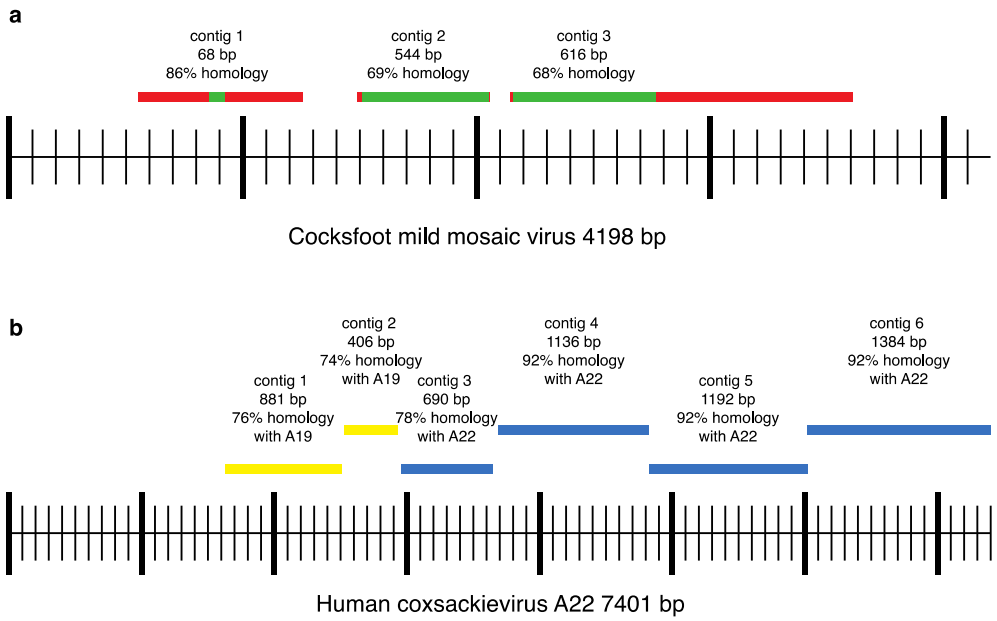
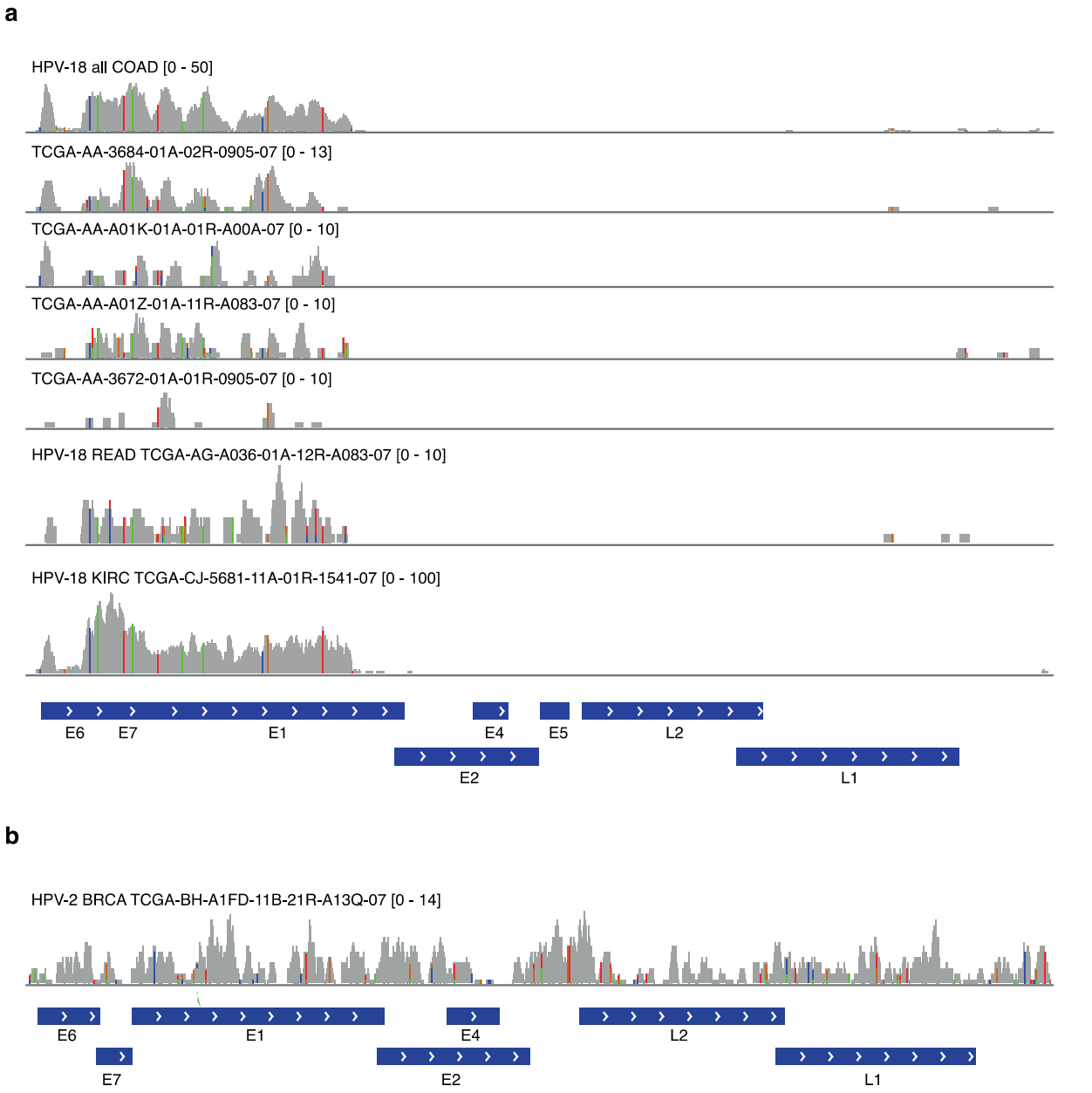
Supplementary Information

