

ARTICLE

<https://doi.org/10.1038/s42003-019-0352-3>

OPEN

# Dynamic host immune response in virus-associated cancers

Song Cao<sup>1,2</sup>, Kristine M. Wylie<sup>2,3</sup>, Matt A. Wyczalkowski<sup>1,2</sup>, Alla Karpova<sup>1</sup>, Jessica Ley<sup>4</sup>, Sam Sun<sup>1,2</sup>, R. Jay Mashl<sup>1,2</sup>, Wen-Wei Liang<sup>1,2</sup>, Xiaowei Wang<sup>5,6</sup>, Kimberly Johnson<sup>7</sup>, John F. DiPersio<sup>1,4</sup>, Hiram Gay<sup>4</sup>, Lee Ratner<sup>1</sup>, Feng Chen<sup>1,4</sup>, Douglas R. Adkins<sup>1,4</sup> & Li Ding<sup>1,2,4,8</sup>

Viruses drive carcinogenesis in human cancers through diverse mechanisms that have not been fully elucidated but include promoting immune escape. Here we investigated associations between virus-positivity and immune pathway alteration for 2009 tumors across six virus-related cancer types. Analysis revealed that for 3 of 72 human papillomavirus (HPV)-positive head and neck squamous cell carcinoma (HNSC) the HPV genome integrated in immune checkpoint genes *PD-L1* or *PD-L2*, driving elevated expression in the corresponding gene. In addition to the previously described upregulation of the PD-1 immunosuppressive pathway in Epstein-Barr virus (EBV)-positive stomach tumors, we also observed upregulation of the PD-1 pathway in cytomegalovirus (CMV)-positive tumors. Furthermore, we found signatures of T-cell and B-cell response in HPV-positive HNSC and EBV-positive stomach tumors and HPV-positive HNSC patients were associated with better survival when T-cell signals were detected. Our work reveals that viral infection may recruit immune effector cells, and upregulate PD-1 and CTLA-4 immunosuppressive pathways.

<sup>1</sup> Department of Medicine, Washington University in St. Louis, St. Louis, MO 63110, USA. <sup>2</sup> McDonnell Genome Institute, Washington University in St. Louis, St. Louis, MO 63108, USA. <sup>3</sup> Department of Pediatrics, Washington University in St. Louis, St. Louis, MO 63110, USA. <sup>4</sup> Siteman Cancer Center, Washington University in St. Louis, St. Louis, MO 63110, USA. <sup>5</sup> Department of Radiation Oncology, Washington University in St. Louis, St. Louis, MO 63110, USA.

<sup>6</sup> Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO 63110, USA. <sup>7</sup> Brown School Master of Public Health Program, Washington University in St. Louis, St. Louis, MO 63130, USA. <sup>8</sup> Department of Genetics, Washington University in St. Louis, St. Louis, MO 63110, USA. Correspondence and requests for materials should be addressed to L.D. (email: [lding@wustl.edu](mailto:lding@wustl.edu))

Since the sequencing of the first genomic DNA from a leukemia patient<sup>1</sup>, various studies have identified somatic and germline variants in key cancer genes<sup>2–5</sup>. These genomic biomarkers may aid in therapy selection<sup>6</sup>. Although large-scale sequencing projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) continue to catalog variants across cancer types, only a minority of patients harbor tumors with genomic aberrations associated with sensitivity to targeted therapy.

Complementary to targeted therapy, cancer immunotherapy utilizes the host immune response to kill tumor cells<sup>7,8</sup>. PD-L1 and PD-L2 on tumor cells or antigen-presenting cells suppress T-cell immune response by binding to PD-1 on T-cells<sup>9,10</sup>. To escape attack by immune cells, tumor cells overexpress PD-L1 by gene amplification, utilization of an ectopic promoter, and disruption of 3' untranslated regions (3' UTRs)<sup>11</sup>, in addition to PTEN loss-of function<sup>12</sup> and EGFR mutations<sup>13</sup>. Other studies indicate that EGFR mutations are not associated with an increased PD-L1 expression and a better clinical response of PD-L1 immune checkpoint inhibitors<sup>14,15</sup>. Elevated PD-L1 expression creates an immunosuppressive microenvironment that facilitates tumor progression<sup>16</sup>. Anti-PD-1 and anti-PD-L1 immune checkpoint blockades show favorable clinical outcome for treating patients with high PD-1 and PD-L1 expression<sup>17–20</sup>. Another important immune checkpoint pathway involves CTLA-4 and its ligands CD80 and CD86. CTLA-4 serves as a negative regulator of T-Cell activity. The anti-CTLA-4 blockade is also an effective therapeutic strategy to kill tumor cells<sup>21,22</sup>.

Immune infiltration of the tumor microenvironment correlates with improved survival in cancer patients<sup>23,24</sup>. Despite the importance of immune infiltrates and their theoretical associations with viral-positivity<sup>25</sup>, there is no systematic study of associations between virus-positive samples and the immune response except for some limited studies on human papillomavirus (HPV) in head and neck squamous cell carcinoma (HNSC)<sup>26–29</sup>. Our previous large-scale study demonstrated viral positivity in multiple cancer types<sup>30</sup>. Here, using TCGA RNA-Seq data for six virus-associated tumors, we systematically study the associations between virus-positivity and the tumor microenvironment, as measured by expression of *PD-L1*, *PD-L2*, *PD-1*, *CD80*, *CD86*, *CTLA-4*, *Tim-3*, *LAG3*, and *4-1BB* and the prevalence of infiltrating immune cells across multiple types of human cancers. Specifically, we found the enrichment of a T-cell immune signature in HPV-positive and EBV-positive tumors compared to non-viral tumors. The increase of T-cell immune response is associated with a better prognosis in HPV-positive patients. In addition, we found HPV integrations at *PD-L1* and *PD-L2* are associated with high expression of these genes. Higher levels of *PD-L1*, *PD-L2*, *PD-1*, *CD80*, *CD86*, *CTLA-4*, *Tim-3*, *LAG3*, and *4-1BB* expression were found in Epstein-Barr virus (EBV)-positive stomach adenocarcinoma (STAD) and cytomegalovirus (CMV)-positive colon and rectum adenocarcinoma (COADREAD) tumors compared to virus-negative tumors, providing the rationale for treating virus-positive tumors by anti-PD-1, and anti-CTLA-4 immune therapy. Besides *PD-L1*, *PD-L2*, *PD-1*, *CD80*, *CD86*, *CTLA-4*, *Tim-3*, *LAG3*, and *4-1BB*, we found the increase of an inducible co-stimulator (ICOS) expression in both HPV-positive head and neck squamous cell carcinoma and EBV-positive stomach adenocarcinoma tumors. ICOS is the immune checkpoint protein, functionally and structurally related to CD28<sup>31</sup>. A positive ICOS signature may indicate a better clinical outcome of anti-CTLA-4 immune therapy in HPV-positive head and neck squamous cell carcinoma and EBV-positive stomach adenocarcinoma patients<sup>32</sup>.

