Supplemental Methods

Samples and sequence data
Five HTS datasets derived from three cancer types were analyzed. We first analyzed the WGS data of a metastatic cervical carcinoma sample with paired normal tissue. This sample was described in our recent paper (Liang et al. 2014). The sequence data is available through the database of Genotypes and Phenotypes (dbGaP, accession number phs000628). We then analyzed the RNA-seq data of three HCC cell lines. These samples were also described in our recent paper (Lau et al. 2014). The sequence data is deposited in the Sequence Read Archive (accession number SRP023539). We then analyzed the WGS data of 99 liver cancer samples, which included 88 samples reported previously (Sung et al. 2012) and 11 new samples. The WGS data from both tumor and adjacent normal tissues is available through the European Nucleotide Archive (accession number ERP001196). We then analyzed the WGS datasets of 106 bladder urothelial carcinoma tumor samples (The Cancer Genome Atlas Research Network 2014b) and 29 stomach adenocarcinoma tumor samples with paired normal tissues (The Cancer Genome Atlas Research Network 2014a) from TCGA. Lastly, we analyzed the WGS datasets of 12 diffuse large B-cell lymphoma samples (Morin et al. 2011). The WGS data is available through the dbGaP (accession number phs000532).

Curation of virome-wide reference genome library
We created a large virome-wide viral reference genome database by combining data from a total of 15 viral sequence repositories (Kuiken et al. 2005; Lefkowitz et al. 2005; Adams and Antoniw 2006; Chang et al. 2007; Hsu et al. 2007; Huang et al. 2008; Bhaduri et al. 2012; Pickett et al. 2012b; Squires et al. 2012; Chen et al. 2013; Naeem et al. 2013; Van Doorslaer et al. 2013). This database contains all whole and partial viral genomes available to date. To eliminate redundancy, we removed identical sequences that presented across the repositories. This led to 921,550 unique viral sequences. To further reduce sequence homology and computational time, we aligned each viral sequence against the rest of the database using BLAT (Kent 2002) with parameters “-minIdentity=0 -minScore=0”. For viral sequences with no more than five
mismatches, we kept only the longest sequence as the representative. We also removed all known synthetic sequences from the library, including sequences of vector and plasmid origin in UniVec (www.ncbi.nlm.nih.gov/tools/vecscreen/univec) and EMVEC (www.ebi.ac.uk/Tools/sss/ncbiblast/vectors.html). This resulted in a total of 411,195 unique and representative viral reference genomes. The taxonomy information of each viral sequence was obtained from the NCBI and incorporated into our library.

**Simulation of viral integration datasets and negative controls**

To simulate virome-wide integration datasets, viral sequences were randomly sampled from our virome-wide reference genome library. Viral sequences of various lengths (≥ 20 bp) were inserted randomly into Chromosome 1 (hg19). Paired-end reads (100 bp in length) of various sequencing depths (1× to 150×) were simulated using pIRS according to the best practices (Hu et al. 2012; Wang et al. 2015). Three sets of data were simulated (200 bp, 500 bp, and 800 bp insert sizes) (Supplemental Table 6). To simulate different integration allele fractions (or cellular proportions), we mixed the simulated WGS datasets, with and without integrations, at three different integration allele fractions (5%, 25%, and 50%). This was repeated for three sequencing depths (5×, 30×, and 60×). All the above simulations had three replications, and the average values were calculated. To evaluate the specificity of VIcaller, we simulated negative controls using pIRS by generating WGS data at 30× sequencing depth with no inserted viral sequences. In addition, we included a publicly-available dataset simulated by VirusFider2 (Wang et al. 2015) containing one HPV-16 whole genome integration (30×) in an ~35 Mb genomic region of Chromosome 1.

**Extraction of soft-clipped sequences**

The extractSoftclipped script from SE-MEI (https://github.com/dpryan79/SE-MEI) was used to extract the soft-clipped sequences. Specifically, the SAM/BAM file was used as the input to search for reads containing soft-clipped sequences, which were then extracted to a separate single-end FASTQ file. The header of each read contains chromosome and soft-clipped sequence breakpoint information. We then modified the extractSoftclipped script to also include additional alignment information in the read header, including read name, combination of bitwise FLAGS, MAPping quality (the probability that a read is misplaced), CIGAR string (a string describing
how the read aligns to the human reference), and position of the mate read. More details can be found in the SAM format specification (https://samtools.github.io/hts-specs/SAMv1.pdf). All the soft-clipped sequences of 20 bp or longer were extracted with the parameter “-l 20”, and only the obtained soft-clipped sequences from paired-end reads mapped in proper pairs were kept for subsequent analyses.

Quality control
The four bioinformatics methods or settings for verification included two BWA-MEM (Li 2013) running modes; one with the default settings and the other with tuned (i.e., more stringent) parameters, BLAT (Kent 2002) with parameters ‘-minIdentity=0 -minScore=0 -stepSize=4’, and BLASTN (Camacho et al. 2009) with parameters ‘-word_size 10 -evalue 1e-5 -outfmt 6’. The latter two serve as independent verification as they are based on basic local alignment while BWA-MEM is based on the Burrows-Wheeler transform (Li 2013).