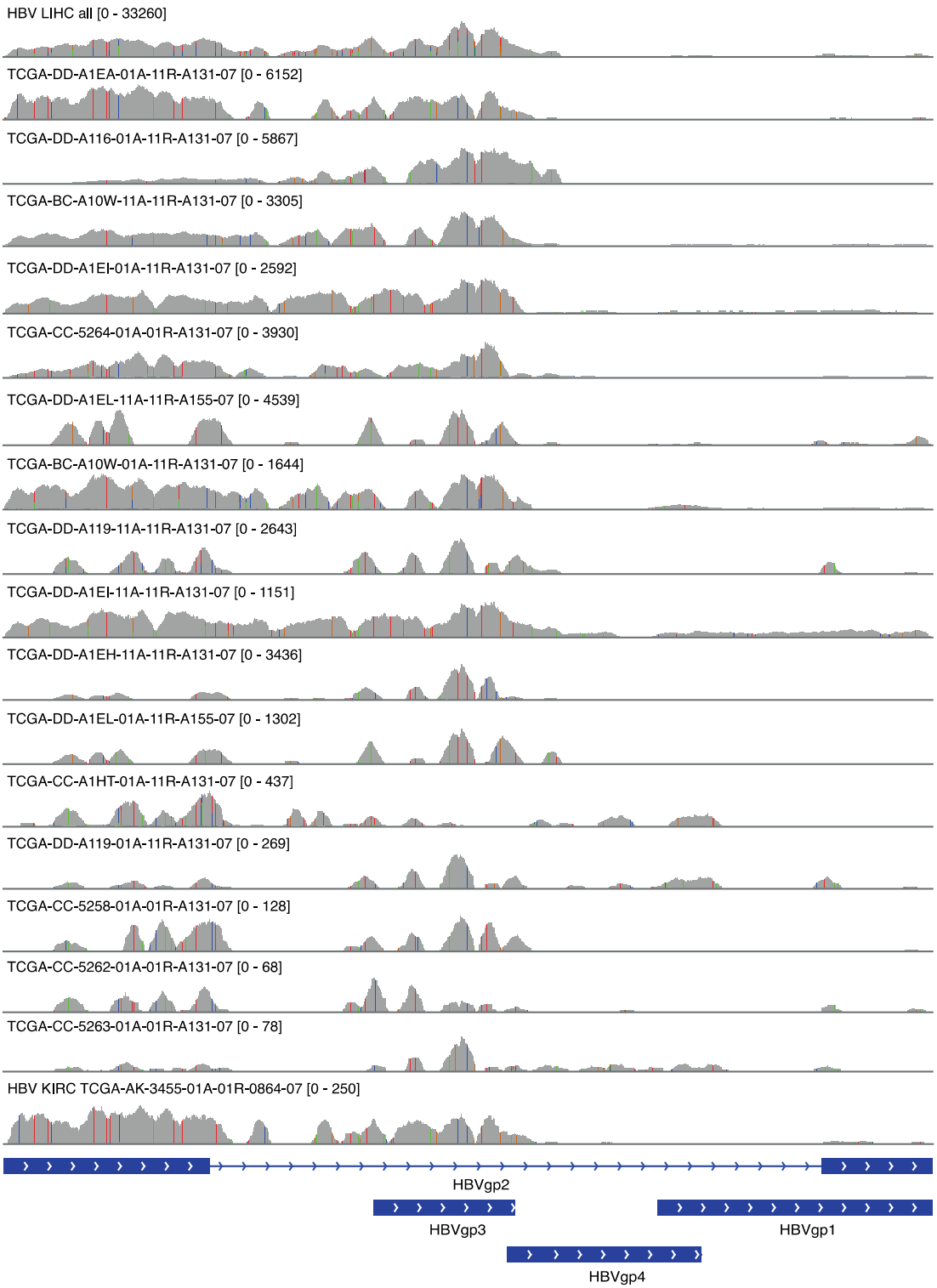
Supplementary Figures



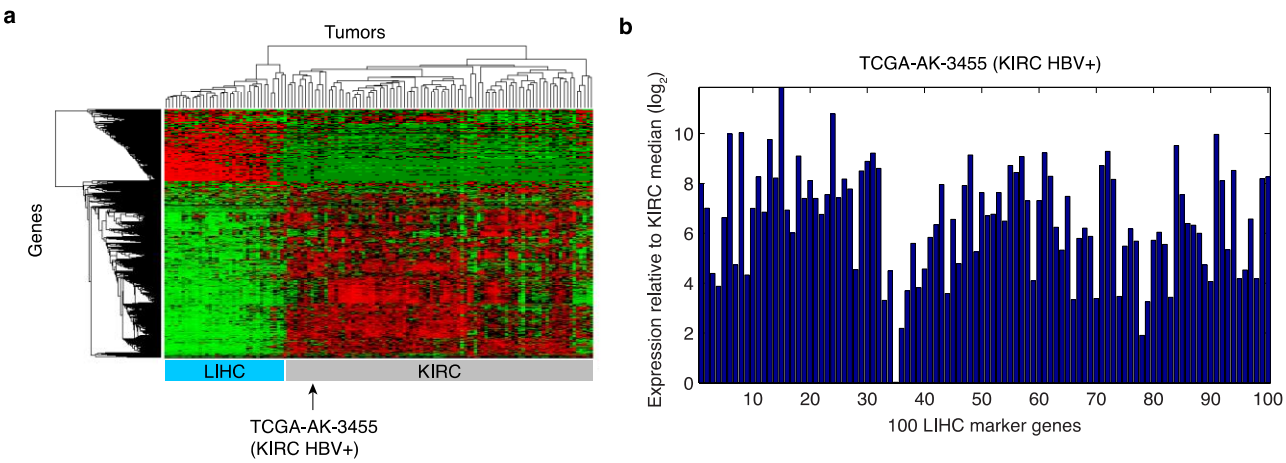
Supplementary Figure S1. Graphical illustration of contig coverage of novel assembled genomes. (a) Contig (red and green bars) positions in relation to the Cocksfoot mild mosaic virus genome (Genbank: EU081018) (CMMV). Red bars correspond to sequences that primarily align to CMMV or other mosaic viruses, but with poor sequence identity. Green bars represent strong homology. (b) Contigs (yellow and blue bars) relative to Human coxsackievirus A22 (Genbank: DQ995647). Contigs 1-2 share homology with A19. Contig 3 displays homology to both A19 and A22, most likely reflecting region conserved between the two viral strains. Contigs 4-6 are more similar to A22 than A19.



Supplementary Figure S2. Graphical illustration of viral gene expression patterns in HPV-positive COAD, READ, KIRC and BRCA. (a) Viral mRNA from all (top row) and individual COAD as well as individual READ and KIRC aligned to HPV18 (Genbank: NC_001357.1). Low count reads (numbers within brackets) and E1 expression indicate few infected cells and a virus in replicative phase respectively. (b) Viral mRNA from BRCA aligned to HPV2 (Genbank: NC_001352.1). High expression levels of capsid-proteins L1 and L2 suggest late stage of infection and assembly of progeny virions. Colored rows in mRNA pattern profile represent nucleotide mismatches to reference sequence.

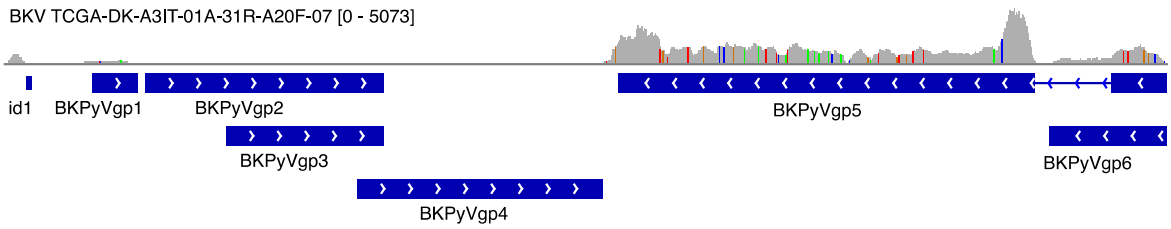


Supplementary Figure S3. Graphical illustration of viral gene expression patterns in HBV-positive tumors. Viral mRNA from all (top row) and individual LIHC and KIRC (bottom row) aligned to HBV (Genbank: NC_003977.1) Majority of reads align to HBVgp2 (S-antigen) and HBVgp3 (X-antigen). Colored rows in mRNA pattern profile represent nucleotide mismatches to reference sequence.

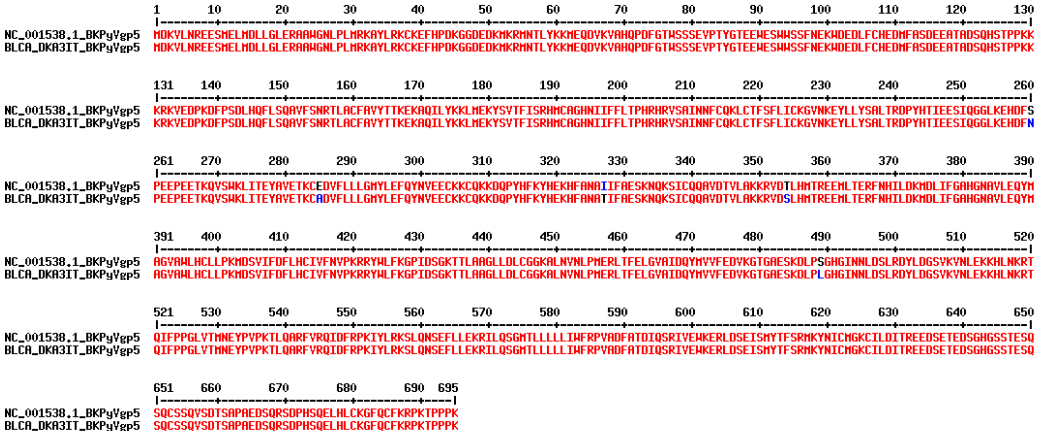


Supplementary Figure S4. An HBV-positive KIRC tumor shows a typical KIRC mRNA profile, but with consistent weak induction of LIHC markers. (a) Unsupervised hierarchical clustering of 34 LIHC tumors and 100 KIRC tumors (Pearson's correlation coefficient, and using top 2500 most variable genes based on standard deviation across included samples). The single HBV-positive tumor (TCGA-AK-3455) had a gross expression profile similar to other KIRC tumors. (b) Closer study of TCGA-AK-3455 revealed consistent induction of LIHC marker genes (top 50 genes induced in LIHC relative to KIRC samples). ND, not detected.