## Results

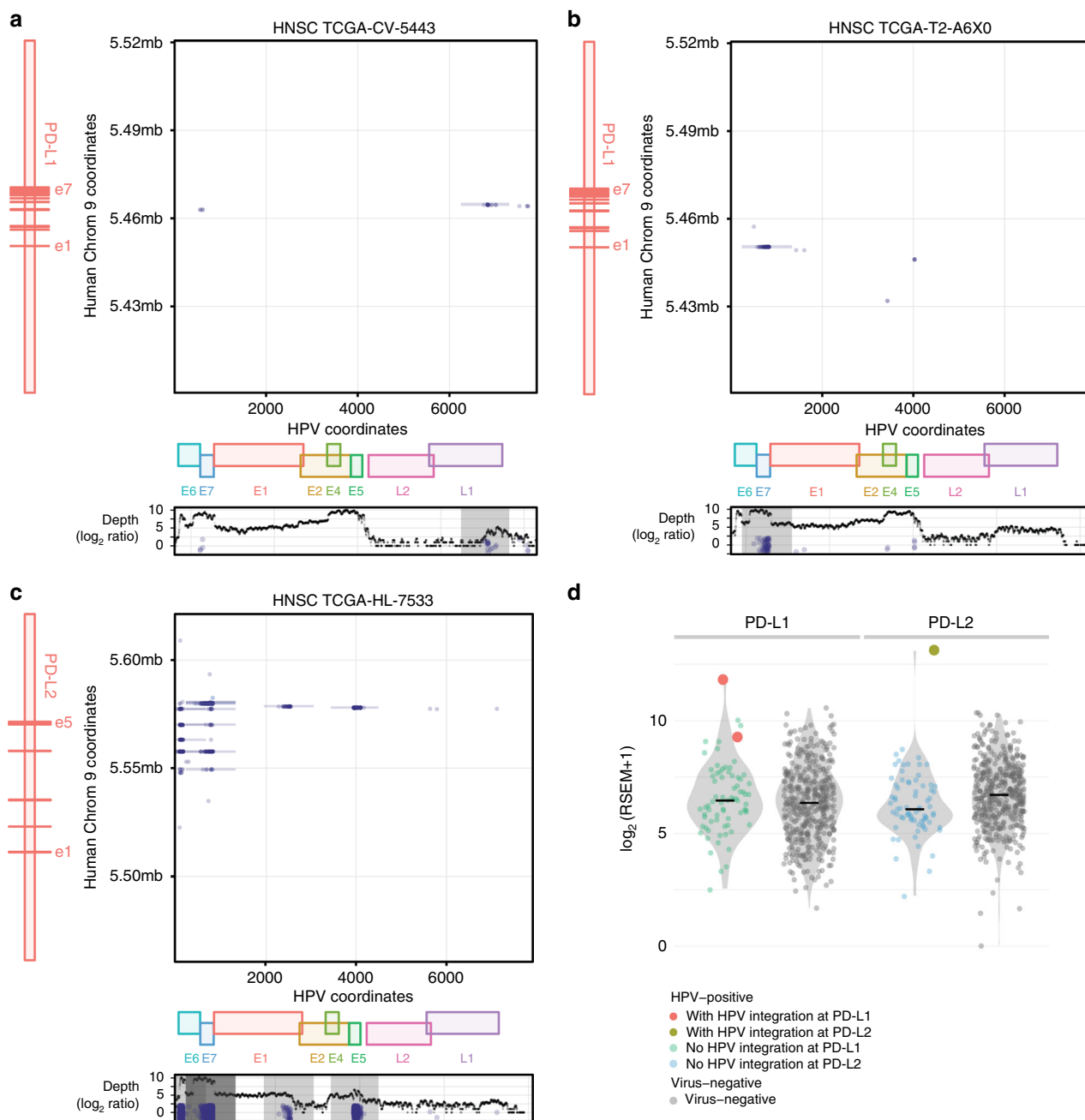
**Recurrent HPV Integrations at *PD-L1* or *PD-L2* in HNSC.** We analyzed 498 TCGA HNSC tumors using VirusScan and identified 72 HPV-positive tumors (numbers of virus-supporting reads per hundred million reads mapped (RPKM) > 100) and 341 virus-negative tumors (RPKM < 5)<sup>30</sup>. The 413 HNSC tumors with clear HPV status were most common in two ethnicity groups: 364 Caucasians and 30 African Americans. There was no significant difference in HPV status between the two ethnicity groups. We additionally found that HPV-positive HNSC tumors were mostly from males (92%). See Supplementary Table 1. Of these, we identified three tumors with HPV integrations at *PD-L1* or *PD-L2* by using discordant read pair analysis (Methods, Fig. 1a–c). In tumor TCGA-CV-5443, HPV integration sites were localized to intron 4 of *PD-L1*. The same integration site in the same sample was also reported in a previous study<sup>33</sup>. Given a larger cohort size, we also found additional previously unidentified HPV integration sites at *PD-L1* and *PD-L2* in tumors TCGA-T2-A6X0 and TCGA-HL-7533, respectively. Inspection of the detailed discordant read pairs showed that the viral E7 gene integrates into the 5' UTR region of *PD-L1*; see Fig. 1b. HPV integrations at *PD-L2* appeared more complicated than at *PD-L1*, revealing multiple HPV integration sites in or after intron 3 of *PD-L2*. These three tumors originated in different anatomic sites (larynx, tonsil, and oral cavity). Although different integration patterns and anatomic sites were observed in the three tumors, HPV integrations at *PD-L1* or *PD-L2* were all accompanied by higher expression levels of these genes compared with those in virus-negative tumors (Fig. 1d), and *PD-L1* or *PD-L2* with HPV integrations are expression outliers (see Methods).

To examine the prevalence of HPV integrations at *PD-L1* or *PD-L2* in other tumor types, we analyzed 229 TCGA HPV-positive cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) tumors and found no putative HPV integrations at either gene. We also checked for integrations of viruses, including HPV, hepatitis B virus (HBV), EBV, and cytomegalovirus, in other cancer types and found no integration events at *PD-L1* or *PD-L2*. This result indicates that high *PD-L1* or *PD-L2* expression induced by HPV integration is a phenomenon that is selectively associated with HPV in HNSC tumors, with *PD-L1* or *PD-L2* integrations occurring in 4.2% of HNSC HPV-positive tumors.

We further looked the relationship between HPV integration and expression in T-cell and B-cell genes. Supplementary Fig. 1 shows the distribution of expression for those genes with HPV integrations. Interestingly, we found that *NR4A2*, *TBC1D1*, *BTNL9*, *DTX1*, *FOXP1*, *INPP4B*, *PDE4D*, and *STAT4* with HPV integration events are expression outliers (see Methods) and all these events are associated with the increase of expression.

## Effect of viral infection on levels of immune checkpoint genes.

Having correlated gene-specific viral integration with elevated *PD-L1* and *PD-L2* expression in HPV-positive HNSC, we next correlated any viral infection with *PD-L1*, *PD-L2*, *PD-1*, *CD80*, *CD86*, *CTLA-4*, *Tim-3*, *LAG3*, and *4-1BB* expressions. In Fig. 2, we compare expression levels of these genes in four tumor types positive for different viruses: HPV in HNSC, EBV in STAD, HBV in liver hepatocellular carcinoma (LIHC), and cytomegalovirus in colon and rectum adenocarcinoma (COADREAD). For the three viruses, we only observed an association between the ethnicity groups for HBV status: 98% are from ASIAN group. We observed most of EBV and HBV tumor patients are Male; See Supplementary Table 1. In HNSC, we found that three tumors with HPV integrations had high expression of *PD-L1* or *PD-L2*, i.e., TCGA-CV-5443 and TCGA-T2-A6X0 with 11.8 and 9.2 for *PD-*

