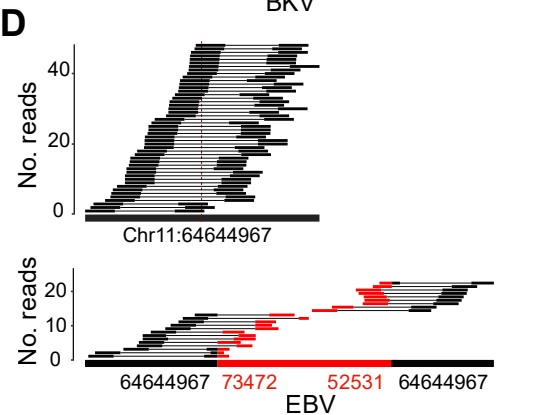
A

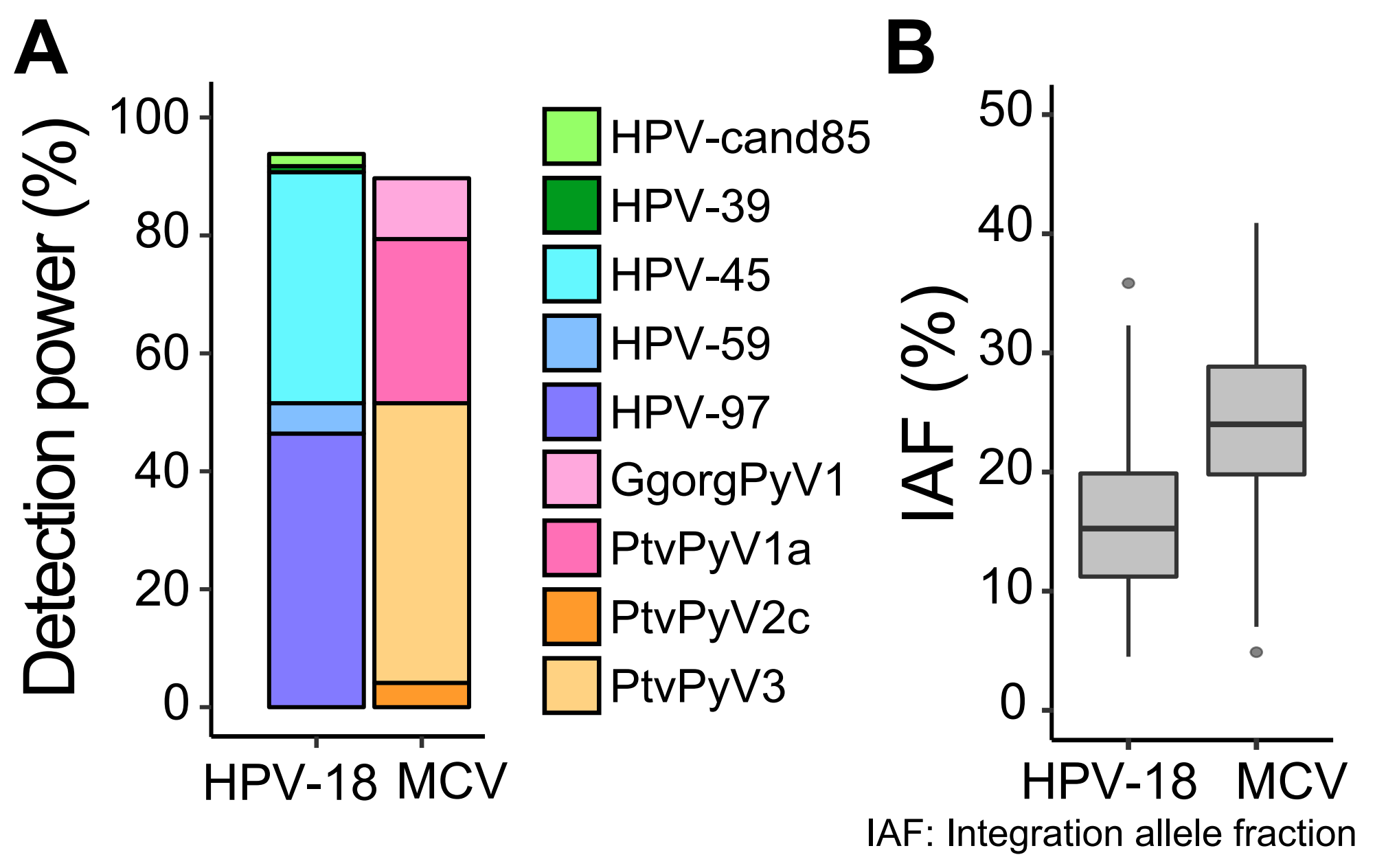
Project	Sample	Virus	Events	Highest IAF
BLCA	FD-A3B4	HPV-56	2	64.7%
BLCA	GC-A3I6	HPV-16	2	48.3%
BLCA	DK-A3IT	BKV	5	39.0%
BLCA	BT-A20V	HPV-45	1	30.0%
DLBCL	09-33003	EBV	1	18.6%
STAD	CD-5801	EBV	1	4.0%

IAF: Integration allele fraction








A
Detection power (%)
100
80
60
40
20
0
HPV-18
MCV
HPV-cand85
HPV-39
HPV-45
HPV-59
HPV-97
GgorgPyV1
PtvPyV1a
PtvPyV2c
PtvPyV3
B
IAF (%)
50
40
30
20
10
0
HPV-18
MCV
IAF: Integration allele fraction