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. When a split read is found, the precise locations of the breakpoints on both human and viral genomes are determined at nucleotide resolution (Supplemental Figs. S17A and S17D). The length of the integrated viral sequence is calculated for events with both upstream (5’) and downstream (3’) breakpoints detected on the viral genome. On the human genome, if two different coordinates are detected by split reads for the same integration event, it indicates that an insertion or deletion has likely occurred within the integration. When there is no split read spanning a breakpoint, the coordinates will be estimated from chimeric reads only. Specifically, if both upstream and downstream breakpoints of an integration are identified, the middle nucleotide of the two human-mapped chimeric reads (one upstream read and one downstream read) nearest to the breakpoints are considered the chromosomal location of the integration (Supplemental Fig. S17C); and if only one breakpoint is identified, the first nucleotide of the chimeric read nearest to the breakpoint is considered the chromosomal location (Supplemental Fig. S17B). In both situations, the first nucleotides of the chimeric read nearest to the breakpoint on the viral genome is considered the viral genome location. Due to the rarity of split reads, the estimated location is a short window of HTS “inner distance” (usually two to three hundred base pairs in length). In
this study, a 200 bp chromosomal distance was used to determine whether two adjacent viral integrations represent the same event in an individual.
Supplemental Figure S1 Approach used by VIcaller to determine the top candidates among the viruses and viral integration events with multiple alignments. The shaded boxes indicate the reads which are removed in the subsequent step; (A) All chimeric and split reads are mapped to our virome-wide reference genome library; (B) After alignment, for the reads mapped to multiple viral genome locations, VIcaller keeps the top hits (loci with highest alignment scores) having adjacent uniquely mapped reads located within a window of one insert size; (C) For multiple references belonging to the same virus, VIcaller keeps the reference with the largest number of reads and removes redundant references; (D) VIcaller determines the candidate virus based on the ranking of the total number of chimeric and split reads, and their average alignment scores (viral reads were not used).
Supplemental Figure S2 Calculation of integration allele fraction. The viral integration-carrying human genome was first randomly fragmented and then sequenced using the Illumina next-generation sequencing technology. After sequencing, the paired-end reads that can be mapped to a viral genome are used for detecting viral integrations (a); and the paired-end mapped reads crossing the integration breakpoint are used to represent no integration (b). Integration allele fraction is calculated as the ratio of the number of integration supporting reads divided by the total number of reads (i.e., those supporting integration and those supporting no integration).
Supplemental Figure S3 All 12 possible combinations of supporting reads considered by VIcaller for detecting viral integrations. Supporting reads include chimeric, split, and viral reads. These scenarios can be divided into two categories: viral integrations with length $\geq$ insert size and those with length $<$ insert size. For the former, one paired-end read covers one breakpoint only; while for the latter, one paired-end read is able to cover both breakpoints. Reads with undetermined sequences (i.e., grey color in this figure) are not considered split reads.
Supplemental Figure S4 All eight possible scenarios of viral integrations detectable by VIcaller. (A) Integration with only upstream breakpoint identified; (B) Integration with only downstream breakpoint identified; (C) Integration with both upstream and downstream breakpoints identified, and the integrated viral sequence is longer than the insert size (e.g., 500 bp); (D) Integration where the length of the integrated viral sequence is shorter than the read length (e.g., 100 bp); (E) Integration where the length of the integrated viral sequence is shorter than the insert size (e.g., 500 bp); (F) Integration with a small deletion at the integration breakpoint; (G) Integration with a small insertion at the integration breakpoint; (H) Two or more viral integrations adjacent to each other (e.g., the distance between the two integrations is smaller than the insert size).
Supplemental Figure S5 Distribution of viral sequence lengths of the simulated germline integrations. A total of 6,133 randomly-selected viral sequences, regardless of viral species, were inserted into the human genome (hg19). This resulted in 5,535 integrations in genomic regions with nucleotide information. The remaining 598 integrations were in regions with “N” nucleotides (no nucleotide information available).
Supplemental Figure S6 Detection power by length of viral reference genome. Only viruses with full viral reference genomes were used for the analysis. Left: 0-100% scale; Right: 92-100% scale.
Supplemental Figure S7 Distribution of the lengths of integrated HBV sequences. Only the HBV integrations with both upstream and downstream breakpoints identified were used to construct the distribution.
Supplemental Figure S8 Genome-wide distribution of HBV integrations identified by VIcaller. The vertical axis represents the number of HBV integrations within a 1-Mb window, with a maximum value of 10. Two genomic regions harbored more than 10 HBV integrations: the TERT gene (Chr5) and the MLL4 gene (Chr19). “Consistent” refers to integrations detected by both VIcaller’s virome-wide approach and the previous HBV-specific approach, while “novel” refers to integrations newly identified by VIcaller.
Supplemental Figure S9 HBV integration allele fractions for all tumors with integrations in the *MLL4* gene. The top shows the highest integration allele fraction in each sample. The bottom-left shows all HBV integrations in each sample, including those in *MLL4* and other regions (except *TERT*). The bottom-right shows the violin plot distribution of allele fractions of integrations in *MLL4* compared to those in other regions (except *TERT*). The paired t-test *P* value was 0.12 and fraction difference was 7.43 (-2.50-17.37) when the average integration allele fraction values of the integrations in *MLL4* and those in other regions was compared. The *P* value was 0.0002 and fraction difference was 17.05 (8.79-25.31) when *TERT* and *MLL4* were combined for analysis.
**Supplemental Figure S10** Sequence read alignment of an EBV (AB850658.1) integration event in a TCGA sample (TCGA-CD-5801). A total of two chimeric reads were found crossing the two breakpoints.
Supplemental Figure S11 Single nucleotide resolution reconstructed integration map of the chimeric and split reads aligned to a simulated HPV-16 (FJ610151.1) integration detected by VIcaller. The HPV-16 integration, which was comprised of a total of 6,500,971 paired-end reads, was originally simulated on Chromosome 1 (hg19) using “pIRS” (Hu et al. 2012) in the VirusFinder 2 study. Using VIcaller, we detected a total of 1,957 supporting reads, including 1,923 viral reads, 10 chimeric reads, and 24 split reads. Compared to VirusFinder 2, VIcaller identified six additional chimeric and split reads that were overlooked by VirusFinder 2. The integration breakpoint identified by VIcaller was at Chr1:24,020,701.
Supplemental Figure S12 Reciprocal alignment. VIcaller re-aligns all identified chimeric and split reads back to the human reference genome using less stringent criteria. Reads that are mappable to the human genome are not used for the discovery of viral integrations.
1) Viral integration to be located:

2) Multiple potential integration sites due to repeat sequences:

3) Potential integration location 1:

4) Potential integration location 2:

- Non-repeat region in the human genome
- Repeat sequence in the human genome
- Viral genome
- Reads map to the viral genome
- Reads map to the repeat region in human genome
- Reads map to the non-repeat region in the human genome

Supplemental Figure S13 Viral integrations derived from repeat regions may be mapped to multiple locations.
Supplemental Figure S14 Supporting reads flanking an integration breakpoint in an insert size window. VIcaller removes reads mapped outside this window during the integration discovery stage to eliminate potentially false supporting (chimeric and split) reads.
Supplemental Figure S15 VIcaller’s approach for determining viral integration statuses (four possibilities are shown in this figure) based on the presence of viral, chimeric, and split reads.
Supplemental Figure S16 Approach for distinguishing between viral infection/contamination and viral integration statuses based on the distribution of viral read depths when there is only one integration event.
Supplemental Figure S17 Approach for fine-mapping integration breakpoints on the human genome. (A) Both the upstream and downstream breakpoints detected, with the breakpoint coordinate(s) precisely determined based on split reads (human); (B) Breakpoint detected only in one direction, with the breakpoint coordinate precisely determined based on split reads (human); (C) Both the upstream and downstream breakpoints detected, with the breakpoint coordinate estimated according to the middle nucleotide between the two chimeric reads (human) closest to the breakpoint; (D) Breakpoint detected only in one direction, with the breakpoint coordinate estimated according to the first nucleotide of the chimeric read (human) closest to the breakpoint. The same approach was used to fine-map the two breakpoints on the viral genome.
## Supplemental Tables

### Supplemental Table S1 Resources included in our virome-wide reference genome library

| Database and Resource                        | Sequence type             | No. sequences | Reference                           |
|----------------------------------------------|---------------------------|---------------|-------------------------------------|
| Virus Pathogen Resource (ViPR)               | Whole & partial genome    | 545,884       | (Pickett et al. 2012a)             |
| NCBI Influenza Database                      | Whole & partial genome    | 366,978       | (Bao et al. 2008)                  |
| Influenza Research Database (IRD)            | Whole & partial genome    | 360,305       | (Squires et al. 2012)              |
| Readscan*                                    | Whole & partial genome    | 328,922       | (Naeem et al. 2013)                |
| Descriptions of Plant Viruses (DPV)          | Whole & partial genome    | 57,380        | (Adams and Antoniw 2006)           |
| Influenza Virus Database (IVDB)               | Whole & partial genome    | 43,874        | (Chang et al. 2007)                |
| RINS*                                        | Whole & partial genome    | 32,102        | (Bhaduri et al. 2012)              |
| Virus miRNA Target (ViTa)                    | Whole & partial genome    | 30,090        | (Hsu et al. 2007)                  |
| VirusSeq*                                    | Whole & partial genome    | 25,525        | (Chen et al. 2013)                 |
| NCBI Viral Genomes                           | Whole genome              | 6,009         | (Brister et al. 2015)              |
| Viral Bioinformatics Resource Center (VBRC)  | Whole & partial genome    | 5,406         | (Upton and Lefkowitz)              |
| Hepatitis C Virus Database Project (HCV)      | Whole genome              | 1,798         | (Kuiken et al. 2005)               |
| Papillomavirus Episteme (PaVE)               | Whole genome              | 321           | (Van Doorslaer et al. 2013)        |
| Coronavirus Database (CoVDB)                 | Whole genome              | 268           | (Huang et al. 2008)                |
| Poxvirus Bioinformatics Resource Center (PBR)| Whole genome              | 112           | (Lefkowitz et al. 2005)            |
| **Total**                                    |                           | **1,804,974** |                                     |
| **Total (unique)**                           |                           | **921,550**   |                                     |
| **Total (unique and representative)**        |                           | **411,195**   |                                     |