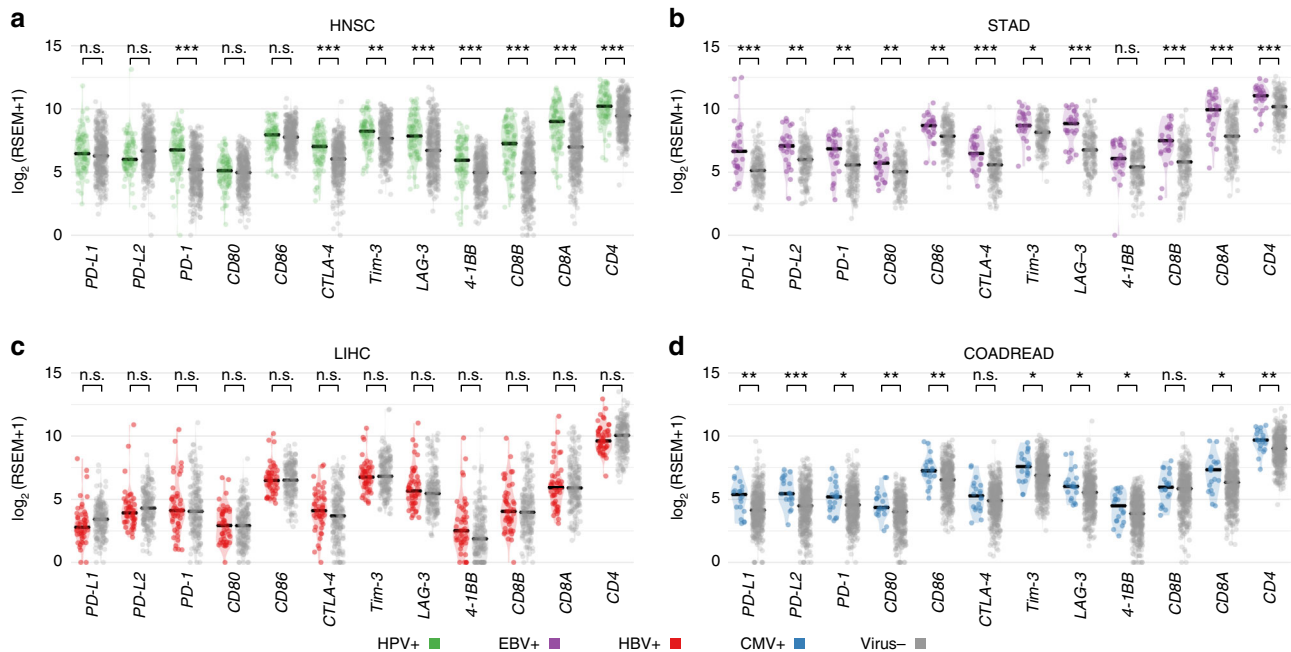


**Fig. 1** HPV integrations at *PD-L1* and *PD-L2*. **a–c** Structural analysis of HPV integration events at *PD-L1* or *PD-L2* for three different HNSC tumors. The dots in the plot show the coordinates in HPV genome (X) and human chromosome (Y) of the breakpoints<sup>67</sup>. The vertical line represents a series of breakpoints clustered together by Pindel<sup>66</sup>. In the figure, e1 represents exon 1. HPV integrates intron 4 and 5' UTR of *PD-L1* in tumor TCGA-CV-5443 and TCGA-T2-A6X0, respectively. In tumor TCGA-HL-7533, HPV can integrate multiple locations in or after intron 3 of *PD-L2*. The bottom panel shows the number of mapped HPV reads ( $\log_2$  ratio) along with HPV coordinates and E6, E7, E1, E2, E4, E5, L2, and L1 are different genes in HPV genome. **d** The expression of *PD-L1* or *PD-L2* for HPV-positive samples with and without HPV integrations and virus-negative tumors

*L1* and TCGA-HL-7533 with 10 for *PD-L2* (RSEM in  $\log_2$  scale). RSEM stands for RNA-Seq reads by expectation maximization, which is widely used for quantifying gene expression<sup>34</sup>. Overall, no significant difference in *PD-L1*, *PD-L2*, *CD80*, or *CD86* expression between HPV-positive and virus-negative HNSC tumors was found. A similar observation was also made for HBV (Fig. 2c). However, we found a higher level of *PD-L1*, *PD-L2*, *CD80*, *CD86*, *Tim-3*, *LAG3*, and *4-1BB* in EBV-positive STAD and cytomegalovirus-positive colon and rectum adenocarcinoma than in virus-negative tumor samples (Fig. 2b–d). To leverage the

new findings of elevated immune escape pathways of cytomegalovirus-positive colon and rectum adenocarcinoma to other cytomegalovirus-positive tumors, we compared cytomegalovirus-positive and negative tumor samples from stomach and esophageal carcinoma (STES). Supplementary Fig. 2a shows a higher *PD-L1*, *PD-L2*, or *CD80* expression in cytomegalovirus-positive stomach and esophageal carcinoma.

In our previous study, we found a high prevalence of EBV-positive and cytomegalovirus-positive esophageal cancers<sup>30</sup>. In the current study, we found upregulation of *PD-L2* expression



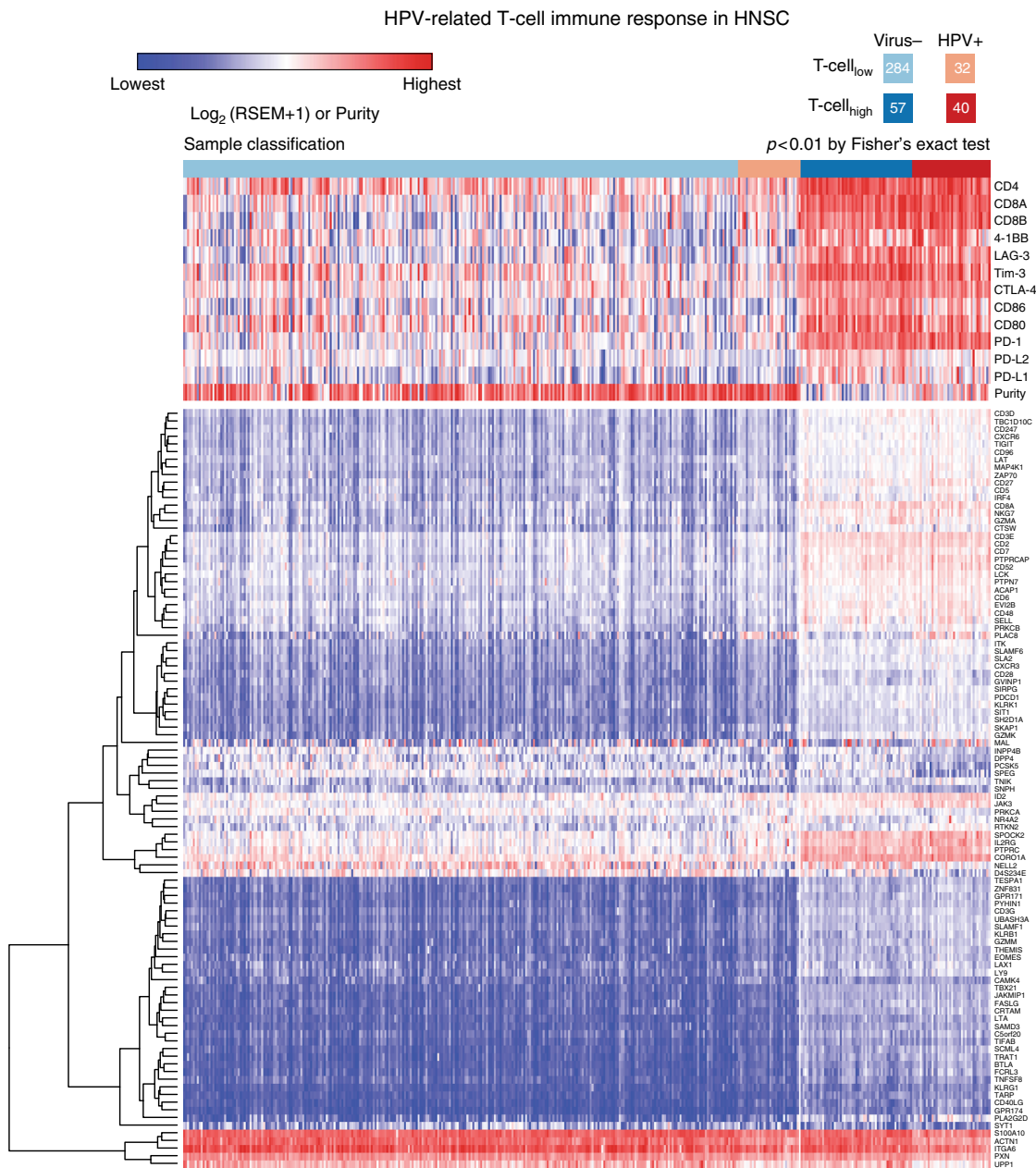
**Fig. 2** Expression of immune checkpoint genes in virus-positive and negative tumors. The comparison of the expressions of immune checkpoint genes *PD-L1*, *PD-L2*, *PD-1*, *CD80*, *CD86*, *CTLA-4*, *Tim-3*, *LAG-3*, and *4-1BB* as well as T-cell *CD4* and *CD8* markers for **a** HPV-positive and virus-negative tumors in HNSC, **b** EBV-positive and virus-negative tumors in STAD, **c** HBV-positive and virus-negative tumors in liver hepatocellular carcinoma and **d** Cytomegalovirus-positive and virus-negative tumors in colon and rectum adenocarcinoma. The “\*”, “\*\*\*”, and “\*\*\*\*” symbols indicate *p*-value in the regions [0.05, 0.01], (0.01,0.001), and <0.001, respectively

( $p = 0.02$ ), *CD80* ( $p = 0.01$ ), and *CD86* (0.04) associated with EBV or cytomegalovirus positivity (Supplementary Fig. 2b). In this analysis, we combined the EBV and cytomegalovirus-positive tumor samples together to improve the statistical power. In esophageal cancers, we obtained 10 EBV or cytomegalovirus-positive and 89 virus-negative tumors for the statistical analysis. We also noted that *PD-1*, *CTLA-4*, *CD4*, and *CD8* expressions were also higher in the EBV and cytomegalovirus-positive samples, suggesting a higher level of infiltrating T-cells compared with virus-negative tumors (Fig. 2). Although an elevated *PD-L1* expression in EBV-positive samples has been reported in other studies<sup>35,36</sup>, our study shows that EBV or cytomegalovirus infection increases expression of genes encoding *PD-L1*, *PD-L2*, *PD-1*, *CD80*, *CD86*, *CTLA-4*, *Tim-3*, *LAG3*, and *4-1BB* immune checkpoint genes together with other T-cell markers such as *CD4* and *CD8* in tumors along the gastrointestinal tract, including esophagus, stomach, and intestine.