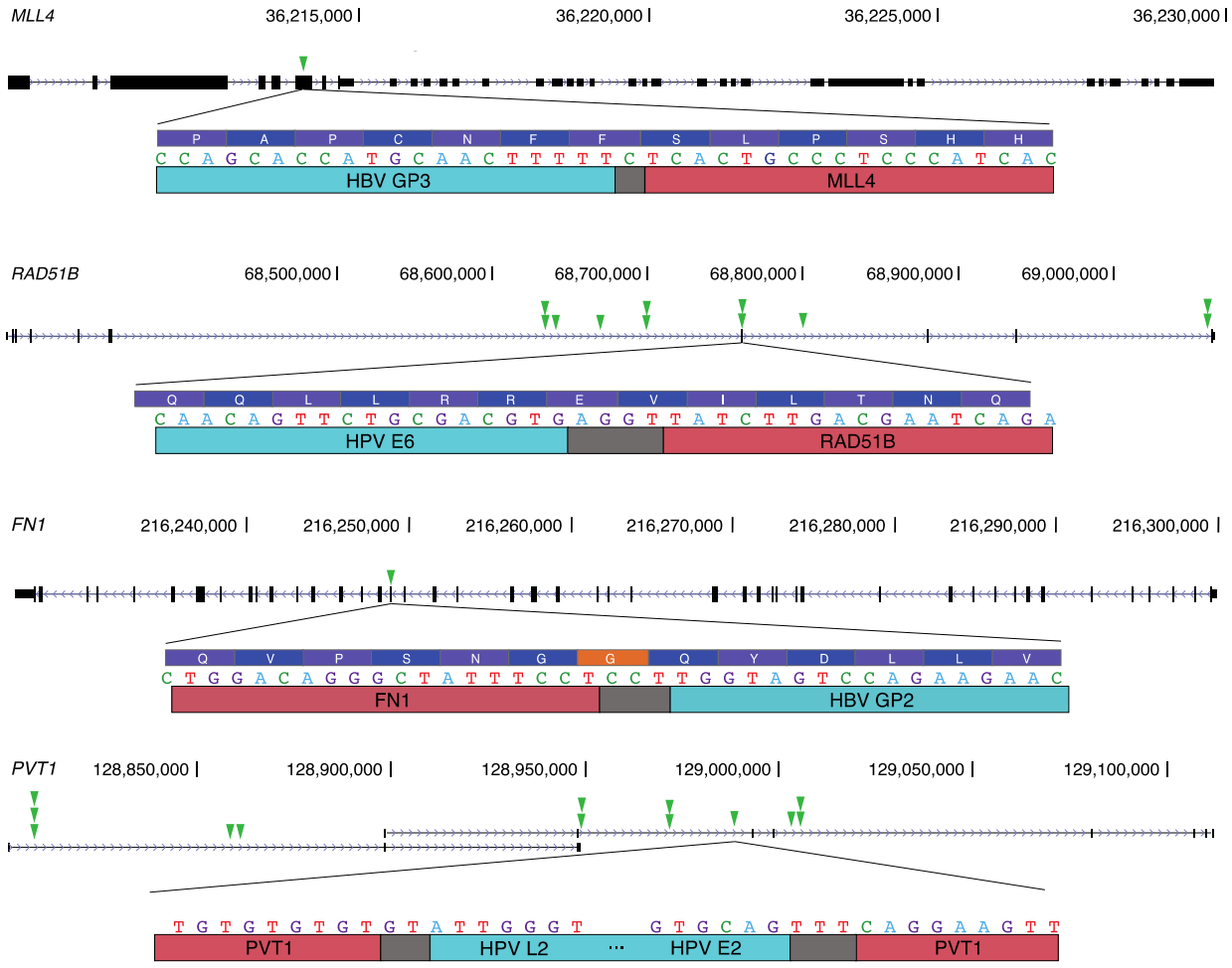
a



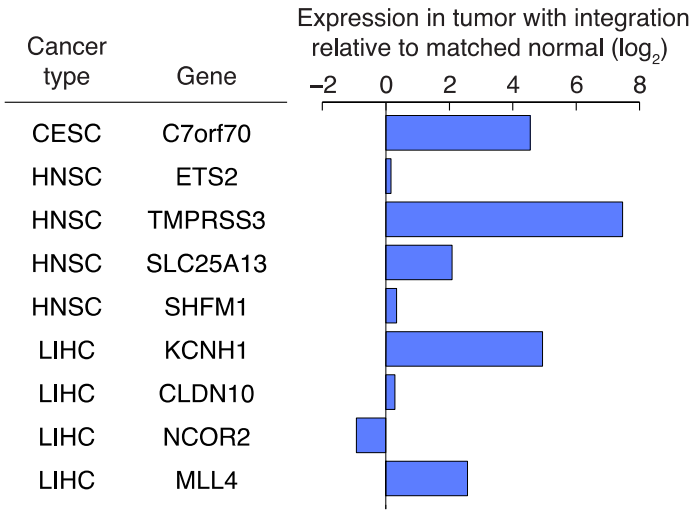
b



Supplementary Figure S5. Graphical illustration of viral gene expression patterns in BKV-positive BLCA and protein sequence of BKPyVgp5. (a) Viral mRNA from BLCA-tumor aligned to BKV (Genbank: NC_001538.1). The oncogenic proteins BKPyVgp6 (small T-antigen) and to greater extent BKPyVgp5 (large T-antigen) are mainly expressed in this tumor. Colored rows in mRNA pattern profile represent nucleotide mismatches to reference sequence. (b) Protein sequence of BKPyVgp5 in BLCA compared to reference sequence display five substitutions, but no truncations.



Supplementary Figure S6. Close-up view of selected fusion sites in recurrent genes. Green arrows indicate fusion breakpoints that were fine-mapped with the help of breakpoint-spanning reads (as presented in Supplementary Table 5). Close-up views are shown for select cases, including in-frame fusions. Light blue, viral transcript; red, human transcript; grey, possible positions for breakpoint; orange, amino-acid introduced during fusion.