*Bioinformatics software which contains a viral genome database.
## Supplemental Table S2 List of bioinformatics tools embedded in VIcaller

| Bioinformatics tool (version) | Download link                                                                 | Reference                                      |
|-------------------------------|------------------------------------------------------------------------------|------------------------------------------------|
| BLAT (v.35)                   | [http://genomic-identity.wikidot.com/install-blat](http://genomic-identity.wikidot.com/install-blat) | (Kent 2002)                                    |
| Bowtie2 (v2.2.7)              | [https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.2.7/](https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.2.7/) | (Langmead and Salzberg 2012)                  |
| BWA (v0.7.12)                 | [https://github.com/lh3/bwa/tree/master/bwakit](https://github.com/lh3/bwa/tree/master/bwakit) | (Li 2013)                                      |
| FastUniq (v1.1)               | [https://sourceforge.net/projects/fastuniq/](https://sourceforge.net/projects/fastuniq/) | (Xu et al. 2012)                              |
| HYDRA (v0.5.3)                | [https://code.google.com/archive/p/hydra-sv/downloads](https://code.google.com/archive/p/hydra-sv/downloads) | (Quinlan et al. 2010)                         |
| MEME (v4.11.1)                | [http://web.mit.edu/meme_v4.11.4/share/doc/download.html](http://web.mit.edu/meme_v4.11.4/share/doc/download.html) | (Bailey et al. 2009)                          |
| BLAST+ (2.2.30)               | [http://mirrors.vbi.vt.edu/mirrors/ftp.ncbi.nih.gov/blast/executables/blast%2B/2.2.30/](http://mirrors.vbi.vt.edu/mirrors/ftp.ncbi.nih.gov/blast/executables/blast%2B/2.2.30/) | (Camacho et al. 2009)                         |
| NGS QC Toolkit (v2.3.3)       | [http://www.nipgr.res.in/ngsqctoolkit.html](http://www.nipgr.res.in/ngsqctoolkit.html) | (Patel and Jain 2012)                         |
| RepeatMasker (v4.0.5)         | [http://www.repeatmasker.org/](http://www.repeatmasker.org/)                  | (Smit et al. 2015)                            |
| TRF (v4.07b)                  | [https://tandem.bu.edu/trf/trf.html](https://tandem.bu.edu/trf/trf.html)        | (Benson 1999)                                  |
| SAMtools (v1.6)               | [https://sourceforge.net/projects/samtools/](https://sourceforge.net/projects/samtools/) | (Li et al. 2009)                              |
| SE-MEI* (Modified)            | [https://github.com/dpryan79/SE-MEI (original version)](https://github.com/dpryan79/SE-MEI (original version)) | -                                              |
| TopHat (v2.1.1)               | [http://ccb.jhu.edu/software/tophat/index.shtml](http://ccb.jhu.edu/software/tophat/index.shtml) | (Kim et al. 2013)                              |

*The original C code for the output was modified, and then included in the VIcaller installer.
**Supplemental Table S3** Terms and definitions

| Term            | Definition                                                                                                                                 |
|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| SE read         | Single-end read                                                                                                                          |
| PE read         | Paired-end read                                                                                                                           |
| VI              | Viral integration                                                                                                                         |
| Human read      | Read that is fully mapped to the human reference genome                                                                                  |
| Unmapped read   | Read that is not fully aligned to the human reference genome, including paired-end unmapped reads, one-end unmapped reads (one-end mapped reads are also extracted), and soft-clipped sequences |
| Supporting read | Read that supports a viral integration event; Supporting reads include chimeric, split, and viral reads                                    |
| Split read      | Read that contains a viral integration breakpoint; the read is aligned to both the human and viral genomes                                  |
| Chimeric read   | Read pair that does not contain a viral integration breakpoint; one read of the pair is aligned to the human genome and the other read is aligned to a viral genome |
| Viral read      | Read that is not aligned to the human reference genome but is fully aligned to a viral reference genome                                   |
**Supplemental Table S4** List of known viral integrations and fusion transcripts from commonly-used cell lines collected from literature. Currently, we include a total of 25 cell lines. Among them, 11 cell lines have viral integrations, 16 cell lines have fusion transcripts, and two have both, e.g., four HPV-18 integration events in the HeLa cells and two HPV-16 integration events in the SiHa cells ([Excel table](#))

**Supplemental Table S5** Viral integration format file ([Excel table](#))
### Supplemental Table S6 Simulated viral integration sequencing datasets for measuring the accuracy of VIcaller

|                     | Virome-wide germline viral integrations | Virome-wide somatic viral integrations | 10 HPV-16 integration$^1$ | One HPV-16 integration$^2$ | Negative control |
|---------------------|----------------------------------------|----------------------------------------|---------------------------|---------------------------|-----------------|
| Sequencing type     | WGS                                    | WGS                                    | WGS                       | WGS                       | WGS             |
| Sequencing depths (×)| 1 - 150                                | 5, 30 and 60                           | 5, 30 and 60              | 30                        | 30              |
| Simulated allele fraction | 100%                                   | 5%, 25%, 50% and 100%$^3$              | 5%, 25%, 50% and 100%     | 100%                      | 100%            |
| Read type           | Paired-end                              | Paired-end                              | Paired-end                | Paired-end                | Paired-end      |
| Read length (bp)    | 100                                    | 100                                    | 100                       | 75                        | 100             |
| Insert size (mean ± SD) (bp) | 200 ± 10; 500 ± 25; 800 ± 40           | 500 ± 25                               | 500 ± 25                  | 200 ± 10                  | 500 ± 25        |
| Human reference sequence for simulation | Chr1                                    | Chr1                                    | Chr1                      | Chr1                      | Whole genome    |
| Total detectable viral integrations | 5,535                                   | 90$^4$,89 and 78                        | 10                        | 1                         | 0               |
| Integrated viral sequences | Partial genome                          | Partial genome                          | Partial genome            | Whole genome              | -               |
| Minimum length of integrated viral genomes (bp) | 20                                      | 500                                    | 500                       | 7,906$^5$                 | -               |
| Replications        | 3                                      | 3                                      | 1                         | 1                         | 1               |

Viral sequences were randomly inserted into the human reference genome using our in-house script. Reads were simulated using pIRS (Hu et al. 2012).

1,$^1$This series of data was only used for software comparison. The HPV-16 (FJ006723.1) genome with various length was randomly inserted into Chromosome 1 (hg19).