**Effect of viral infection on host immune response.** *CD4+* and *CD8+* T-cells and B-cells play important roles in fighting infection and cancer. Immune infiltration is frequently observed in solid tumors, and is associated with improved host survival<sup>23</sup>. Here, we evaluated associations between viral infection and immune cell infiltration in the tumor microenvironment. We collected a list of genes corresponding to T-cell and B-cell signatures (Supplementary Table 2) from previous publications<sup>37–39</sup>. We then identified 99 genes with significant differential expression (FDR < 0.05, see Methods) between HPV-positive and virus-negative HNSC tumors (Supplementary Data 1, Fig. 3). In the 99 selected genes, we also required that the difference of median values of gene expression ( $\log_2$ ) is larger than 1 between the two cohorts. Overall, HPV-positive tumors displayed higher levels of T-cell signatures than virus-negative tumors (Fig. 3). In Fig. 3, we separated samples into four different groups based on supervised

clustering results, i.e., Virus-/T-cell<sub>low</sub>, HPV+/T-cell<sub>low</sub>, Virus-/T-cell<sub>high</sub>, and HPV+/T-cell<sub>high</sub>. Overall, HPV-positive tumors displayed higher levels of T-cell signatures than virus-negative tumors (Fig. 3). The expressions of 99 T-cell genes in HPV-positive HNSC tumor are higher than the values in virus-negative samples (Supplementary Fig. 3). GSEA<sup>40,41</sup> also shows the enrichment of T-cell gene set in HPV-positive HNSC tumors (Supplementary Fig. 4a). By using gene ontology (GO) database<sup>42,43</sup>, we found that most of the 99 genes are classified as gene sets related to immune response, lymphocyte and leukocyte, indicating infiltrated immune cells, and etc. No obviously different clusters were observed in term of GO annotation for these genes (Supplementary Fig. 5). We found that tumors with high T-cell signatures from our clustering method were associated with high *PD-L1*, *PD-L2*, *PD-1*, *CD80*, *CD86*, *CTLA-4*, *Tim-3*, *LAG3*, *4-1BB*, *CD8*, and *CD4* expression. These tumors also associated with high immune scores and lower tumor purity, indicative of a high level of immune infiltration (Supplementary Fig. 6a, c). Tumor purity and immune score were calculated based on the method used by Aran et al.<sup>44</sup>. In addition, HPV-positive HNSC tumors had elevated levels of *PD-1*, *CTLA-4*, *CD8*, and *CD4* compared to virus-negative tumors (Fig. 2a). Similarly, we found an enrichment of B cell signatures in HPV-positive HNSC tumors (Supplementary Figs. 4b, 7).

Next, we identified 78 T-cell genes with significant differential expression (FDR < 0.05, see Methods) between EBV-positive and virus-negative STAD tumors (Fig. 4 and Supplementary Data 1). In the 78 selected genes, we also required that the difference of median values of gene expression ( $\log_2$ ) is larger than 1 between the two cohorts. GSEA analysis indicates an enrichment of T-cell in EBV-positive STAD tumors (Supplementary Fig. 4c). GO database annotation shows that 77 differential expression are mostly related to immune response, lymphocyte, cell activation, etc. and we did not observe significant different clusters among these genes in term

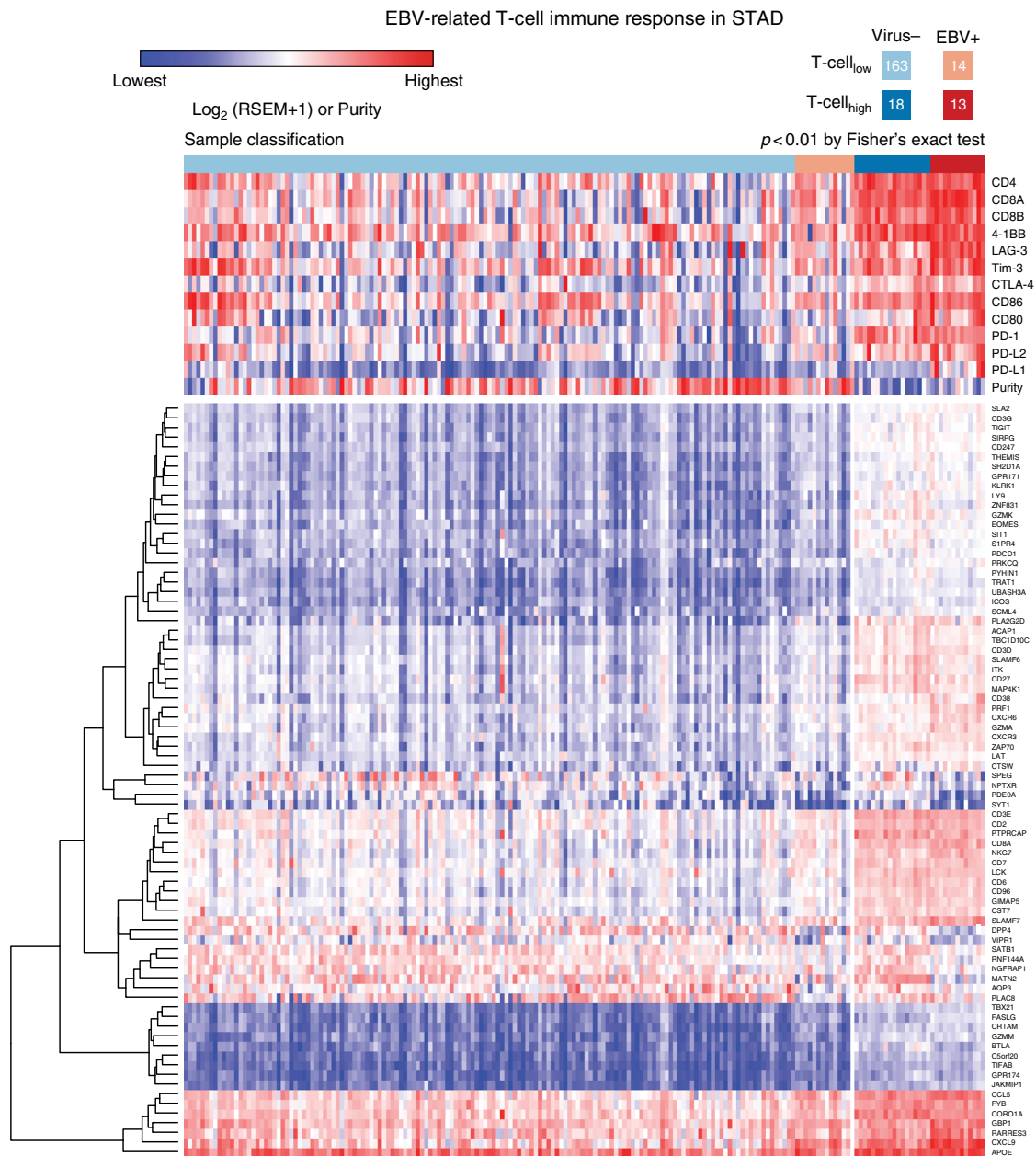


**Fig. 3** The supervised clustering of T-cell genes in HPV-positive and virus-negative HNSC tumors. In the upper panel, we show the expression of T-cell markers such CD4, CD8 and immune checkpoint genes *PD-L1*, *PD-L2*, *PD-1*, *CD80*, *CD86*, *CTLA-4*, *Tim-3*, *LAG-3*, and *4-1BB*. Tumors are clustered to four different groups based on the T-cell and virus statuses, i.e., Virus-/T-cell<sub>low</sub>, HPV+/T-cell<sub>low</sub>, Virus-/T-cell<sub>high</sub>, and HPV+/T-cell<sub>high</sub> from left to right

of GO annotation (Supplementary Fig. 8). Tumors with high T-cell signatures from our clustering analysis are concordant with high expression of *PD-L1*, *PD-L2*, *PD-1*, *CD80*, *CD86*, *CTLA-4*, *Tim-3*, *LAG3*, *4-1BB*, *CD8*, and *CD4*, EBV-positive status (Fig. 5), and low tumor purity and high immune score (Supplementary Fig. 6b, d). B-cell response signatures were also observed in EBV-positive STAD tumors based on differential expression of 34 genes compared with non-viral tumors (Supplementary Fig. 9, 4d). This includes human leukocyte antigen (*HLA*) genes such as *HLA-DQB1*, *HLA-DQA1*, *HLA-DMA*, *HLA-DMB*, *HLA-DRB5*, and *HLA-DOA*. These genes play an important role for the formation of major histocompatibility complex (MHC) class II/peptide complex, which recognizes microbial antigens and cancer neoantigens<sup>45</sup>.