Supplementary Figure S7. Relative expression of integration target genes in tumors with integration relative to matched normals without integration. We identified a total of 9 cases of viral integration in gene loci for which a matched normal control without integration in the same gene was available. For these genes and tumors, we calculated relative expression levels in the tumors compared to the normals.

a


[0 - 115036]
TCGA-C5-A3HD-01B-11R-A213-07 [0 - 907]
TCGA-C5-A1M9-01A-11R-A13Y-07 [0 - 730]
TCGA-C5-A1BI-01B-11R-A13Y-07 [0 - 10430]
TCGA-JX-A3Q8-01A-11R-A21T-07 [0 - 6754]
TCGA-FU-A3HY-01A-11R-A21T-07 [0 - 10193]
TCGA-C5-A2LZ-01A-11R-A213-07 [0 - 3994]
TCGA-C5-A3HL-01A-11R-A213-07 [0 - 4453]
TCGA-C5-A2M2-01A-21R-A18M-07 [0 - 3230]
TCGA-EA-A1QT-01A-11R-A14Y-07 [0 - 6076]
TCGA-DS-A0VM-01A-11R-A10U-07 [0 - 4395]
TCGA-FU-A2QG-01A-11R-A18M-07 [0 - 4640]
TCGA-EA-A3HU-01A-11R-A213-07 [0 - 4200]
TCGA-EA-A3HT-01A-61R-A21T-07 [0 - 3371]
TCGA-BI-A20A-01A-11R-A14Y-07 [0 - 2579]
TCGA-C5-A1ML-01A-11R-A14Y-07 [0 - 1611]
TCGA-C5-A2LX-01A-11R-A18M-07 [0 - 1875]
TCGA-EX-A1H5-01A-31R-A13Y-07 [0 - 2879]
TCGA-EK-A2RB-01A-11R-A18M-07 [0 - 6435]
TCGA-EA-A3HS-01A-11R-A213-07 [0 - 2817]
TCGA-DG-A2KL-01A-11R-A180-07 [0 - 2680]
TCGA-DS-A0VK-01A-21R-A10U-07 [0 - 3617]
TCGA-DS-A1OC-01A-11R-A14Y-07 [0 - 1385]
TCGA-DS-A0VL-01A-21R-A10U-07 [0 - 5319]
TCGA-IR-A3LL-01A-11R-A213-07 [0 - 3389]
TCGA-C5-A1MN-01A-11R-A14Y-07 [0 - 2670]
TCGA-EK-A2PI-01A-11R-A18M-07 [0 - 2622]
TCGA-FU-A3NI-01A-11R-A21T-07 [0 - 2847]
id1 E6
E1
E4
E5id3
L2
id5
E7
E2
id4
L1
TCGA-DR-A0ZL-01A-11R-A10U-07 [0 - 1942]
TCGA-EK-A2RL-01A-11R-A18M-07 [0 - 3878]
TCGA-C5-A2M1-01A-11R-A18M-07 [0 - 3253]
TCGA-C5-A2LY-01A-31R-A18M-07 [0 - 2810]
TCGA-DS-A0VN-01A-21R-A10U-07 [0 - 1582]
TCGA-C5-A1MP-01A-11R-A14Y-07 [0 - 1901]
TCGA-C5-A1BJ-01A-11R-A13Y-07 [0 - 5652]
TCGA-EK-A2IP-01A-11R-A180-07 [0 - 1495]
TCGA-DS-A1OD-01A-11R-A14Y-07 [0 - 1763]
TCGA-C5-A2LV-01A-11R-A18M-07 [0 - 3162]
TCGA-EA-A3QE-01A-21R-A21T-07 [0 - 2826]
TCGA-C5-A1BL-01A-11R-A13Y-07 [0 - 1235]
TCGA-BI-A0VR-01A-11R-A10U-07 [0 - 1415]
TCGA-FU-A3EO-01A-11R-A213-07 [0 - 2260]
TCGA-C5-A1MH-01A-11R-A14Y-07 [0 - 2582]
TCGA-EK-A2RN-01A-12R-A213-07 [0 - 1527]
TCGA-C5-A0TN-01A-21R-A14Y-07 [0 - 647]
TCGA-BI-A0VS-01A-11R-A10U-07 [0 - 1331]
TCGA-EK-A2RE-01A-11R-A18M-07 [0 - 1761]
TCGA-EK-A2H1-01A-11R-A180-07 [0 - 668]
TCGA-HM-A3JJ-01A-21R-A21T-07 [0 - 1051]
TCGA-C5-A1BE-01B-11R-A13Y-07 [0 - 790]
TCGA-HG-A2PA-01A-11R-A213-07 [0 - 651]
TCGA-EK-A2RC-01A-11R-A18M-07 [0 - 473]
TCGA-EK-A2RK-01A-11R-A18M-07 [0 - 363]
TCGA-DG-A2KM-01A-11R-A180-07 [0 - 737]
TCGA-C5-A1M8-01A-21R-A13Y-07 [0 - 511]
TCGA-EK-A2RO-01A-11R-A18M-07 [0 - 177]
id1 E6
E1
E4
E5id3
L2
id5
E7
E2
id4
L1

Supplementary Figure S8. Graphical illustration of viral gene expression patterns in HPV-positive CESC. HPV positive CESC mRNA aligned to (a) HPV16 (b) HPV18 (c) HPV31 (d) HPV33 (e) HPV39 (f) HPV45 (g) HPV52 (h) HPV58 (i) HPV59 (j) HPV68b (k) HPV69 and (l) HPV70 (Genbank accession numbers found in Supplementary Table S1). Colored rows in mRNA pattern profile represent nucleotide mismatches to reference sequence.

The landscape of viral expression and integration in human cancer

Supplementary Tables

Annotated viruses	
Virus type	Genbank
Human papillomavirus type 31	J04353.1
Human papillomavirus type 33	M12732.1
Human papillomavirus type 39	M62849.1
Human papillomavirus type 45	X74479.1
Human papillomavirus type 52	X74481.1
Human papillomavirus type 56	X74483.1
Human papillomavirus type 58	D90400.1
Human papillomavirus type 59	X77858.1
Human papillomavirus type 68b	FR751039.1
Human papillomavirus type 69	AB027020.1
Human papillomavirus type 70	U21941.1
Human coxsackievirus A22 strain ban99-10427	DQ995647.1
Assembled viruses	
Assembled genome 1: Coxsackie-like virus	
TAAGGTACCTATTACAATTACTATAGCACCAATGTGCTGTGAGTTTAATGGGCTCAGATCTCTCACAGTGC CTTTCACACAAGGATTACCGGTGATGGCAACACCAGGGTCTAACCAGTTTTTGACATCTGATAACTTCCAA TCGCCATGCGCTCTCCCAGACTTTGATGTTACTCCCGAGATACACATACCAGGGGAAGTGAAGAATATGAT GGAGCTGGCTGAAATAGATACAATTATCCCTATGAATGCAGTCCCAACCAAAGTAAATACTATGGAAGCAT ATCCAATACCTCTAGCAGCAGGGGTGCAGAATAATAACAAATCCATATTTTCAATTAGTTTAAGTCCTGCC GCTGATCAGAGATTGTCTAGAACAATGTTAGGTGAAATTCTTAATTACTACACACACTGGACTGGCTCCAT TAAATTCACATTCCTATTTTGTGGTAGTATGATGGCCACAGGAAAGTTACTGCTGAGTTACAGCCCCCCAG GAGCAAAACCCCCGACGACTAGAAAAGAGGCCATGTTAGGCACACATATTGTGTGGGATTTAGGCTTACAA TCCAGTGCAACTATGGTCGCGCCTTGGATTTCAAATGTGAACTACAGGCGTTGTGTCAAAGATGATTTTAC AGAAGGTGGTTACATATGCTGCTTCTACCAGACTGCAATTGTAGTGCCTCCAAACACTCCAACGGACATGT ATATGCTTGCATTTGTTAGTGCGTGTAATGACTTCTCAGCAAGGCTTTTGAAAGACACTCCATTTGTCGTC CAGACCACACCTGTGGCTCGCACTCAGGGGATTGATGATGTTATTGACACAGTCGTAGCTAATGCACTTAA	

GGTGTCCATGCCACAAGTACAAGACACCCG

TCAACTCAAAAGAAGTCCCTGCGCTAACTGCTGTTGAGACAGGTGCCACCAGTCAAATAGAGCCTTCAGAT
GTGATCGAGACGCGCCATGTCATAAACCAAAGGTTAAGATCAGAATGCACCGTAGAATCCTTCTTTGGAAG
ATCTGCTTGTGTGGCCATAATAGGTTTGAGCAACAAGAAGCCCACAGACACCAACGGTAAAGAATTGTTTG
CAACATGGCCTATCTCATATCTAGACACATATCAATTGAGGAGAAAGTTGGAGATGTTCACTTACTCTAGA
TTTGACATTGAGATGACTTTTGTTGTAACAGAGAGATTCTTTACCTCAACATCTGCAGCTGCTAGGGACTA
TGTATACCAGATCATGTATGTGCCACCAGGTGCCCCAATCCCCCAGTCGTGG