2,$^2$The dataset was derived from a previous study (Wang et al. 2015), in which one HPV-16 whole-genome was inserted into Chromosome 1.

3,$^3$Allele fraction of 100% was only used for software comparison.

4,$^4$A series of datasets carrying 90 virome-wide somatic viral integrations was also used for software comparison with various sequencing depths and allele fractions.

5,$^5$The genome length of HPV-16 (NC_001526.2).
### Supplemental Table S7 Numbers of chimeric and split reads of simulated viral integrations

| Depth (×) | No. integrations simulated (mean ± SD)* | No. integrations detected (mean ± SD) | Lengths of integrated viral sequences (bp) (mean ± SD) | No. chimeric reads (mean ± SD) | No. split reads (mean ± SD) |
|-----------|----------------------------------------|--------------------------------------|-------------------------------------------------------|-------------------------------|-----------------------------|
| 1         | 461.3 ± 23.3                           | 299.4 ± 117.8                       | 7585.6 ± 1553.8                                       | 3.6 ± 1.5                    | 0.8 ± 0.1                   |
| 2         | 461.3 ± 23.3                           | 372.9 ± 74.0                        | 7607.0 ± 1376.0                                       | 6.3 ± 3.3                    | 1.4 ± 0.2                   |
| 3         | 461.3 ± 23.3                           | 403.1 ± 49.6                        | 7391.3 ± 1086.3                                       | 8.9 ± 4.9                    | 2.0 ± 0.4                   |
| 4         | 461.3 ± 23.3                           | 416.3 ± 42.4                        | 7381.1 ± 1105.3                                       | 11.7 ± 6.6                   | 2.7 ± 0.5                   |
| 5         | 461.3 ± 23.3                           | 423.1 ± 40.0                        | 7375.7 ± 1049.4                                       | 14.4 ± 8.3                   | 3.3 ± 0.6                   |
| 6         | 461.3 ± 23.3                           | 426.5 ± 38.1                        | 7355.6 ± 1052.2                                       | 17.1 ± 9.8                   | 3.9 ± 0.8                   |
| 7         | 461.3 ± 23.3                           | 427.8 ± 37.6                        | 7321.3 ± 1050.8                                       | 19.8 ± 11.4                  | 4.5 ± 0.9                   |
| 8         | 461.3 ± 23.3                           | 430.2 ± 36.4                        | 7310.5 ± 1039.9                                       | 22.4 ± 13.0                  | 5.2 ± 1.0                   |
| 9         | 461.3 ± 23.3                           | 430.8 ± 36.4                        | 7307.5 ± 1032.5                                       | 25.1 ± 14.6                  | 5.8 ± 1.1                   |
| 10        | 461.3 ± 23.3                           | 431.6 ± 35.8                        | 7300.1 ± 1042.8                                       | 27.7 ± 16.0                  | 6.4 ± 1.2                   |
| 11        | 461.3 ± 23.3                           | 432.0 ± 35.7                        | 7276.5 ± 1008.6                                       | 30.4 ± 17.7                  | 7.1 ± 1.4                   |
| 12        | 461.3 ± 23.3                           | 432.3 ± 35.4                        | 7265.0 ± 971.5                                       | 32.9 ± 19.2                  | 7.7 ± 1.5                   |
| 13        | 461.3 ± 23.3                           | 432.5 ± 35.6                        | 7277.5 ± 1001.1                                       | 35.6 ± 20.7                  | 8.3 ± 1.6                   |
| 14        | 461.3 ± 23.3                           | 432.5 ± 35.3                        | 7263.8 ± 991.0                                       | 38.1 ± 22.2                  | 8.9 ± 1.7                   |
| 15        | 461.3 ± 23.3                           | 433.0 ± 35.4                        | 7262.3 ± 1001.7                                       | 40.6 ± 23.6                  | 9.5 ± 1.8                   |
| 16        | 461.3 ± 23.3                           | 433.5 ± 35.1                        | 7273.2 ± 1010.9                                       | 43.1 ± 25.1                  | 10.1 ± 1.9                  |
| 17        | 461.3 ± 23.3                           | 433.6 ± 34.9                        | 7264.0 ± 989.0                                       | 45.6 ± 26.5                  | 10.7 ± 2.0                  |
| 18        | 461.3 ± 23.3                           | 433.2 ± 35.2                        | 7263.1 ± 991.0                                       | 48.1 ± 28.0                  | 11.3 ± 2.2                  |
| 19        | 461.3 ± 23.3                           | 433.4 ± 35.2                        | 7276.9 ± 1027.4                                       | 50.7 ± 29.5                  | 11.9 ± 2.2                  |
| 20        | 461.3 ± 23.3                           | 434.2 ± 34.5                        | 7261.9 ± 1023.8                                       | 53.0 ± 30.8                  | 12.4 ± 2.3                  |
| 25        | 461.3 ± 23.3                           | 434.7 ± 34.2                        | 7260.3 ± 1023.4                                       | 65.0 ± 37.8                  | 15.3 ± 2.8                  |
| 30        | 461.3 ± 23.3                           | 434.7 ± 34.4                        | 7268.5 ± 1036.6                                       | 76.5 ± 44.4                  | 18.0 ± 3.4                  |
| 35        | 461.3 ± 23.3                           | 435.3 ± 34.0                        | 7257.5 ± 1026.8                                       | 87.7 ± 50.9                  | 20.7 ± 3.7                  |
| 40        | 461.3 ± 23.3                           | 435.7 ± 33.8                        | 7242.8 ± 1011.1                                       | 98.5 ± 57.2                  | 23.1 ± 4.1                  |
| 45        | 461.3 ± 23.3                           | 435.4 ± 34.0                        | 7247.0 ± 1015.1                                       | 108.9 ± 63.1                 | 25.7 ± 4.5                  |
| 50        | 461.3 ± 23.3                           | 436.4 ± 33.8                        | 7234.8 ± 1010.4                                       | 118.9 ± 68.8                 | 28.0 ± 4.9                  |
| 100       | 470.7 ± 31.5                           | 462.3 ± 31.4                        | 7040.1 ± 608.0                                       | 231.3 ± 18.8                 | 48.5 ± 3.5                  |
| 150       | 470.7 ± 31.5                           | 462.3 ± 31.4                        | 7040.1 ± 608.0                                       | 303.4 ± 26.0                 | 63.5 ± 4.0                  |

*The simulated viral integrations located in regions with “N” nucleotides were excluded.