From our analyses, 78% of T-cell signature genes in EBV-positive tumors overlapped with those from HPV-positive HNSC, indicating that similar T-cell immune responses are associated with EBV and HPV. For instance, an ICOS signature was identified in EBV-positive STAD (Supplementary Table 2) as well as HPV-positive HNSC; The expression of *ICOS* increases about two-fold in EBV-positive STAD and HPV-positive HNSC compared to virus-negative samples. *ICOS* is an immune checkpoint gene in *CD28* and *CTLA-4* family, which plays important role in regulating the immune response and enhances the antitumor immune response in anti-*CTLA-4* blockade<sup>31,32</sup>.

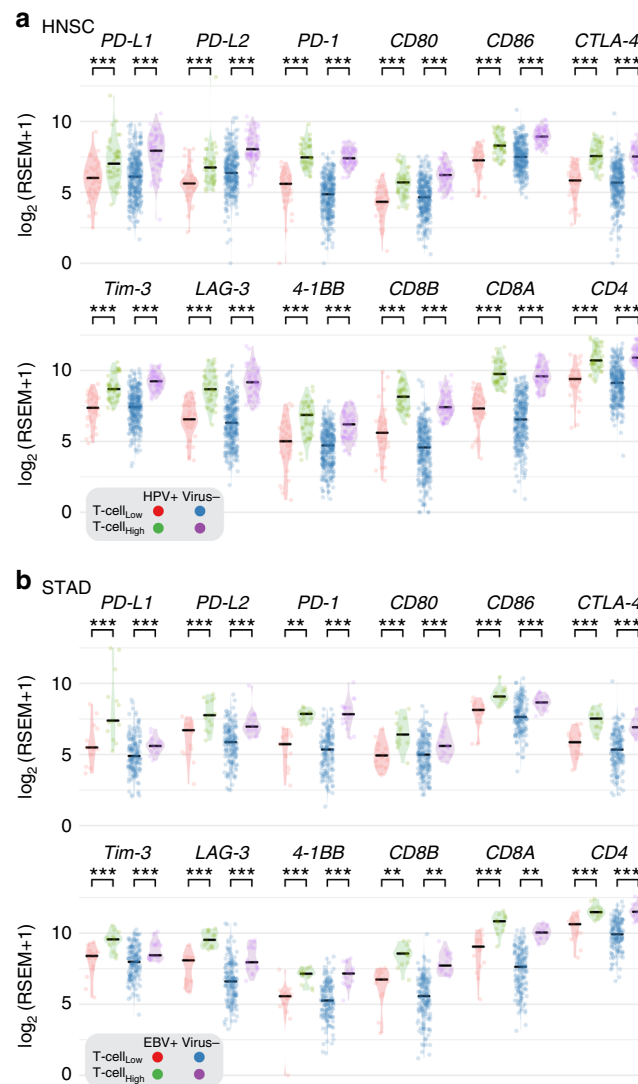
Furthermore, we selected genes, which showed at least four-fold difference in the relative expression between HPV-positive and EBV-positive samples compared to the corresponding virus-negative cohorts. In the T-cell gene list, six genes (*PLAC8*,



**Fig. 4** The supervised clustering of T-cell genes in EBV-positive and virus-negative STAD tumors. In the upper panel, we show the expression of T-cell markers such CD4, CD8 and immune checkpoint genes *PD-L1*, *PD-L2*, *PD-1*, *CD80*, *CD86*, *CTLA-4*, *Tim-3*, *LAG-3*, and *4-1BB*. Tumors are clustered to four different groups based on the T-cell and virus statuses, i.e., Virus-/T-cell<sub>low</sub>, EBV+/T-cell<sub>low</sub>, Virus-/T-cell<sub>high</sub>, and EBV+/T-cell<sub>high</sub> from left to right

*CXCL9*, *SATB1*, *PDE9A*, *NPTXR*, and *NELL2*) passed the cut-off. We note that for *PLAC8*, a gene that is associated with pancreatic cancer progression<sup>46</sup>, was highly elevated in HPV-positive samples with about 16-fold increase compared to the virus-negative HNSC samples. However, in EBV-positive STAD sample, the expression of *PLAC8* was downregulated by about four-fold compared to virus-negative STAD sample (Supplementary Table 2). In the B-cell gene list, we identified an additional 18 genes fitting these criteria, such as *STAG3*, *MET1*, *CDKN2A*, *COCH*, *BTNL9*, *SPIB*, *MS4A1*, *CD19*, *CR2*, *BLK*, *VPREB3*, *CXCR5*, *MIR600HG*, *BANK1*, *F5*, *BACH2*, *KYNU*, and *SLC22A3*. *CDKN2A* showed distinct expression alteration in HPV-positive and EBV-positive samples. In HPV-positive samples, the expression of *CDKN2A* was 16-fold higher than virus-negative HNSC samples, but in EBV-positive samples, its expression was

three-fold lower than the virus-negative samples. *CDKN2A* is a tumor suppressor, which is highly mutated in virus-negative HNSC samples<sup>47</sup>, but not in HPV-positive HNSC samples. The low expression of *CDKN2A* found in EBV-positive STAD and HPV-negative HNSC samples reflect two different mechanisms inactivating the gene function by EBV infection and somatic mutation. We note that *CDKN2A* is mostly involved in the regulation of cell cycle<sup>48</sup>, not necessarily related to the immune infiltration. The latter is mainly associated with a low tumor purity and increased overall expression of T-cell genes (Figs. 3, 4). In contrast to HPV-positive HNSC and EBV-positive STAD, we did not find significant enrichment of T-cell and B-cell signatures in HBV-positive liver hepatocellular carcinoma. For colon and rectum adenocarcinoma, we find higher levels of *CD4*, *CD8A*, and *PD-1* in cytomegalovirus-positive tumors than in virus-negative

