Analysis of the immune landscape in virus-induced cancers using a novel integrative mechanism discovery approach

Lindsay M. Wong [a,b,1], Wei Tse Li [a,b,1], Neil Shende [a,b,1], Joseph C. Tsai [a,b], Jiayan Ma [a,b], Jaideep Chakladar [a,b], Aditi Gnanasekar [a,b], Yuanhao Qu [a,b], Kypros Dereschuk [a,b], Jessica Wang-Rodriguez [c,d], Weg M. Ongkeko [a,b,*]

[a] Department of Surgery, Division of Otolaryngology-Head and Neck Surgery, University of California, San Diego, La Jolla, CA, 92093, USA
[b] Research Service, VA San Diego Healthcare System, San Diego, CA 92161, USA
[c] Department of Pathology, University of California San Diego, La Jolla, CA 92093, USA
[d] Pathology Service, VA San Diego Healthcare System, San Diego, CA 92161, USA

A B S T R A C T

Background: The mechanisms of carcinogenesis from viral infections are extraordinarily complex and not well understood. Traditional methods of analyzing RNA-sequencing data may not be sufficient for unraveling complicated interactions between viruses and host cells. Using RNA and DNA-sequencing data from The Cancer Genome Atlas (TCGA), we aim to explore whether virus-induced tumors exhibit similar immune-associated (IA) dysregulations using a new algorithm we developed that focuses on the most important biological mechanisms involved in virus-induced cancers. Differential expression, survival correlation, and clinical variable correlations were used to identify the most clinically relevant IA genes dysregulated in 5 virus-induced cancers (HPV-induced head and neck squamous cell carcinoma, HPV-induced cervical cancer, EBV-induced stomach cancer, HBV-induced liver cancer, and HCV-induced liver cancer) after which a mechanistic approach was adopted to identify pathways implicated in IA gene dysregulation.
Results: Our results revealed that IA dysregulations vary with the cancer type and the virus type, but cytokine signaling pathways are dysregulated in all virus-induced cancers. Furthermore, we also found that important similarities exist between all 5 virus-induced cancers in dysregulated clinically relevant oncogenic signatures and IA pathways. Finally, we also discovered potential mechanisms for genomic alterations to induce IA gene dysregulations using our algorithm.
Conclusions: Our study offers a new approach to mechanism identification through integrating functional annotations and large-scale sequencing data, which may be invaluable to the discovery of new immunotherapy targets for virus-induced cancers.

© 2021 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Background

Virus-induced cancers account for approximately 15–20% of human cancers [1]. Regarded as the second most important etiol-

ogy of cancer, viruses can directly and indirectly affect major cellular processes and pathways at varying stages of tumorigenesis [1,2]. The mechanism by which viruses induce cancer, while extensively investigated, is extremely complicated, with many viruses known to have more than one way of contributing to cancer development [3]. Suggested potential mechanisms of viral oncogenesis include altered levels of protein expression, genomic modifications, chronic inflammation, and immunosuppression [2]. The immune mediation by viruses, also known as the indirect mechanism, may be more amenable to therapeutic targeting or preventive care because the immune system is more druggable than unseen viruses hidden in cells [3]. By studying well-known viruses in virus-induced cancers such as human papillomavirus (HPV) in head and neck squamous cell carcinoma (HNSCC) and cervical squamous cell carcinoma and endocervical adenocarcinoma

(CESC), hepatitis B (HBV) and C (HCV) in liver hepatocellular carcinoma (LIHC), and Epstein-Barr virus (EBV) in stomach adenocarcinoma (STAD), we aim to elucidate these mechanisms and provide potential targets for therapy.

Large-scale gene expression sequencing data exist on the above virus-induced cancers in TCGA. With sequencing data from thousands of patients, the TCGA database is a powerful resource for analysis and comparison of specific phenotypes. However, analysis of large-scale sequencing experiments is fraught with difficulties. The traditional approach of analyzing genes differentially expressed in one cohort vs. another (i.e. virus-induced cancer samples vs. normal samples) would yield a long list of statistically significant genes that may not reveal which genes are the true drivers of phenotype, which are dysregulated as a consequence of another gene being dysregulated, and which are purely statistical noise that have little biological significance. Out of these concerns, the idea of analyzing gene pathways was born [4]. Analysis of pathways allows analysis of the transcriptome using known biological pathways, or sets of genes with known functions. Thus, dysregulated genes are grouped into functional clusters. This grouping theoretically reveals the underlying biological mechanism suggested by the transcriptome that is not easily discovered by looking at a list of dysregulated genes. Furthermore, if many genes within a pathway are dysregulated in the same direction, the dysregulation of these genes would more likely be biologically relevant rather than simply statistical artifacts [5]. Gene set enrichment analysis (GSEA) is the most popular software for pathways analysis [6]. However, despite its popularity and promise, significant shortcomings exist in its methods. First, the pathways and gene sets assembled for use with GSEA may not represent actual coherent expression [7]. Gene sets group together genes that upregulate the pathway and genes that downregulate the pathway, so a concerted expression of all genes in the gene set may never be observed. Some gene sets also represent multiple biological processes that may not be dysregulated concurrently in any cell [7]. Second, the gene sets within the database frequently contain overlapping genes, and these overlapping gene sets may describe processes that are related but decidedly not the same, such as B cell receptor signaling and the complement cascade [4]. Third, there is the problem of gene-gene correlation within gene sets, where gene sets with genes that are frequently co-expressed are more likely to produce an artificially strong correlation of the pathway to a target [8]. Fourth, choice of significance cutoff, such as the GSEA-recommended false discovery rate (FDR) < 0.25 rather than the traditional FDR < 0.05, can affect results dramatically [6]. Therefore, there is a need to develop new computational methods for more accurate determination of the true biological processes dysregulated when analyzing sequencing data.