Viral reads were not calculated because of the homologous sequences among viral genomes.
**Supplemental Table S8** Detection power and precision of the simulated somatic viral integrations detected by VIcaller

| Simulated dataset | Total No. simulated integrations | Depth (×) | Integration allele fraction | Total No. detected integrations | No. correctly detected integrations | Detection power | Precision |
|-------------------|----------------------------------|-----------|-----------------------------|--------------------------------|------------------------------------|----------------|-----------|
| Dataset-1         | 90                               | 5         | 5%                          | 13                              | 11                                 | 12.22%         | 84.62%    |
| Dataset-1         | 90                               | 5         | 25%                         | 67                              | 67                                 | 74.44%         | 100%      |
| Dataset-1         | 90                               | 5         | 50%                         | 81                              | 81                                 | 90.00%         | 100%      |
| Dataset-1         | 90                               | 30        | 5%                          | 70                              | 70                                 | 77.78%         | 100%      |
| Dataset-1         | 90                               | 30        | 25%                         | 87                              | 87                                 | 96.67%         | 100%      |
| Dataset-1         | 90                               | 30        | 50%                         | 88                              | 88                                 | 97.78%         | 100%      |
| Dataset-1         | 90                               | 60        | 5%                          | 84                              | 84                                 | 93.33%         | 100%      |
| Dataset-1         | 90                               | 60        | 25%                         | 86                              | 86                                 | 95.56%         | 100%      |
| Dataset-1         | 90                               | 60        | 50%                         | 87                              | 87                                 | 96.67%         | 100%      |
| Dataset-2         | 89                               | 5         | 5%                          | 9                               | 8                                  | 8.99%          | 88.89%    |
| Dataset-2         | 89                               | 5         | 25%                         | 63                              | 63                                 | 70.79%         | 100%      |
| Dataset-2         | 89                               | 5         | 50%                         | 85                              | 85                                 | 95.51%         | 100%      |
| Dataset-2         | 89                               | 30        | 5%                          | 75                              | 74                                 | 83.15%         | 98.67%    |
| Dataset-2         | 89                               | 30        | 25%                         | 89                              | 89                                 | 100%           | 100%      |
| Dataset-2         | 89                               | 30        | 50%                         | 88                              | 88                                 | 98.88%         | 100%      |
| Dataset-2         | 89                               | 60        | 5%                          | 85                              | 85                                 | 95.51%         | 100%      |
| Dataset-2         | 89                               | 60        | 25%                         | 88                              | 88                                 | 98.88%         | 100%      |
| Dataset-2         | 89                               | 60        | 50%                         | 89                              | 89                                 | 100%           | 100%      |
| Dataset-3         | 78                               | 5         | 5%                          | 10                              | 10                                 | 12.82%         | 100%      |
| Dataset-3         | 78                               | 5         | 25%                         | 58                              | 58                                 | 74.36%         | 100%      |
| Dataset-3         | 78                               | 5         | 50%                         | 71                              | 71                                 | 91.03%         | 100%      |
| Dataset-3         | 78                               | 30        | 5%                          | 64                              | 64                                 | 82.05%         | 100%      |
| Dataset-3         | 78                               | 30        | 25%                         | 77                              | 77                                 | 98.72%         | 100%      |
| Dataset-3         | 78                               | 30        | 50%                         | 77                              | 77                                 | 98.72%         | 100%      |
| Dataset-3         | 78                               | 60        | 5%                          | 74                              | 74                                 | 94.87%         | 100%      |
| Dataset-3         | 78                               | 60        | 25%                         | 78                              | 78                                 | 100%           | 100%      |
| Dataset-3         | 78                               | 60        | 50%                         | 78                              | 78                                 | 100%           | 100%      |
| **Total**         | **2,313**                        |           |                             | **1,921**                       | **1,917**                          | **82.88%**     | **99.79%** |
### Supplemental Table S9 Results of VIcaller detection of viral integrations from simulated WGS negative control data

|                                | Negative control (expected) | Reads and integrations (detected) |
|--------------------------------|----------------------------|----------------------------------|
| **Total simulated reads**      | 470,618,138                |                                  |
| **Reads mapped to VIcaller library** |                           |                                  |
| Chimeric reads                 | 0                          | 0                                |
| Split reads                    | 0                          | 177*                             |
| Viral reads                    | 0                          | 0                                |
| **Number of reads mapped to HPV-16** |                           |                                  |
| Chimeric reads                 | 0                          | 0                                |
| Split reads                    | 0                          | 0                                |
| Viral reads                    | 0                          | 0                                |
| **Virus**                      | 0                          | 0                                |
| **Integration with two or more chimeric & split reads** | 0                          | 0                                |
| **Number of integrations**     | 0                          | 0                                |

The whole-genome paired-end reads (30×) were simulated from the human reference genome (hg19) using pIRS (Hu et al. 2012).

*All 177 split reads were distributed randomly across the human genome and no two reads were clustered together, indicating these were random noise. VIcaller found no integrations in the negative control dataset.
**Supplemental Table S10** Comparison of viral integration detection power and precision among VIcaller and three of the commonly used tools (Virus-Clip, VirusSeq, and VirusFinder 2). A series of datasets containing 10 HPV integrations and a series of datasets containing 90 virome-wide integrations were simulated, with each set having various sequencing depths and allele fractions. Each software was applied to all datasets. Virus-Clip was candidate virus-based, and thus was applied for detecting the 10 HPV integrations using the HPV-16 viral genome as its viral genome reference. For VirusFinder 2 and VirusSeq, we also applied them to detect the 90 virome-wide integrations. We provided VirusFinder 2 our virome-wide reference genome library as its reference genome for comparisons among tools. However, VirusFinder 2 showed very low detection power (~1%) for the virome-wide integration detection; and VirusSeq could not finish the analysis jobs in three days, even for the analysis of a small number of candidate integrations (by comparison, the computational time/duration for VIcaller was no more than half an hour). Thus, VirusFinder 2 and VirusSeq are not appropriate for virome-wide integration detection (Excel Table)

**Supplemental Table S11** Viral integrations in liver cancer, cervical cancer, and bladder cancer samples detected by VIcaller. The table shows the information for integration breakpoints on both human and viral genomes, integration direction, number of chimeric and split reads, integration allele fraction and bioinformatics validation results for each integration event (Excel Table sheets A-D)
**Supplemental Table S12** VIcaller detection of viral integrations in a metastatic cervical carcinoma patient

| Sample ID (dbGaP) | Sequencing type | Tissue type | Read length (bp) | Virus | #Known integrations | #Known integrations detected by VIcaller | #Novel integrations detected by VIcaller | #Novel integrations validated by Sanger sequencing |
|-------------------|----------------|-------------|-----------------|-------|---------------------|------------------------------------------|-------------------------------------------|------------------------------------------------|
| SRR1016778        | WGS            | Tumor       | 83              | HPV-18| 2                   | 2 (100%)                                 | 3                                         | 3                                             |
|                   | (Liang et al. 2014) |             |                 |       |                     |                                          |                                           |                                               |
| SRR1016780        | WGS            | Normal      | 83              | HPV-18| 0                   | 0                                        | 0                                         | -                                             |

The sample was described in a previous study (Liang et al. 2014). HPV-18 represents human papillomavirus type 18.
Supplemental Table S13  BLASTN validation of 425 randomly selected chimeric and split reads detected by VIcaller

|                        | Detected by VIcaller | Validated by BLASTN alignment against whole NCBI nucleotide database |
|------------------------|----------------------|---------------------------------------------------------------------|
| No. reads aligned to VIcaller virome-wide library\(^1\) | 425                  | 425 (100%)                                                          |
| No. reads aligned to the HBV viral genome\(^2\)       | 227                  | 227 (100%)                                                          |
| No. reads with consistent nucleotide locations in a 100 bp window for the same Accession version | 227                  | 227 (100%)                                                          |
| No. reads aligned to the human genome\(^3\)           | 198                  | 198 (100%)                                                          |
| No. reads with consistent nucleotide locations in a 100 bp window | 198                  | 198 (100%)                                                          |
| No. reads aligned uniquely to the human genome with stringent E-value | 198                  | 198 (100%)                                                          |

\(^1\) Each end of the paired-end reads was considered as a read.