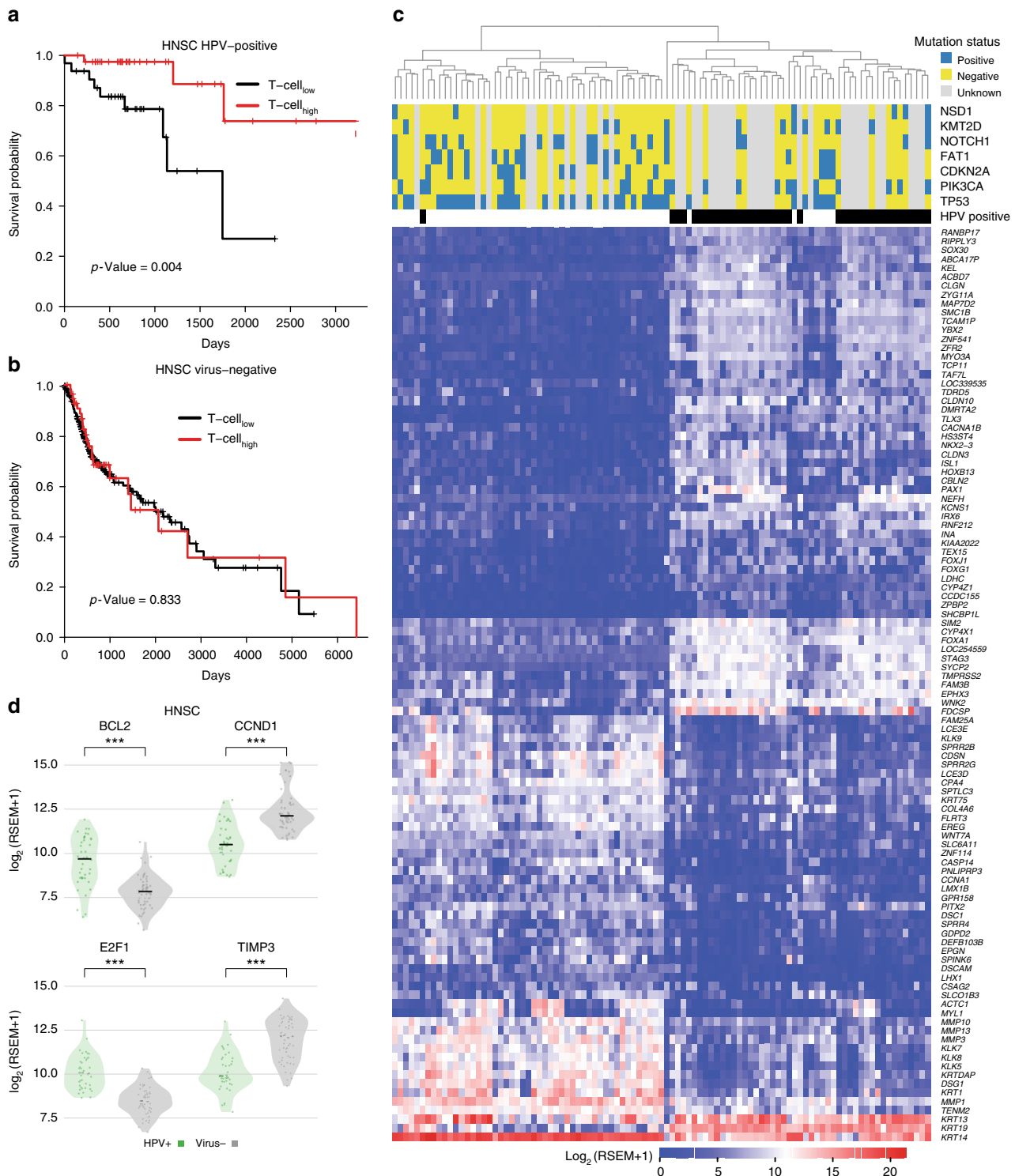


**Fig. 5** Expression of immune checkpoint genes in T-cell<sub>high</sub> and T-cell<sub>low</sub> tumors. The comparison of the expressions of immune checkpoint genes *PD-L1*, *PD-L2*, *PD-1*, *CD80*, *CD86*, *CTLA-4*, *Tim-3*, *LAG-3*, and *4-1BB* as well as T-cell CD4 and CD8 markers between T-cell<sub>high</sub> and T-cell<sub>low</sub> tumors in HPV-positive HNSC, EBV-positive STAD and virus-negative cohorts

tumors (Fig. 2), though other T-cell genes only showed modest differential expression in cytomegalovirus-positive and virus-negative tumors.

We further compared the T-cell differentiation phenotype in virus-positive and negative tumors for different viruses by using markers such as CD28, CD27, CD45, CD103, perforin, GMP-17, and granzymeA<sup>49</sup> (see Supplementary Fig. 10). High expression of CD28 and CD27 in HPV-positive HNSC and cytomegalovirus-positive colon and rectum adenocarcinoma tumors suggest an increase in the presence of CD28+CD27+ T-cells compared to virus-negative HNSC tumors and virus-negative colon and rectum adenocarcinoma tumors. CD28 and CD27 are markers for precursor or early differentiation T-cells<sup>49</sup>. The HPV-positive HNSC tumors and cytomegalovirus-positive colon and rectum adenocarcinoma tumors also had higher expression of NK T-cell markers (perforin, GMP-17, and granzymeA) and CD45 compared with virus-negative HNSC and colon and rectum adenocarcinoma tumors. Also, HPV-positive HNSC tumors and EBV-positive tumors had a higher expression of CD103, a marker for resident T-cells, compared to virus-negative samples, but not in cytomegalovirus-positive colon and rectum adenocarcinoma and HBV-positive liver hepatocellular carcinoma tumors.

**Clinical relevance of virus-associated immune response.** Based on our clustering analysis, HPV-positive tumors were more likely to have an elevated immune response (40 of 72 tumors, 55%) compared with virus-negative tumors (57 of 341 tumors, 16%) ( $p < 0.01$ , Fisher's exact test, two-tailed) (Fig. 3). We first performed survival analysis to evaluate how immune response correlates with clinical outcome in HPV-positive and virus-negative cohorts (Fig. 6a, b, respectively). We found that immune response is associated with a positive prognosis in patients with HPV-positive HNSC, but not in those with virus-negative HNSC (Fig. 6a). We then examined mutational and gene expression patterns in two cohorts: HPV-positive HNSC with elevated immune response and virus-negative HNSC with elevated immune response (Fig. 6c); in total, we found 2695 genes with at least two-fold differential expression (FDR < 0.05). We identified 103 genes with 16-fold or higher differential expression (Fig. 6c). Notably, we found that the expression of *FOXA1*, a gene associated with better survival in breast cancer<sup>50</sup>, is about 16-fold higher in HPV-positive tumors. We also identified seven highly mutated genes (*TP53*, *PIK3CA*, *CDKN2A*, *FAT1*, *NOTCH1*, *KMT2D*, and *NSD1*) with frequency greater than 10% in HNSC tumors with elevated immune response. The HPV-positive cohort



**Fig. 6** Survival and gene expression analyses in HPV-positive and negative tumors. The comparison of overall survival in T-cell<sub>high</sub> and T-cell<sub>low</sub> tumors between two different cohorts: **a** HPV-positive and **b** virus-negative. **c** The mutational status for seven mutated genes in HNSC and the supervised clustering based on 103 genes which demonstrate strong significantly differential expression among HPV-positive and virus-negative HNSC tumors (FDR < 0.05, see Methods). **d** The differentially expressed genes involved in p53 signaling pathway between HPV-positive and virus-negative tumors

contained no variants in either *TP53* or *CDKN2A* and only one in *FAT1*; in the HPV-negative cohort, these genes were frequently mutated. HNSC tumors with wild-type *TP53* are more sensitive to radiation therapy than tumors with *TP53* mutations<sup>51</sup>. In

addition, we examined differentially expressed genes in the p53 signaling pathway between the two cohorts and found 4 of 16 p53 pathway genes showing substantial expression alteration (see Fig. 6d). Levels of B-cell CLL/lymphoma 2 (*BCL2*) and E2F



transcription factor 1 (E2F1) are higher in HPV-positive tumors. Previous studies show favorable prognosis with a high expression of BCL2<sup>52</sup> and poor prognosis with overexpressed CCND1<sup>53</sup>. The distinct expression pattern of p53 signaling pathway genes may also drive different clinical outcome of the two cohorts, though both cohorts are associated with an infiltrated immune cell microenvironment.

## Discussion

In this study, we systematically investigated associations between virus infection or integration and alteration of the tumor microenvironment. We found a significant difference ( $p = 0.01$ , Fisher's exact test) between HPV integration status at *PD-L1* or *PD-L2* in HPV-positive HNSC ( $N = 72$ ) and CESC ( $N = 229$ ) tumors. Specifically, we found three integrations among the HNSC samples with high expression of *PD-L1* or *PD-L2* and no integrations among the CESC samples. It is likely that HPV has co-evolved to target *PD-L1* or *PD-L2* to create an immunosuppressive tumor microenvironment in some head and neck cancers. Further investigation with a large sample set is important for leveraging the observed relationship between virus integration at *PD-L1* or *PD-L2* in HNSC and the increased expression. We also found samples with HPV integrations in other immune-related genes (*NR4A2*, *TBC1D1*, *BTNL9*, *DTX1*, *FOXP1*, *INPP4B*, *PDE4D*, and *STAT4*) have an increased expression of these genes. Previous studies show that high expression of *NR4A2*, *BTNL9*, *FOXP1*, or *PDE4D* can antagonize immune response or is associated with tumor progression<sup>54–57</sup>. In EBV-positive STAD, cytomegalovirus-positive colon and rectum adenocarcinoma and EBV or cytomegalovirus positive stomach and esophageal carcinoma, we found that viral infection associates with high expression of *PD-L1* or *PD-L2*, *CD80*, *CD86*, *PD-1*, *CTLA-4*, *Tim-3*, *LAG3*, and *4-1BB* without integrating into the human genome. Moreover, our study indicates that EBV and cytomegalovirus elevate *PD-L1* or *PD-L2*, *CD80*, *CD86*, *PD-1*, *CTLA-4*, *Tim-3*, *LAG3*, and *4-1BB* expression in multiple cancers along the gastrointestinal tract including stomach and esophageal carcinoma and colon and rectum adenocarcinoma.