CAACCCTTCAATATTTTATACAACTGGAAATGCCAGTCCAAGGATTTCGATCCCGTTTGTAGGCATTGCTG
CAGCTTACTCACACTTTTATGATGGTTTCTCCGTAGTCCCTTTCAATTCAGTGGATGCTGGTGCATCCAAC
AAGTATGGCTATTCTTCAGTCAATGACTTTGGCACACTAGCGATCAGGATTGTTAACGAATATGACCCAGT
AACTATAGATGCAAAAGTTAGGGTGTACCTCAAACCCAAACATGTAAAAGTGTGGTGCCCCAGACCGCCCA
GAGCTGTAGCTTACAATGGACCAACTGTGAACTTTTCAGAAAACCCATCAGTCATGACACAGGTTGCTGAT
ATCAGAACTTATGGATTTGGACATCAAAACAAAGCAGTCTACACTGCTGGGTACAAAATTTGTAATTATCA
TTTAGCCACACCTGAAGACATGGAAAAAGCTGTCAATATCATGTGGGACAGAGATCTCCTAGTCTGTGAGA
GTGGTGCTCAAGGCACCGATACCATTGCAAGGTGTTCATGCAAAGCAGGAGTCTACTATTGTGAGTCTAAA
AGAAAGTATTATCCAGTTACGGTTTGTGGACCCACATTCCAGTATATGGAGGCTAATGACTTCTACCCACC
TAGATATCAATCCCACATGTTAATTGGTTATGGCTTTGCCAACCCTGGTGAC

CAACATGGAGTAATGGGCATCATTACTGCTGGTGGTCAAGGGGTTGTGGCTTTTGCAGACATTAGGGATTT
GTACATGTATGAGGAGGAAGCTATGGAACAAGGTGTTTCAGACTATGTTAATAGGTTAGGAATGGCATTTG
GTGCCGGTTTTAGTGCTGAAGTAGCAAACAAAATTTCAGAGATTCAAACCACGGTCCAGAGTGTCCTTACA
GAAAAGCTCCTGAAGAATCTAATAAAAATTGTTTCAGCTCTTGTTATTGTTGCCAGAAATTATGAAGATTC
GATTACAGTACTAGCCACACTTTCTCTATTAGGATGTGACGCCTCGCCGTGGCAATGGCTCAAGGAAAAAC
TCTGTAATATGATTGGCATCCCCTACGTCATGAAACAAGGAGATACTTGGCTAAAAAAGTTCACAGAAGCT
TGCAATGCCGCCAAGGGGCTGGAGTGGATTGCAAATAAAATTAGTAAATTTATAGATTGGATCAAGGAGAA
AGTGCTCCCAGAGGCAAAGGACAAATTGGAGTATCTTTCAAAGATGAAACAACTTGAAATGATAGAAAACC
AGATGGCCACGTTACATCAATCATGCCCTAGCCAAGAAGAGCAGGAAGTTCTCTTTAACAACATCAGATGG
TTAGCAATAAAGGCAAGAAAGTTTGCGCCCTTGTATGCAGCTGAGGCCCGTAGAATATTTAAGCTAGAAAC
ATCCATCAACAATTATGTACAATTCAAGACCAAACACCGCATTGAACCAGTATGTTTGTTGATTCATGGAA
CCCCCGGTACAGGCAAGTCGGTTGCTACAGGCCTCATTGGAAGATCAATTGCAAAGCAAGCAAACACCAGT
ACTTATTCACTGCCCCCGGATCCTTCACATTTTGATGGCTATAAGCAACAGGGAGTTGTGATTATGGATGA
CCTGAACCAAAACCCAGATGGGGAAGATATGAAGTTGTTTTGTCAAATGGTCTCCACAGTGGAGTTCATAC
CACCAATGGCCAGCTTGGAGGAAAAGGGAATATTGTTCACCTCAGATTATGTGCTTGCTTCCACCAATTCT
AACACAATAACACCTCCCACAGTATCAGCATCTGATGCATTATCTAGAAGATTTGCTTTTGACATGGACAT
A

ATACCCAATGTCTGAATATACAACAAAAGGAAAGTTGAACATGGCTTTGGCTACCCAATTGTGTAAAGACT
GCCATAAACCAGCAAACTTCTCTACATGCTGCCCCTTGGTGTGTGGGAAAGCTATTCAATTGATGGACAAA
AATACCAGGATGAGATACTCTCTGGACCAGATTACCACTATGATGATTAACGAGAGAAATAGGAGGTACAA
CATTGGAGGCTGCTTAGAGGCACTATTCCAAGGCCCAATTGAATTCAAAGATCTCAAGATCGATGTGGTTT
CCACACCACCGCCCTCAGCAATATCAGACCTGCTCAAATCAGTAGATAATCAGGAGGTTAGGGATTATTGC
AAACAACAGGGTTGGGTGGTGGAGATCCCTTCAAATTCTGTCACCATTGAAAGACACATCTCCCGTGCAAC
TAGTATTTTACAATCACTCTCCACTTTTGCAGTGGTGGCGGGTATGGTCTATGTAGTTTACAAACTTTTTG
CAGGATTTCAGGGCGCATACACAGGACTCCCTAATGCAAAACCAAAAATCCCTACAATTAGAGCAGCCAAA
GTTCAAGGACCTGTATTTGATTATGCAGTTGCCATGGCTAAGAAGAATATACTCACTGCTACCACAGAGAA
GGGTGAATTCACAATGCTGGGTGTCTATGACAGGGTAGCTGTGCTTCCTACTCACTCAAACCCAGGTGAAA
CCATAGTTATATCAGGAAAAGAAGTGAAGATCTTAGATGCTAAAAGATTAGTTGATAGTGATGAAGTTAAC
CTTGAAATTACTATTGTAACCTTAGACAGAAATGAAAAGTTTAGGGATATTAGAACTCACTTACCAACCCA
AATCCATGAAACTAATGATGCAGTCCTTGCCGTAAACACCTCCAAATTTCCTAACATGTACATTCCAGTTG
GATCAGTTATCGAACAGGGTATGCTAAATCTTGGGGGTAGGCCCACAAACAGAACCTTGATGTATAACTTC
CCAACCAAGGCAGGACAGTGTGGTGGTGTTCTGATGGCAACTGGCAAAGTGATTGGCATACACGTTGGTGG
CAATGGATCACATGGTTTTGCTGCCGCTCTCAAGAGAAGTTACTTCACTGAAGAACAAGGGGAGATCCAGT
GGATGAGGCCAAGCAAAGATGCAGGTTATCCCATGATTAATGCACCCTCAAAGACAAA