\(^2\) Seven reads were aligned to the NCBI Nucleotide database with parameters specifically for short sequences.

\(^3\) Repeats were not filtered or masked during the BLASTN analysis against the human genome.
**Supplemental Table S14** Comparison of VIcaller virome-wide approach with a previous HBV-specific approach

| **Previous HBV-specific approach** (Sung et al. 2012) | **Our VIcaller virome-wide approach** |
|------------------------------------------------------|---------------------------------------|
| Use only one HBV reference genome (NC_003977)        | Use a virome-wide reference genome library |
| Align to a hybrid HBV and human reference genome     | Subtract reads mapped to the human genome, and then align the unmapped reads to the VIcaller virome-wide reference genome library |
| Remove PCR duplicate reads (those mapped to the same genomic locations, i.e., ± 2 bp) | Remove PCR duplicate reads (those mapped to exactly the same genomic locations, i.e., ± 0 bp) |
| Do not remove reads with repeat sequences (high false positive rate) | Remove reads with < 20 bp non-repeat sequences |
| Do not remove reads with homologous sequences (high false positive rate) | Reciprocal alignment to the human genome using less stringent parameters to remove reads with homologous sequences |
| Only use reads with insert size of 170 bp for fine-mapping of breakpoints | Use all reads (insert sizes of 170 bp and 800 bp) for fine-mapping of breakpoints |
| Use BLAT for alignment to finely-map breakpoints     | Use BWA-MEM for alignment to finely-map breakpoints |
| Detect only one breakpoint in both human and HBV genome | Detect both the upstream (5’) and downstream (3’) breakpoints of an integration event in both human and HBV genome |
| Do not detect InDels around integration breakpoints   | Detect InDels around integration breakpoints |
**Supplemental Table S15** PCR and Sanger sequencing validation of seven HBV integrations newly detected by VIcaller

| Sample ID | Integration breakpoint | Detection by VIcaller | Detection by HIVID using an HBV sequence enrichment approach | PCR/Sanger sequencing validation by HIVID |
|-----------|-------------------------|-----------------------|---------------------------------------------------------------|----------------------------------------|
| 49T       | Chr2:204,666,726        | Detected              | Detected                                                      | Validated                              |
| 49T       | Chr2:205,395,701        | Detected              | Detected                                                      | Validated                              |
| 65T       | Chr5:1,295,442          | Detected              | Detected                                                      | Validated                              |
| 43T       | Chr7:154,782,385        | Detected              | Detected                                                      | Validated                              |
| 266T      | Chr12:21,035,361        | Detected              | Detected                                                      | Validated                              |
| 32T       | Chr17:14,716,384        | Detected              | Detected                                                      | Validated                              |
| 32T       | ChrX:115,667,431        | Detected              | Detected                                                      | Validated                              |
| 26T       | Chr3:140,571,184        | Detected              | Detected                                                      | Not tested                             |
| 55T       | Chr4:191,039,672        | Detected              | Detected                                                      | Not tested                             |
| 34T       | Chr5:1,295,375          | Detected              | Detected                                                      | Not tested                             |
| 64T       | Chr5:1,307,431          | Detected              | Detected                                                      | Not tested                             |
| 145T      | Chr13:99,639,767        | Detected              | Detected                                                      | Not tested                             |
| 145T      | Chr14:73,450,075        | Detected              | Detected                                                      | Not tested                             |
| 30T       | Chr4:191,040,807        | Detected              | Fail to detect                                                | Not tested                             |
| 95T       | Chr1:47,605,956         | Detected              | Fail to detect                                                | Not tested                             |

HIVID represents a high-throughput viral integration detection method (Li et al. 2013b).
### Supplemental Table S16

HBV integrations detected by HIVID but not by VIcaller or Sanger sequencing

| Sample ID | Integration breakpoint detected by HIVID | Detection by HIVID | PCR and Sanger sequencing | Detection by VIcaller |
|-----------|----------------------------------------|--------------------|---------------------------|----------------------|
| 71T       | Chr1:10,216                            | Detected           | Not validated             | Not detected         |
| 41T       | Chr18:10,111                           | Detected           | Not validated             | Not detected         |
| 43T       | Chr10:42,389,516                       | Detected           | Not validated             | Not detected         |
| 71T       | Chr10:42,389,892                       | Detected           | Not validated             | Not detected         |
| 90T       | ChrY:13,458,418                        | Detected           | Not validated             | Not detected         |

HIVID represents a high-throughput viral integration detection method (Li et al. 2013b).
Supplemental Table S17 Adeno-associated virus (AAV or AAV-like) integrations detected in two liver cancer samples (Excel Table)

Supplemental Table S18 HBV integrations in TERT containing at least one HBV enhancer or promoter (Excel Table sheets A-B)

Supplemental Table S19 List of tumor and matched normal tissues (for viral integration positive samples) of the 106 analyzed TCGA bladder cancer samples (Excel Table)

Supplemental Table S20 List of tumor and matched normal tissues (for viral integration positive samples) of the 12 analyzed CGCI non-Hodgkin lymphoma patients (Excel Table)

Supplemental Table S21 List of tumor and matched normal tissues (for viral integration positive samples) of 29 analyzed TCGA stomach adenocarcinoma samples (Excel Table)
### Supplemental Table S22

Viruses and number of integration events detected after removing the candidate references from our virome-wide database.

| Virus annotated                                      | Accession number | No. integrations | Total |
|------------------------------------------------------|------------------|------------------|-------|
| **MCV (NC_010277.2)**                                |                  |                  |       |
| Gorilla gorilla gorilla polyomavirus 1               | HQ385752.1       | 10               | 87    |
| Pan troglodytes verus polyomavirus 1a                | HQ385746.1       | 27               |       |
| Pan troglodytes verus polyomavirus 2c                | HQ385751.1       | 4                |       |
| Pan troglodytes verus polyomavirus 3                | HQ385748.1       | 46               |       |
| **HPV-18 (NC_001357.1)**                             |                  |                  |       |
| Human papillomavirus cand85                          | AF131950.1       | 2                | 91    |
| Human papillomavirus type 39                         | M62849.1         | 1                |       |
| Human papillomavirus type 45                         | EF202163.1, M38198.1, EU779717.1 & EF202156.1 | 38 | 91 |
| Human papillomavirus type 59                         | X77858.1 & EU918767.1 | 5     | 91    |
| Human papillomavirus type 97                         | DQ080080.1, EF436229.1 & EF202168.1 | 45 | 91    |
### Supplemental Table S23