A previous study<sup>58</sup> demonstrated elevated *PD-L1* expression in both tumor and immune cells across a large number of tumor samples by using immunohistochemistry (IHC) assays. For example, in 101 head-neck tumors, they found 28 and 19% of immune cell and tumor cell, respectively, were positive for *PD-L1* in their samples. The anti-*PD-L1* antibody works well on *PD-L1* positive tumors to neutralize *PD-L1* and make the tumor susceptible to attack by the immune system<sup>58</sup>. Our studies show an elevated *PD-L1* expression in EBV and cytomegalovirus-positive samples, suggesting clinical trials of *PD-L1* immunotherapy in these patients may be beneficial. Also, in a subset of HPV-positive HNSC, *PD-L1* is highly expressed when HPV integrates into the *PD-L1*, suggesting these patients may have responded to anti-*PD-L1* therapies. The previous clinical trial on HNSC has shown longer overall survival in both HPV-positive and *PD-L1*-positive tumors when treating with anti-*PD-1* monoclonal antibody than the standard single-agent therapy<sup>59</sup>. Furthermore, we found that the expression of *PD-L2*, *CD80*, *CD86*, and *CTLA-4* are also elevated in cytomegalovirus and EBV-positive tumor patients, suggesting anti-*PD-L2* and anti-*CTLA4* immunotherapy may be effective in patients with these types of tumors.

In contrast to the clear association between EBV and stomach adenocarcinoma, some controversy exists about the association between cytomegalovirus and colorectal cancer<sup>60–62</sup>. In our previous large-scale study, we found a higher abundance of cytomegalovirus in tumors than in adjacent normal samples<sup>30</sup>. In the current study, we discovered a high level of *PD-L1/PD-L2* in

cytomegalovirus-positive tumors across the gastrointestinal tract suggesting that cytomegalovirus mediates the tumor microenvironment, which helps tumor cells to avoid the attack of immune cells.

In addition, we found distinct immune responses for different viruses in different cancer types. A high level of immune response was observed in HPV-positive HNSC and EBV-positive STAD samples but not in HBV-positive liver hepatocellular carcinoma. One explanation is that HBV promotes cancer in a different way than EBV/HPV, which are directly oncogenic, HBV promotes cancer by making the liver cirrhotic/inflamed chronically. The immune response was measured by gene expression of characteristic T-cell and B-cell markers, including *CD4*, *CD8*, and *PD-1* T-cell markers, and tumor purity, which when low indicates high immune cell infiltration into the tumor microenvironment. We also observed high expression of *ICOS* and *CTLA-4* in both HPV-positive HNSC and EBV-positive STAD, suggesting these tumors may have had an effective clinical response to anti-*CTLA-4* immune therapy. Survival analysis shows high immune response is associated with favorable survival in HPV-positive but not HPV-negative HNSC samples. The different mutational status and expression patterns may lead to different clinical outcomes of the immune response. For instance, we found different expression alteration in key genes involved p53 signaling pathway in two cohorts. The complete retention of wild-type TP53 in HPV-positive HNSC tumors is another key factor driving the difference, as previous studies show better radiotherapy sensitivity in HNSC patients with wild-type TP53<sup>51</sup>. For cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), we separated tumors into low and elevated immune infiltration cohorts according to HPV-positive T-cell signatures (Supplementary Fig. 11). Although tumors with elevated immune response show a better survival rate before eight years, there is no significant difference in overall survival rate based on immune response (Supplementary Fig. 12). Further clustering samples based on a *CD8+* T-cell gene list show improved association survival rate and *CD8+* T-cell status and patients with *CD8+* T-cell status have a higher chance of a tumor-free status (Supplementary Figs. 13, 14).

Our study highlights the importance of viral integration and infection in shaping tumor microenvironments. The current study is necessarily based on gene expression data from RNA-Seq. Proteomics data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) will enable us to investigate the virus-mediated tumor microenvironment at the protein level<sup>63</sup>. The highly immunogenic property of HPV16 virus, the dominant HPV subtype affecting HNSC patients, can help to explain the increased immune response in HPV-positive HNSC samples (Supplementary Fig. 15a). There is no significant difference in terms of mutational burden in HPV-positive and virus-negative samples (Supplementary Fig. 15b). However, why different patients have different responses to different viral presentation requires more detailed work in the future. In addition, though the current work does not identify clear mechanisms by which virus infection affects *PD-L1* and *PD-L2* expression, it nonetheless suggests that viruses may aid tumors in evading the *PD-1* immune checkpoint pathway across multiple cancer types. Our analysis of elevated expression in both *PD-1*, *CTLA-4*, *Tim-3*, *LAG3*, and *4-1BB* checkpoint genes and immune response in virus-positive tumors may contribute to therapy selection in these patients.

## Methods

**Virus integration.** We discovered viruses in the tumor samples by using the VirusScan pipeline<sup>30</sup>, which is available from Github<sup>64</sup>. For the identification of virus integration sites in human genome, we first used BWA<sup>65</sup> to align RNA-Seq

data to the human plus viral reference. From the re-aligned bam file, we extracted the discordant read pairs, where one read of a read pair maps to human, the other to virus. Pindel<sup>66</sup> was used to identify exact breakpoints for all samples with ten or more human-virus discordant reads. The breakpoints in Fig. 1 were visualized by using BreakPointSurveyor<sup>67</sup>.

**Statistical analysis.** Survival analysis was implemented by using R package survival. We used the Student's *t*-test to extract differentially expressed genes in virus-positive and negative samples, using FDR = 0.05 as the cut-off. The FDR value was obtained by p.adjust with Benjamini and Hochberg correction from R package. The heatmap figure was generated by using *Heatmap.3R* package with default parameters.

**Expression outlier analysis.** To investigate if genes with virus integrations are expression outliers, we used the Tukey's standard formula to quantify the outlier score:

Outlier score =  $(x - Q3)/IQR$  for upper tail and  $(x - Q1)/IQR$  for low tail, where IQR is the interquartile range, Q1 and Q3 are the first and third quartiles, respectively and  $x$  is the RSEM value in a log<sub>2</sub> scale. In the current study, genes with an outlier score greater than 1.0 or less than -1.0 are considered to be expression outliers.

**Neoantigen prediction.** Different lengths of epitopes (8mer, 9mer, 10mer, and 11mer) are constructed from HPV16 protein sequences. We use NetMHC3pan<sup>68</sup> to predict the binding affinity between epitopes and MHC based on the HLA type in each tumor. The HLA type was adopted from ref.<sup>69</sup>. Epitopes with binding affinity ≤500 nM which are also not present in Ensembl 70.37 database are extracted for the following neoantigen analysis.

**Reporting Summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

We collected gene expression (RSEM), and clinical data from Broad firehose<sup>70</sup> across six cancer types including cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon adenocarcinoma and rectal adenocarcinoma (COADREAD), esophageal cancer (ESCA), head/neck squamous cell carcinoma (HNSC), stomach adenocarcinomas (STAD), and liver hepatocellular carcinoma (LIHC) from The Cancer Genome Atlas (TCGA). The aligned TCGA RNA-Seq bam files included in this study can be downloaded from the NCI's Genomic Data Commons (GDC). The source gene expression data are available in the Supplementary Data 1.