AAACTAGAACCTAGTGTTTTCTTTGATGTTTTCCCAGGAGAAAAAGAACCAGCTGCTTTAACCAAGAATGA

TCCTAGACTGAAAGTTGATTTCGAAGAAGCCGTCTTTTCCAAATACATTGGGAATAAAATTACAGAGGTGG
ATGAGTACATGAAGGAGGCTGTCGATCACTATGCTGGTCAGCTGATGTCCCTTGATATACCTACAGAACAG
ATGTGCCTGGAGGATGCCATGTATGGCACAGATGGTCTTGAAGCCCTAGATTTGACAACTAGTGCTGGGTA
CCCCTATGTGGCTATTGGAAA
GAAAAAGAGAGACATTCTCAACAAACAAACACGAGACACCAAAGAAATGCAGAAAATGTTAGATAAATATG
GAATAAATCTACCATTGGTGACTTATGTTAAAGATGAACTTAGATCCAAATCAAAAGTGGAACAGGGTAAA
TCTAGATTGATTGAAGCAAGCTCTTTGAATGATTCGGTTGCTATGAGACAGGCTTTTGGTCACCTATATGC
CAAGTTTCACCAGAATCCAGGGATAATAACAGGCTCTGCAGTGGGGTGTGACCCAGATGTCTTCTGGAGTA
AGGTGCCAGTGATGTTAGATGGAGAACTCTTTGCTTTTGATTACACAGGTTATGATGCTTCACTCTCTCCA
GCTTGGTTTGAAGCCTTGAAAATGGTACTGGAAAAGATTGGCTTTGGTGATCGAGTTGACTTCATAGACTA
CTTAAATCACTCACACCACTTATACAGGAACAAATTATACTGTGTTAAAGGTGGAATGCCCTCAGGGTGCT
CTGGCACCTCAATATTCAATTCCATGATCAATAACTTGATAATTAGGACACTAATGTTAAAAACTTACAAA
GGGGTGGATCTAGACTCTCTCCGTATGGTAGCATACGGGGATGACGTCATTGCATCCTACCCACACAAAAT
TGATGCTGGCCTCCTAGCCCAAGCAGGAAAAGACTATGGATTAGTCATGACACCAGCAGATAAGGGTGCTA
CCTTCACAGATGTGGATTGGAGCAATGTAACATTCCTCAAAAGATTCTTCAGAGCAGATGAGCAGTATCCT
TTCCTTGTTCACCCAGTGATGCCCATGAAAGACATTTATGAATCAATCAGGTGGACCAAAGATCCACGAAA
CACACAGGATCATGTCCGATCATTATGCTTATTGGCTTGGCACAATGGAGAAGAAACTTATAACAAATTCC
TAGCTCAAATTAGAAGCGTGCCAGTTGGCCGTGCTCTACTCCTACCGGAGTATTCTACTCTGCGCAGGCGC
TGGCTCGATTCTTTTTAGAGTGACCCTACCTCACCCGAATTGGCTTGGGTTATGGTGTTGTAGGGGTAAAT
TTTCTTTTAATTCGG

Assembled genome 2: Mosaic-like virus

GCTCCTATGGAGTGGCGGGGGGAGCCATCACAGCTCTCACCACCACCAAGCATCTTGGGAGTAAACTCGTG
CTCGGAGTCCTAACCGCTGCTAGCATCTATTGTGCAGTCCAGTTGCACATATATGTTTCGTTGGCTAAAAG
GCGGCAATTCGATAAGACTCGGTCGGACTTCCTTGAACAACAATACCAGGACATCCTCAAATCAATGGAGG
AGCAGAAAGAAAGGGCAACCAGAGAACGGCTAGAAGACTGCCTCACCAAGGAAGTGCTGGAACCGGCCCAA
TTAGCAGAAGACGGAACTGTGCTTAAACAAGAAGTGGCACAATACCTTGTCCACGAACACGGTAAGTTTGT
CAGGGCCTTAGTCTGCGAGGCTAAGATGGAATTTGGGGGAACCCCAAAGACCACCGAAGCCAACCAACTTG
CTGTATGGAGGTTCCTGTACCGAACCTGCGACAAGAGAGGGTTGAATCCAGCTGATGCTCAAAGAAGTATC
ACAGCTGCCCTTCCATTCGTCTTCTTGCCCAGTGCCTATGACGAGTCTGCTGCTATCGGGAGAACTTCAGA
GGAGACTGCAGAGGCTCTTCGTAGGTACAAGTCTCAATTCACCCAAGACACACCACTTCAAAAATTAGTGT
GCAACCCGCTTTCAGGGAAAGCTTGGAGGGCATGGGCACGTAACGTGTTCTATGGAGACCAGG

GAGAGTTGGACGAAGGTCTCCCATATCGTATGAGAAGTTCCTTTCCTACTACACTGGTGGTAAACTCACTA
CGTATAGTAGGGCTGTTGATTCCCTCACTGAACGACCAGTCAACAAGGCTGACGCCAAGCTCTCGACATTT
GTCAAAGCGGAGAAACTAAACTTGTCTCTGAAGTCTGATCCAGTCCCGCGTGTCATCCAACCCAGACACCC
CAGGTACAATGTGGAAGTTGGTAGATATCTCAAACCAATAGAGCACGACATCTACACTGCTATCGACGGGT
TATTTGGGTCCAAAACCATTTTCAAGGGACTGTCTGTCGAGGCAATGGGTTCTCTGATCCACCAGAAAATG
CGCAAATTCTTGAGGCCGTGCGCTATTGGATTCGACGCCTCTCGCTTCGACCAACACGTGTCTGTAGATGC
TCTGAAGTATGAACACTCCATTTACAAGGGAATCTATTCCCACTCCAAAACCCTCAACACCCTACTCAAAT
GGCAGATCCATAACAAAGGGGTGGCAATCGCAAAGGATGGGTTTTTCCGATATTCTGTGGATGGGTGCC

GCTGAACTCATCAACAACGGTGATGACAATGTGTTAATTTGCTCGGAGGATGATGAGGAAGTAGTGAAGAG
ACACTTGTATGATCACTGGCTCAAGTATGGGTTTGAAGTGGTTGCGGAAGAGCCTGTATATATAACAGAAC
AAGTAGAGTTTTGCCAGATGAAGCCTGTGTTTGATGGAACCAACTACGTCATGGTTCGGAAACCCGATGTC
TCGATGTCCAAAGATTGTCATAGCATCACCCCATTTTACACTACCAAAACTGCCAAGAAGTGGGTGCATGC
CGTCGGAGAATGTGGGCTATCTCTAACTGGAGGCATACCAATCAAGCAAGAGTATTACACGTGTATGATCA
GAAATGGCGACCAACATGGGGAAATTGACAAAAGTAAGGAGTTCATTTCAGGGTTTACTCGGCTCAGCAAG
TGCAGCAATCGCAAGTATCGACAAATTTCTAGCGAGACCCGATACTCGTTCTACCTGGCCTTTGGTTACAC
ACCAGACGAACAGGTAGCTATCGAGAACTACTTCCGGACGCTAGAGTTGCCATGGCACTATGGGCTGTCGG
GTACTCCGGCAAGAGCACCTGAATGTCTACTCCTAACTACGATCCCCCAACCACCGACGTTCAACAAGTCA
ACACCACAAAGGAGCCACGAGCCCAGCGACAATCAAGCAGGCAACAACTCGCGCACAGCGCCAGCAGAGCA
GTGGGACAGAGGACTAACAGTTCTCAGGACACTGCGTCTGCCAATTTCGTGATTGTCGCAGAAAGGGTCGA
AGTCACCAACAATTTCAACTTCTAGGCTGGCTACAGGTAAGTGCCACTGTCCAGATTCTCTCTGGACACCA
GGACTGCTTGCCATCTGCATAGTCTTAGCCCTTTGCTTGGTGGCTAGATCCACATCTGAACCTCCCATCAT
CTCACCTCCAGTCTTCCATACTGTATACCATTGTGAGAAATACCAGAACATCGAGGTTCAGAAATGAATGG

```
CTCCCGTGGTCGTCGTCAACAACCTGCCCCCTCCCGGAACAGTGGCGGCCAGAGAAATCAAGGTTCTCGGC
GCAAGAGTCGTAGAAGCAGATCTGTGGAACGAGCAATGCCAGTTGCCACCTCGTACTCTGTGGGTCCGACG
GGACCCCCGGGCATGGGGAGTAGACAGGGATGGACAAGTTTGGCTCACAAGGAAGTCATACTTCAAGTCAC
TGCTAGTACCAGCAGTGACACCATACTCACCATCCCCGTCATTCCAAACCTGCTCTATCCGCCAGACAGTA
CGACTTATTCCGGACGTGCGAAATTCCTCGCAGGGCACGCCCCACTGTACTCCCAACACAAATGGGACATG
TTGGCATTCCAATGGACACCAAGTTGTCCAACGACTACACCTGGAAACGTGGTTTTGAGGTTCATACCCAA
CTACACCACGCCAACACCAACCAACATGTTAGACACAATGGACAGT
```

Supplementary Table S1. Additional missing genomes added to complement the RefSeq viral genome database. Annotated viruses with Genbank accession numbers (top panel) and assembled viruses with contig-sequences (bottom panel).