Viruses and integration events detected after removing the candidate viral reference genomes from our virome-wide database

| Sample ID     | Chr. | Upstream breakpoint on human | Downstream breakpoint on human | Virome-wide analysis | If the target virus was removed from VLcaller library |
|---------------|------|-----------------------------|-------------------------------|----------------------|-----------------------------------------------------|
|               |      |                             |                               |                      | Detected virus | Accession number | No. supporting reads | Integration allele fraction | Detected virus | Accession number | No. supporting reads | Integration allele fraction |
| TCGA-DK-A3IT  | 2    | 164592606                   | 164597503                     | BK virus             | AB485697.1      | 4               | 13.8%             | -                  | BK virus        | AB485698.1      | 23               | 39.0%             | VmPyV2           | AB767299.2      | 12               | 25%              |
|               | 2    | 164602831                   |                                | BK virus             | AB211370.1       | 3               | 5.8%              | -                  | BK virus        | AB211370.1      | 3                | 5.8%              | -                 | -                | -                | -                |
|               | 2    | 192282973                   |                                | BK virus             | AY628234.1       | 6               | 16.2%             | JC virus           | DQ875211.1      | 4                | 16.2%             | -                 | -                | -                | -                |
|               | 4    | 99351267                    |                                | BK virus             | AB211370.1       | 4               | 26.7%             | PteronotusPyV*     | JX520662.1      | 2                | 15.4%             | -                 | -                | -                | -                |
| TCGA-FD-A3B4  | 9    | 139375582                   |                                | HPV-56               | X74483.1         | 10              | 52.6%             | HPV-53*            | GQ472849.1      | 3                | 52.6%             | -                 | -                | -                | -                |
|               | 9    | 139396854                   |                                | HPV-56               | X74483.1         | 11              | 64.7%             | HPV-66*            | EF202154.1      | 2                | 64.7%             | -                 | -                | -                | -                |
| TCGA-BT-A20V  | 9    | 117921629                   | 117921810                     | HPV-45               | EF202163.1       | 12              | 30.0%             | HPV-97             | EF202168.1      | 5                | 30.0%             | -                 | -                | -                | -                |
| TCGA-GC-A3I6  | 20   | 30300027                    | 30300037                      | HPV-16               | U89348.1         | 69              | 48.3%             | HPV-35*            | M74117.1        | 7                | 48.3%             | -                 | -                | -                | -                |
|               | 20   | 30314047                    |                                | HPV-16               | AF536179.1       | 3               | 7.0%              | -                  | -                | -                | -                 | -                | -                | -                | -                |

BK virus: BK polyomavirus; HPV: Human papillomavirus; JC virus: JC polyomavirus; VmPyV2: Vervet monkey polyomavirus 2; PteronotusPyV: Pteronotus davyi polyomavirus 1.

*,Identified when a less stringent threshold was applied.
**Supplemental Table S24** Comparison of software features for virus and viral integration

| Features | Viral sequence detection only | Viral integration detection only | Both |
|----------|-----------------------------|---------------------------------|------|
|          | PathSeq (Kostic et al. 2011) | RINs (Bhaduri et al. 2012)     |      |
|          | CaPSID (Borozan et al. 2012) | READSCAN (Naeem et al. 2013)   |      |
|          | SURPI (Naccache et al. 2014) | SeqMap2.0 (Hawkins et al. 2011) |      |
|          | SummonChimeras (Katz and Pipas 2014) | Virus-Clip (Ho et al. 2015) |      |
|          |                                | SeeksV (Liang et al. 2017)     |      |
|          |                                | BATVI (Tennakoon and Sung 2017) |      |
|          |                                | VirusFinder 1 & 2 (Wang et al. 2013; Wang et al. 2015) |      |
|          |                                | VirusSeq (Chen et al. 2013)    |      |
|          |                                | ViralFusionSeq (Li et al. 2013a) |      |
|          |                                | Vy-PER (Forster et al. 2015)   |      |
|          |                                | VIcaller (Forster et al. 2019)  |      |
| Year     | 2011                         | 2012                            | 2011 |
| Year     | 2013                         | 2013                            | 2013 |
| Year     | 2014                         | 2014                            | 2014 |
| Year     | 2015                         | 2015                            | 2015 |
| Year     | 2016                         | 2016                            | 2016 |
| Year     | 2017                         | 2017                            | 2017 |
| Year     | 2013 & 2015                  | 2013                            | 2013 |
| Year     | 2015                         | 2015                            | 2015 |
| Year     | 2019                         | 2019                            | 2019 |
| Sequencing platforms | Illumina | Illumina & SOLID | Illumina | Illumina | Illumina | Illumina | Illumina |
| Sequence reads | PE & SE | SE | PE & SE | SE | PE & SE | SE | PE & SE | PE | PE & SE |
| Sequence reads | FASTQ | FASTQ | FASTQ | FASTQ | FASTQ | FASTQ | FASTQ | FASTQ | FASTQ & BAM |
| Operation system | Linux | Linux | Linux | Ubuntu | Linux | Linux | Linux | Linux | Linux | Linux |
| Flexibility to replace software components | Y | N | Y | N | N | N | N | Y | N |
| Flexibility to replace software components | N | N | Y | N | N | N | N | Y | N |
| Flexibility to replace software components | N | N | Y | N | N | N | N | Y | N |
| Flexibility to replace software components | N | N | N | N | N | N | N | Y | N |
| Flexibility to replace software components | N | N | N | N | N | N | N | Y | N |
| Input formats | FASTQ & BAM | SAMAL | SNAP | BLAT & Bowtie2 | BLASTN | BWA & BWA-MEM & BLASTN | BWA-MEM | BWA-MEM & TopHat2 |
| Methods to align to human genome | MAQ & BLASTN | BLAT & Bowtie | NovoAlign | SMALT | SNAP | BLAT | Bowtie2 | BLASTN | BWA & BWA-MEM & BLASTN | BWA-MEM & TopHat2 |
| Methods to align to viral genomes | BLASTN & BLASTX | BLASTN | NovoAlign | SMALT | SNAP | BLAT | BLASTN | BWA-MEM & BLASTN | BWA-MEM & TopHat2 |
| Methods to align to viral genomes | De novo assembly | Velvet | Trinity | N | ABySS | ABySS & Minimo | N | N | N | N | N |
| Methods to align to viral genomes | De novo assembly | BLASTN | BLASTN | NovoAlign | SMALT | SNAP | BLAT | BLASTN | BWA-MEM & BLASTN | BWA-MEM & TopHat2 |
| Read distribution on viral genome | N | N | Y | N | N | N | N | N | N | N |
| Read distribution on viral genome | N | N | Y | N | N | N | N | N | N | Y |
| Viral integration event detection | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Taxonomy annotation | N | N | N | N | N | N | N | N | N | Y |
| Taxonomy annotation | N | N | N | N | N | N | N | N | N | Y |
| Output format | Text | Text | Webpage | Text | Text & Pic. | Text | Text | Text | Text | Text |
| Output format | Text | Text | Text | Text | Text | Text | Text | Text | Text | Text |
Supplemental Table S25 Feature comparison between VIcaller and other existing viral integration detection software