Received: 30 August 2018 Accepted: 7 February 2019

Published online: 22 March 2019

## References

- Ley, T. J. et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
- Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
- Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
- Zhang, J. H. et al. Germline mutations in predisposition genes in pediatric cancer. *New Engl. J. Med.* **373**, 2336–2346 (2015).
- Lu, C. et al. Patterns and functional implications of rare germline variants across 12 cancer types. *Nat. Commun.* **6**, 10086 (2015).
- Rubio-Perez, C. et al. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals novel targeting opportunities. *Cancer Res.* **75**, 2983 (2015).
- Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69–74 (2015).
- Stronen, E. et al. Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. *Science* **352**, 1337–1341 (2016).
- Latchman, Y. et al. PD-L2 is a second ligand for PD-1 and inhibits T cell activation. *Nat. Immunol.* **2**, 261–268 (2001).
- Zou, W. P., Wolchok, J. D. & Chen, L. P. PD-L1 (B7-H1) and PD-1 pathway blockade for cancer therapy: mechanisms, response biomarkers, and combinations. *Sci. Transl. Med.* **8**, 328rv4 (2016).
- Kataoka, K. et al. Aberrant PD-L1 expression through 3'-UTR disruption in multiple cancers. *Nature* **534**, 402–406 (2016).
- Parsa, A. T. et al. Loss of tumor suppressor PTEN function increases B7-H1 expression and immunoresistance in glioma. *Nat. Med.* **13**, 84–88 (2007).
- Akbay, E. A. et al. Activation of the PD-1 pathway contributes to immune escape in EGFR-driven lung tumors. *Cancer Discov.* **3**, 1355–1363 (2013).
- Roussel, H. et al. Composite biomarkers defined by multiparametric immunofluorescence analysis identify ALK-positive adenocarcinoma as a potential target for immunotherapy. *Oncotimmunology* **6**, e1286437 (2017).
- Soo, R. A. et al. Immune checkpoint inhibitors in epidermal growth factor receptor mutant non-small cell lung cancer: Current controversies and future directions. *Lung Cancer* **115**, 12–20 (2018).
- Freeman, G. J. et al. Engagement of the PD-1 immunoinhibitory receptor by a novel B7 family member leads to negative regulation of lymphocyte activation. *J. Exp. Med.* **192**, 1027–1034 (2000).
- Eroglu, Z. et al. High response rate to PD-1 blockade in desmoplastic melanomas. *Nature* **553**, 347–350 (2018).
- Gopalakrishnan, V. et al. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* **359**, 97–103 (2018).
- Matson, V. et al. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science* **359**, 104–108 (2018).
- Le, D. T. et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017).
- Leach, D. R., Krummel, M. F. & Allison, J. P. Enhancement of antitumor immunity by CTLA-4 blockade. *Science* **271**, 1734–1736 (1996).
- Wei, S. C. et al. Distinct cellular mechanisms underlie anti-CTLA-4 and anti-PD-1 checkpoint blockade. *Cell* **170**, 1120–1133 e1117 (2017).
- Iglesia, M. D. et al. Genomic analysis of immune cell infiltrates across 11 tumor types. *J. Natl Cancer Inst.* <https://doi.org/10.1093/jnci/djw144> (2016).
- Li, B. et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* **17**, 174 (2016).
- Goldszmid, R. S., Dzutsev, A. & Trinchieri, G. Host immune response to infection and cancer: unexpected commonalities. *Cell Host Microbe* **15**, 295–305 (2014).
- Fakhry, C. et al. Improved survival of patients with human papillomavirus-positive head and neck squamous cell carcinoma in a prospective clinical trial. *J. Natl Cancer Inst.* **100**, 261–269 (2008).
- Lyford-Pike, S. et al. Evidence for a role of the PD-1:PD-L1 pathway in immune resistance of HPV-associated head and neck squamous cell carcinoma. *Cancer Res.* **73**, 1733–1741 (2013).
- Badoual, C. et al. PD-1-expressing tumor-infiltrating T cells are a favorable prognostic biomarker in HPV-associated head and neck cancer. *Cancer Res.* **73**, 128–138 (2013).
- Chakravarthy, A. et al. Human papillomavirus drives tumor development throughout the head and neck: improved prognosis is associated with an immune response largely restricted to the oropharynx. *J. Clin. Oncol.* **34**, 4132–4141 (2016).
- Cao, S. et al. Divergent viral presentation among human tumors and adjacent normal tissues. *Sci. Rep.* **6**, 28294 (2016).
- Hutloff, A. et al. ICOS is an inducible T-cell co-stimulator structurally and functionally related to CD28. *Nature* **397**, 263–266 (1999).
- Fan, X., Quezada, S. A., Sepulveda, M. A., Sharma, P. & Allison, J. P. Engagement of the ICOS pathway markedly enhances efficacy of CTLA-4 blockade in cancer immunotherapy. *J. Exp. Med.* **211**, 715–725 (2014).
- Parfenov, M. et al. Characterization of HPV and host genome interactions in primary head and neck cancers. *Proc. Natl Acad. Sci. USA* **111**, 15544–15549 (2014).
- Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 323 (2011).
- Chen, B. J. et al. PD-L1 expression is characteristic of a subset of aggressive B-cell lymphomas and virus-associated malignancies. *Clin. Cancer Res.* **19**, 3462–3473 (2013).
- Cancer Genome Atlas Research, N. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
- Iglesia, M. D. et al. Prognostic B-cell signatures using mRNA-seq in patients with subtype-specific breast and ovarian cancer. *Clin. Cancer Res.* **20**, 3818–3829 (2014).
- Palmer, C., Diehn, M., Alizadeh, A. A. & Brown, P. O. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genom.* **7**, 115 (2006).
- Schmidt, M. et al. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.* **68**, 5405–5413 (2008).
- Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
- Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
- Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
- Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
- Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).

45. Rimsza, L. M. et al. Loss of MHC class II gene and protein expression in diffuse large B-cell lymphoma is related to decreased tumor immunosurveillance and poor patient survival regardless of other prognostic factors: a follow-up study from the Leukemia and Lymphoma Molecular Profiling Project. *Blood* **103**, 4251–4258 (2004).
46. Kinsey, C. et al. Plac8 links oncogenic mutations to regulation of autophagy and is critical to pancreatic cancer progression. *Cell Rep.* **7**, 1143–1155 (2014).
47. Cancer Genome Atlas, N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).
48. Evan, G. I. & Vousden, K. H. Proliferation, cell cycle and apoptosis in cancer. *Nature* **411**, 342–348 (2001).
49. Appay, V. et al. Memory CD8+T cells vary in differentiation phenotype in different persistent virus infections. *Nat. Med.* **8**, 379–385 (2002).
50. Badve, S. et al. FOXA1 expression in breast cancer—correlation with luminal subtype A and survival. *Clin. Cancer Res.* **13**, 4415–4421 (2007).
51. Kimple, R. J. et al. Enhanced radiation sensitivity in HPV-positive head and neck cancer. *Cancer Res.* **73**, 4791–4800 (2013).
52. Ichim, G. & Tait, S. W. A fate worse than death: apoptosis as an oncogenic process. *Nat. Rev. Cancer* **16**, 539–548 (2016).
53. Lin, R. J. et al. Cyclin D1 overexpression is associated with poor prognosis in oropharyngeal cancer. *J. Otolaryngol. Head Neck Surg.* **42**, 23 (2013).
54. Sekiya, T. et al. The nuclear orphan receptor Nr4a2 induces Foxp3 and regulates differentiation of CD4+T cells. *Nat. Commun.* **2**, 269 (2011).
55. Yamazaki, T. et al. A butyrophilin family member critically inhibits T cell activation. *J. Immunol.* **185**, 5907–5914 (2010).
56. Hsiao, H. W. et al. Deltex1 is a target of the transcription factor NFAT that promotes T cell anergy. *Immunity* **31**, 72–83 (2009).
57. Rahrmann, E. P. et al. Identification of PDE4D as a proliferation promoting factor in prostate cancer using a sleeping beauty transposon-based somatic mutagenesis screen. *Cancer Res.* **69**, 4388–4397 (2009).
58. Herbst, R. S. et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature* **515**, 563–567 (2014).
59. Ferris, R. L. et al. Nivolumab for recurrent squamous-cell carcinoma of the head and neck. *N. Engl. J. Med.* **375**, 1856–1867 (2016).
60. Harkins, L. et al. Specific localisation of human cytomegalovirus nucleic acids and proteins in human colorectal cancer. *Lancet* **360**, 1557–1563 (2002).
61. Bender, C. et al. Analysis of colorectal cancers for human cytomegalovirus presence. *Infect. Agent. Cancer* **4**, 6 (2009).
62. Collins, D., Hogan, A. M. & Winter, D. C. Microbial and viral pathogens in colorectal cancer. *Lancet Oncol.* **12**, 504–512 (2011).
63. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–5 (2016).
64. Cao, S. & Ding, L. VirusScan Pipeline. <https://github.com/ding-lab/VirusScan> (2016).
65. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
66. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. M. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
67. Wyczalkowski, M. A. et al. BreakPoint Surveyor: a pipeline for structural variant visualization. *Bioinformatics* **33**, 3121–3122 (2017).
68. Nielsen, M. & Andreatta, M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* **8**, 33 (2016).
69. Thorsson, V. et al. The immune landscape of cancer. *Immunity* **48**, 812–830 (2018).
70. Broad firehose. [http://gdac.broadinstitute.org/runs/stddata\\_2016\\_01\\_28/](http://gdac.broadinstitute.org/runs/stddata_2016_01_28/).

## Acknowledgements

This work was supported by the National Cancer Institute grant R01CA178383 and National Human Genome Research Institute grant U01HG006517 to L.D. The Cancer Genome Atlas ([cancergenome.nih.gov](http://cancergenome.nih.gov)) was the source of primary data.

## Author contributions

L.D. designed and supervised research. S.C., K.W., M.A.W., A.K., J.L., S.S., and R.J.M. analyzed the data. S.C. performed statistical analysis. S.C., M.A.W., A.K., and W.L. prepared figures and tables. X.W., K.J., J.F.D., H.G., L.R., F.C., D.R.A., and L.D. provided disease specific analysis and guidance. S.C. and L.D. wrote the manuscript. K.W., R.J.M., K.J., J.F.D., H.G., F.C., D.R.A., and L.D. revised the manuscript.

## Additional information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s42003-019-0352-3>.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019