TCGA cancer abbr.	Libraries	Unique samples	Unique tumors	Unique normals	Unique patients	Read length (nt)	Total reads	Average reads per library	Total non-human reads	Average non-human reads	Fraction non-human candidate reads (%)	Total viral aligned reads
BLCA	109	109	96	13	96	2x48	1.74E+10	1.59E+08	1.17E+08	1.07E+06	0.7	160094
BRCA	914	914	810	104	805	2x50	1.55E+11	1.69E+08	2.11E+09	2.31E+06	1.4	3722
CESC	89	89	87	2	87	2x48	1.68E+10	1.89E+08	1.99E+08	2.24E+06	1.2	3232205
COAD	196	194	194	0	194	1x51	5.07E+09	2.59E+07	2.30E+08	1.17E+06	4.5	26745
GBM	169	167	167	0	161	2x76	2.28E+10	1.35E+08	1.67E+09	9.89E+06	7.3	635
HNSC	345	341	304	37	304	2x48	5.97E+10	1.73E+08	2.84E+08	8.24E+05	0.5	1540524
KIRC	519	519	453	66	462	2x50	9.18E+10	1.77E+08	1.70E+09	3.27E+06	1.9	9461
KIRP	74	74	59	15	62	2x48	1.32E+10	1.78E+08	6.96E+07	9.41E+05	0.5	661
LAML	167	167	167	0	167	2x50	2.08E+10	1.25E+08	6.92E+08	4.14E+06	3.3	2823
LIHC	54	54	34	20	36	2x50	8.07E+09	1.49E+08	6.36E+07	1.18E+06	0.8	433531
LUAD	413	412	355	57	355	2x48	6.04E+10	1.46E+08	3.77E+08	9.14E+05	0.6	2494
LUSC	238	237	220	17	221	2x50	4.29E+10	1.80E+08	4.07E+08	1.71E+06	0.9	25623
OV	420	419	419	0	413	2x75	7.74E+10	1.84E+08	1.93E+09	4.60E+06	2.5	6854
PRAD	179	179	140	39	142	2x48	2.99E+10	1.67E+08	1.19E+08	6.66E+05	0.4	484
READ	71	71	71	0	71	1x76	1.90E+09	2.68E+07	7.52E+07	1.06E+06	4.0	1403
SKCM	249	249	249	0	247	2x48	4.20E+10	1.69E+08	1.41E+08	5.67E+05	0.3	4642
STAD	43	43	43	0	43	2x50	6.67E+09	1.55E+08	5.50E+07	1.28E+06	0.8	57463
THCA	279	279	249	30	261	2x48	5.07E+10	1.82E+08	5.36E+08	1.92E+06	1.1	935
UCEC	321	320	316	4	316	1x76	1.00E+10	3.12E+07	5.07E+08	1.58E+06	5.1	9450

Supplementary Table S2. Additional statistics and information about RNA-seq libraries included in this study. RNA-seq libraries were generated by the TCGA using the Illumina Tru-Seq protocol, on tissue samples selected based on stringent quality requirements (>80% or >60% tumor nuclei, see http://cancergenome.nih.gov/cancersselected/biospeccriteria). Paired-end data was generated on the Illumina HiSeq 2000 platform, while single-end reads (COAD, READ, and UCEC) were generated on the Illumina Genome Analyzer IIx. Read lengths varied between 48 and 76 nt, but were quality-trimmed before further analysis as described in Methods.

The landscape of viral expression and integration in human cancer

ID	HBV genotype
TCGA-BC-A10W-01A-11R-A131-07 Contig: ”17270”	C2
TCGA-CC-5258-01A-01R-A131-07 Contig: ”15238”	B4
TCGA-CC-5258-01A-01R-A131-07 Contig: ”15244”	B4
TCGA-CC-5262-01A-01R-A131-07 Contig: ”8683”	B
TCGA-CC-5262-01A-01R-A131-07 Contig: ”8685”	B
TCGA-CC-5263-01A-01R-A131-07 Contig: ”12025”	B
TCGA-CC-5263-01A-01R-A131-07 Contig: ”12029”	B
TCGA-CC-5263-01A-01R-A131-07 Contig: ”12051”	B4
TCGA-CC-5264-01A-01R-A131-07 Contig: ”10755”	C1
TCGA-DD-A119-01A-11R-A131-07 Contig: ”14856”	B2
TCGA-DD-A119-01A-11R-A131-07 Contig: ”14860”	B2
TCGA-DD-A1EI-01A-11R-A131-07 Contig: ”21462”	C2
TCGA-DD-A1EI-01A-11R-A131-07 Contig: ”21452”	C2
TCGA-DD-A1EL-01A-11R-A155-07 Contig: ”3970”	A
TCGA-DD-A1EL-01A-11R-A155-07 Contig: ”3974”	A2
TCGA-CC-A1HT-01A-11R-A131-07 Contig: ”112109”	E or G
TCGA-CC-A1HT-01A-11R-A131-07 Contig: ”112077”	C or B
TCGA-CC-A1HT-01A-11R-A131-07 Contig: ”112009”	C5
TCGA-DD-A1EA-01A-11R-A131-07 Contig: ”13008”	C1

TCGA-DD-A1EA-01A-11R-A131-07 Contig: ”12958”	C1
TCGA-DD-A116-01A-11R-A131-07 Contig: ”25740”	C2
TCGA-AK-3455-01A-01R-0864-07 (KIRC) Contig: ”18800”	C1
TCGA-AK-3455-01A-01R-0864-07 (KIRC) Contig: ”18810”	C1

Supplementary Table S3. HBV genotype classification. Contigs assembled from HBV-positive LIHC and KIRC were phylogenetically analysed. Individual phylogenetic trees were constructed by maximum likelihood analysis after bootstrapping to 1000 replicates using the MEGA5 software (Tamura-Nei substitution setting, bootstrap values shown at nodes). The comparison includes published sequences representing all HBV genotypes and relevant subgenotypes.