| Software            | SeqMap2.0 (Hawkins et al. 2011) | SummonChimera (Katz and Pipas 2014) | VirusSeq (Chen et al. 2013) | ViralFusionSeq (Li et al. 2013a) | VirusFinder 1&2 (Wang et al. 2013; Wang et al. 2015) | Virus-Clip (Ho et al. 2015) | Vy-PER (Forster et al. 2015) | SeeksV (Liang et al. 2017) | BATVI (Tennakoon and Sung 2017) | VIcaller |
|---------------------|---------------------------------|-------------------------------------|-----------------------------|-----------------------------------|-------------------------------------------|-----------------------------|-------------------------------|--------------------------------|--------------------------------|-----------|
| Detect presence of viral sequences | -                               | -                                   | Y                           | Y                                 | Y                                         | -                           | -                            | -                              | -                              | Y         |
| Use of chimeric reads to detect viral integration breakpoints | -                               | Y                                   | Y                           | Y                                 | Y                                         | -                           | Y                            | Y                              | Y                              | Y         |
| Use of split reads to detect viral integration breakpoints | Y                               | Y                                   | -                           | Y                                 | Y                                         | Y                           | Y                            | Y                              | Y                              | Y         |
| Able to handle multiply-mapped reads | -                               | -                                   | -                           | -                                 | -                                         | -                           | -                            | -                              | -                              | Y         |
| Able to address sequence repeats | -                               | -                                   | -                           | -                                 | Y                                         | -                           | -                            | -                              | -                              | Y         |
| Able to utilize partial viral reference genomes | -                               | -                                   | -                           | -                                 | -                                         | -                           | -                            | -                              | -                              | Y         |
| Detect integrations of routinely-tested candidate viruses | Y                               | Y                                   | Y                           | Y                                 | Y                                         | Y                           | Y                            | Y                              | Y                              | Y         |
| Detect integrations of undiagnosed viruses | -                               | -                                   | -                           | -                                 | Y                                         | -                           | -                            | -                              | -                              | Y         |
| Virome-wide detection of all possible integrated viruses | -                               | -                                   | -                           | -                                 | -                                         | -                           | -                            | -                              | -                              | Y         |
| Able to detect both upstream and downstream breakpoints | -                               | Y                                   | -                           | -                                 | -                                         | -                           | -                            | -                              | -                              | Y         |
| Calculation of integration allele fraction | -                               | -                                   | -                           | -                                 | -                                         | -                           | Y                            | -                              | -                              | Y         |
| Viral integration characterization and annotation | -                               | -                                   | -                           | -                                 | Y                                         | -                           | -                            | -                              | -                              | Y         |

1. The methods/software used for breakpoint detection in the existing software include: Cluster (SeqMap2.0 and VirusSeq), ClipCrop (ViralFusionSeq), CREST (ViralFusionSeq, VirusFinder 1 and VirusFinder 2), and SVDetect (VirusFinder 1 and VirusFinder 2). The latter three software: ClipCrop (Suzuki et al. 2011), CREST (Wang et al. 2011), and SVDetect (Zeitouni et al. 2010) were designed for structural variation detection. In contrast, VIcaller uses in-house scripts specifically designed for viral integration breakpoint detection.

2. Reads aligned to multiple viruses or locations.

3. Repeat sequences detected by RepeatMasker (Smit et al. 2015), TRF (Benson 1999), and DUST (Morgulis et al. 2006).

4. Existing viruses not routinely tested for.

5. VirusFinder 1 and 2 can detect only one virus at a time.

6. The taxonomy information is obtained from the NCBI.
| Quality controls before viral genome alignment | Trim low-quality nucleotides (< Q20) from the 3’ end of reads |
|-------------------------------------------------|----------------------------------------------------------|
|                                                 | Remove reads that contain > 20% nucleotides with quality scores < Q20 |
|                                                 | Remove redundant reads (e.g., PCR duplicates) |
|                                                 | Remove paired-end reads if either read is < 20 bp |
| Quality controls after viral genome alignment   | If multiple reads map to identical positions on a viral genome, only keep one |
|                                                 | Remove hits (mapping loci) of multiply mapped reads having no uniquely mapped reads within a window of one insert size |
|                                                 | Remove reads containing < 20 bp of non-repeat sequences |
|                                                 | Remove homologous regions by reciprocal (secondary) alignment to the human genome |
| Quality controls after candidate viral integration calling | Remove viral integrations with less than two chimeric or split reads |
|                                                 | Remove integrations that have no chimeric read at either breakpoint |
|                                                 | Remove integrations whose combined supporting reads span < 50 bp of the human genome |
|                                                 | Remove integrations with average alignment score < 30 |
|                                                 | Remove chimeric and split reads with abnormal directions and locations |
References

Adams MJ, Antoniw JF. 2006. DPVweb: a comprehensive database of plant and fungal virus genes and genomes. *Nucleic Acids Res* **34**: D382-385.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-208.

Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D. 2008. The Influenza Virus Resource at the National Center for Biotechnology Information. *Journal of Virology* **82**: 596-601.

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.

Brister JR, Ako-adjei D, Bao Y, Blinkova O. 2015. NCBI Viral Genomes Resource. *Nucleic Acids Research* **43**: D571-D577.

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

Chang S, Zhang J, Liao X, Zhu X, Wang D, Zhu J, Feng T, Zhu B, Gao GF, Wang J et al. 2007. Influenza Virus Database (IVDB): an integrated information resource and analysis platform for influenza virus research. *Nucleic Acids Res* **35**: D376-380.

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.

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

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.

Hsu PW, Lin LZ, Hsu SD, Hsu JB, Huang HD. 2007. ViTa: prediction of host microRNAs targets on viruses. *Nucleic Acids Res* **35**: D381-385.

Hu X, Yuan J, Shi Y, Lu J, Liu B, Li Z, Chen Y, Mu D, Zhang H, Li N et al. 2012. pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics* **28**: 1533-1535.

Huang Y, Lau SK, Woo PC, Yuen KY. 2008. CoVDB: a comprehensive database for comparative analysis of coronavirus genes and genomes. *Nucleic Acids Res* **36**: D504-511.

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

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656-664.

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.

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.

Kuiken C, Yusim K, Boykin L, Richardson R. 2005. The Los Alamos hepatitis C sequence database. *Bioinformatics* **21**: 379-384.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

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.

Lefkowitz EJ, Upton C, Changayil SS, Buck C, Traktman P, Buller RM. 2005. Poxvirus Bioinformatics Resource Center: a comprehensive Poxviridae informational and analytical resource. *Nucleic Acids Res* **33**: D311-316.
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.

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.

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.

Pickett BE, Greer DS, Zhang Y, Stewart L, Zhou L, Sun G, Gu Z, Kumar S, Zaremba S, Larsen CN et al. 2012a. Virus Pathogen Database and Analysis Resource (ViPR): A Comprehensive Bioinformatics Database and Analysis Resource for the Coronavirus Research Community. *Viruses* 4: 3209-3226.

Pickett BE, Greer DS, Zhang Y, Stewart L, Zhou L, Sun G, Gu Z, Kumar S, Zaremba S, Larsen CN et al. 2012b. Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. *Viruses* 4: 3209-3226.

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.

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

Squires RB, Noronha J, Hunt V, Garcia-Sastre A, Macken C, Baumgarth N, Suarez D, Pickett BE, Zhang Y, Larsen CN et al. 2012. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza Other Respir Viruses* 6: 404-416.

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.

Suzuki S, Yasuda T, Shiraishi Y, Miyano S, Nagasaki M. 2011. ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC bioinformatics* 12: S7.

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.
Upton C, Lefkowitz E. Viral Bioinformatics Resource.
Van Doorslaer K, Tan Q, Xirasagar S, Bandaru S, Gopalan V, Mohamoud Y, Huyen Y, McBride AA. 2013. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. Nucleic Acids Res 41: D571-578.
Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L et al. 2011. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods 8: 652-654.
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.
Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S. 2012. FastUniq: a fast de novo duplicates removal tool for paired short reads. PLoS One 7: e52249.
Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-ne P, Nicolas A, Delattre O, Barillot E. 2010. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics 26: 1895-1896.