	RNAseq-based clusters, >= 10 unique reads required					WGS-based clusters, >=2 unique reads shown, low-confidence clusters (<4) displayed in italic				
TCGA barcode	Position	Width	Unique reads	Total reads	Support in WGS	Position	Width	Unique reads	Total reads	RNA-seq overlap
TCGA-BA-4077	chr14:68699685-68700029	344	111	485		chr14:68699686-68699986	300	25	26	Yes
	chr14:68701635-68701835	200	77	579		chr14:68741265-68741699	434	64	82	
	chr14:68703891-68704286	395	121	865						
	chr14:68706681-68706868	187	35	49	Yes					
	chr14:68741465-68741699	234	39	77						
	chr14:68758598-68758699	101	38	146						
	chr14:68878139-68878302	163	33	104						
	chr14:68963844-68964298	454	22	28						
TCGA-BA-5153						*chr1:246154478-246154615*	*137*	*2*	*2*	*No*
TCGA-BA-5558						*chr1:177235263-177235316*	*53*	*2*	*2*	*No*
TCGA-BA-5559	chr1:67217967-67218254	287	17	18		chr1:67229037-67229256	219	7	7	
	chr1:67220192-67220460	268	38	79		chr1:67229822-67230084	262	12	12	Yes
	chr1:67221328-67221428	100	16	18		chr1:67230867-67231062	195	4	4	
	chr1:67229016-67229254	238	52	97						
	chr1:67229674-67230085	411	46	91	Yes					
	chr1:67230825-67231081	256	16	21						
	chr1:67231402-67231610	208	18	23						
	chr1:67236068-67236326	258	84	290						
	chr1:67241961-67242088	127	13	16						
TCGA-BB-4223										
TCGA-BB-4225	chr2:89156699-89157141	442	15	18	No*					
TCGA-CN-4741	chr11:76091791-76091920	129	17	29	Yes	chr11:76087547-76087873	326	8	8	
	chr13:73619740-73619879	139	19	55	Yes	chr11:76095856-76096214	358	23	24	Yes
	chr3:189239616-189239985	369	58	140	Yes	chr13:73619741-73619860	119	5	5	
						chr13:73620066-73620338	272	16	17	Yes
						chr3:189239695-189239981	286	7	7	Yes
TCGA-CV-5971	chr5:84619761-84620324	563	26	46		chr20:33629909-33630290	381	49	55	
	chr5:84624624-84625143	519	111	398	Yes	chr20:33638325-33638718	393	40	45	No
						chr21:30912638-30912969	331	13	13	
						chr21:30933491-30933624	*133*	*2*	*2*	No
						chr5:84619616-84620400	784	20	21	Yes
TCGA-CV-6939	chr7:95783231-95783547	316	36	43	Yes**	chr7:96291514-96291703	189	4	4	Yes
	chr7:96251630-96251952	322	46	68						
	chr7:96274882-96275252	370	77	159						
	chr7:96279540-96279832	292	29	40	Yes					
	chr7:96282120-96282524	404	19	22						
	chr7:96285739-96285900	161	15	19						
	chr7:96291613-96291706	93	13	18						

*Known integration locus (IGKC gene)
**Supported by one discordant read pair in WGS

Supplementary Table S4. Comparison of integrations detected using RNA-seq with DNA-level results in 9 virus-positive HNSC tumors with available whole-genome sequencing data. 9 HPV-positive HNSC tumors had available low/medium-coverage whole-genome sequencing (WGS) data (297.4 to 557.9 million reads per sample). WGS datasets were processed with the same pipeline used for RNA-seq. 8/9 RNA-seq-based integrations were supported by at least one discordant read pair in WGS. WGS-based clusters with >=2 unique reads are indicated in the table. Integration clusters were grouped by genomic locus (indicated by light/dark blue).

Gene	Virus:gene	Hg19 gene:region	Hg19 breakpoint	Virus breakpoint	Correct orientation	Frame	Fused viral end	Sample ID
MLL4	HBV:GP3	MLL4: Exon	chr19:36214028	1827	Yes	In-frame	3'	TCGA-CC-5258-01
FN1	HBV:GP2	FN1: Exon	chr2:216248908	460	Yes	In-frame	3'	TCGA-BC-A10W-11
PVT1	HPV:E2	PVT1:Intron	chr8:128869002	3821	Yes	Non-coding	5'	TCGA-BB-7866-01
	HPV:E1	PVT1:Exon	chr8:128867401	883	Yes		3'	TCGA-BB-7866-01
	HPV:L2	PVT1:Intron	chr8:128993412	4385	Yes		5'	TCGA-C5-A1M9-01
	HPV:E2	PVT1:Intron	chr8:128993412	3117	Yes		3'	TCGA-C5-A1M9-01
	HPV:E1	PVT1:Intron	chr8:128944606	881	Yes		3'	TCGA-C5-A1M9-01
	HPV:E6	PVT1:Intron	chr8:128944606	227	Yes		3'	TCGA-C5-A1M9-01
	HPV:E1	PVT1:Exon	chr8:129001407	881	Yes		3'	TCGA-C5-A1M9-01
	HPV:E6	PVT1:Intron	chr8:129001858	217	Yes		5'	TCGA-C5-A1M9-01
	HPV:E1	PVT1:Intron	chr8:129001858	880	Yes		5'	TCGA-C5-A1M9-01
	HPV:L2	PVT1:Exon	chr8:128806916	4911	Yes		5'	TCGA-C5-A2M1-01
	HPV:E2	PVT1:Exon	chr8:128806916	3739	Yes		5'	TCGA-C5-A2M1-01
	HPV:E1	PVT1:Intron	chr8:128978455	2118	No		5'	TCGA-C5-A1MJ-01
	HPV:E1	PVT1:Intron	chr8:128978455	2118	No		5'	TCGA-C5-A1MJ-01
LOC727677	HPV:E6	LOC727677:Intron	chr8:128404748	226	Yes	Non-coding	5'	TCGA-EK-A2H1-01
	HPV:E1	LOC727677:Intron	chr8:128404748	877	Yes		5'	TCGA-EK-A2H1-01
	HPV:E2	LOC727677:Intron	chr8:128404748	3631	Yes		5'	TCGA-EK-A2H1-01
	HPV:E6	LOC727677:Exon	chr8:128302307	236	Yes		5'	TCGA-FU-A23K-01
	HPV:E1	LOC727677:Exon	chr8:128302307	932	Yes		5'	TCGA-FU-A23K-01
	HPV:E1	LOC727677:Exon	chr8:128302307	1361	Yes		5'	TCGA-FU-A23K-01
	HPV:E6	LOC727677:Exon	chr8:128302306	222	Yes		5'	TCGA-FU-A23L-01
	HPV:E1	LOC727677:Exon	chr8:128302306	932	Yes		5'	TCGA-FU-A23L-01
	HPV:E6	LOC727677:Intron	chr8:128312059	222	Yes		5'	TCGA-FU-A23L-01
	HPV:E1	LOC727677:Intron	chr8:128312059	932	Yes		5'	TCGA-FU-A23L-01
RAD51B	HPV:E1	RAD51B:Intron	chr14:68646030	929	No	N/A	5'	TCGA-EK-A2R8-01
	HPV:E6	RAD51B:Intron	chr14:68646030	230	No	N/A	5'	TCGA-EK-A2R8-01
	HPV:E1	RAD51B:Intron	chr14:68651472	931	No	N/A	5'	TCGA-EK-A2R8-01
	HPV:E1	RAD51B:Intron	chr14:68671846	2669	No	N/A	5'	TCGA-EK-A2R8-01
	HPV:E1	RAD51B:Exon	chr14:69149653	840	Yes	Out-of-frame	5'	TCGA-EK-A2H0-01
	HPV:E7	RAD51B:Exon	chr14:69149653	131	Yes	Out-of-frame	5'	TCGA-EK-A2H0-01
	HPV:E1	RAD51B:Intron	chr14:68701639	881	Yes	N/A	5'	TCGA-BA-4077-01
	HPV:E1	RAD51B:Intron	chr14:68703892	881	Yes	N/A	5'	TCGA-BA-4077-01
	HPV:E1	RAD51B:Exon	chr14:68758600	883	Yes	Out-of-frame	3'	TCGA-BA-4077-01
	HPV:E6	RAD51B:Exon	chr14:68758600	228	Yes	In-frame	3'	TCGA-BA-4077-01
	HPV:E1	RAD51B:Intron	chr14:68801478	880	No	N/A	5'	TCGA-C5-A1BE-01
ERBB2	HPV:E7	ERBB2:Exon	chr17:37873744	576	No	N/A	5'	TCGA-C5-A1M9-01
	HPV:E6	ERBB2:Exon, 5' UTR	chr17:37855840	408	Yes	N/A	5'	TCGA-DS-A0VM-01

Supplementary Table S5. Detailed mapping of fusion breakpoints in recurrent genes. Integration breakpoints in recurrent genes were further fine-mapped by identification of breakpoint-spanning reads (partly human and partly viral). Where applicable, we determined whether fusions were in-frame or out-of-frame. 2 and 1 spanning reads, respectively, supported sites in *ERBB2*, while remaining listed sites were all supported by at least 10 reads. In cases where the exact breakpoint could not be determined due to identity between human and viral sequences (at most a 4 nt span), the position closest to human is presented.