In this study, we developed a novel gene dysregulation analysis pipeline that integrates both pathways analysis and individual-gene analysis. Using this pipeline, we analyzed immune-associated (IA) gene dysregulation resulting from viral infection in each of the five comparisons in order to locate common or unique pathways that may explain the similarities or differences between virus positive and virus negative tumorigenesis. Using data from TCGA, we identified significantly dysregulated genes and correlated their expressions with patient survival, clinical variables, and immune relevance to identify a panel of clinically relevant dysregulated IA genes for each comparison. We utilized these dysregulated genes to identify the most clinically relevant IA pathways, immunologic states, and oncogenic signatures for each cohort and found overlapping pathways and signatures between the five virus-induced cancers. Finally, using a new algorithm we developed, we explored the potential mechanisms of viruses to induce IA gene dysregulation through genomic alterations and reprogramming of cancer pathways. Our novel approach could accelerate the identification of previously unknown mechanisms for viral-induced cancers that may be clinically actionable, which may improve the efficacy of existing therapeutic methods or provide new targets.

2. Results

2.1. Conceptual framework

Viral infections can significantly alter the immunologic landscape in human tissues, and certain viruses could increase the risk of cancer development dramatically. The mechanisms of viral carcinogenesis are diverse and multi-faceted but revolve around immune mediation and direct stimulation of oncogenic properties [3]. While viruses have different and diverse effects on the immune system, it remains to be discovered which immune pathways are the primary targets of oncogenic viruses and the mechanism through which these viruses modulate the immune system. Using data from TCGA, we will focus on immune-associated (IA) genes' dysregulation at cancer sites to deduce the effects of each virus on the immune response and inflammatory pathways. Furthermore, we aim to computationally derive the most probable pathway causing such immune dysregulation.

To address the shortcomings in traditional analyses of RNA sequencing data, we developed a new conceptual framework, or a computational workflow, that will allow us to consider a diverse array of data and analysis methods before formulating definitive analysis results. This framework seeks to integrate two major branches of analysis, pathways-level analysis and gene-level analysis, before taking in genomic alterations data to deduce a probable pathway (Fig. 1A). The purpose of our workflow is to computationally identify a mechanism for immune-associated gene dysregulations by integrating analyses examining different aspects of gene regulation. First, differential expression analysis (edgeR, $p < 0.05$) is performed between the virus-induced cancer samples and control samples. Second, the fold changes for each gene are input into GSEA on the pathways-level arm (left side of circle in Fig. 1A). Third, the individual significantly dysregulated genes are filtered for immune relevance and then are correlated with patient survival followed by clinical variables, composing our gene-level arm (right side of circle in Fig. 1A). Individual genes that are significantly dysregulated and correlated with survival and other clinical variables will be retained as candidates for the next step and termed as clinically relevant IA genes. Fourth, we identified genes that are close in function to these clinically relevant IA genes using ReactomeFIViz, a software that provides functional annotations of genes in a network. These genes, close in function, are termed neighboring genes. Fifth, we query these genes in significantly dysregulated cancer or immune-associated pathways to identify the pathways that are most clinically and biologically relevant. Finally, a list of genes containing clinically relevant IA genes, neighboring genes, and genes within the clinically/biologically relevant pathways is inputted into our pathways inference algorithm, along with correlations between genomic alterations and IA gene expression. The algorithm will perform a breadth-first search to determine the shortest pathway between genomic alteration and IA gene dysregulation.

Our conceptual framework sidesteps traditional limitations in gene-level and pathways-level analysis by integrating both types of analyses. In pathways analysis, while numerous pathways could be statistically significant, it may be difficult to locate the true pathways that are biologically dysregulated. In gene-level analysis, pathways and functional structures are hard to obtain, although the most dysregulated genes are still apparent. In our approach, we identified the most significant dysregulated IA genes and
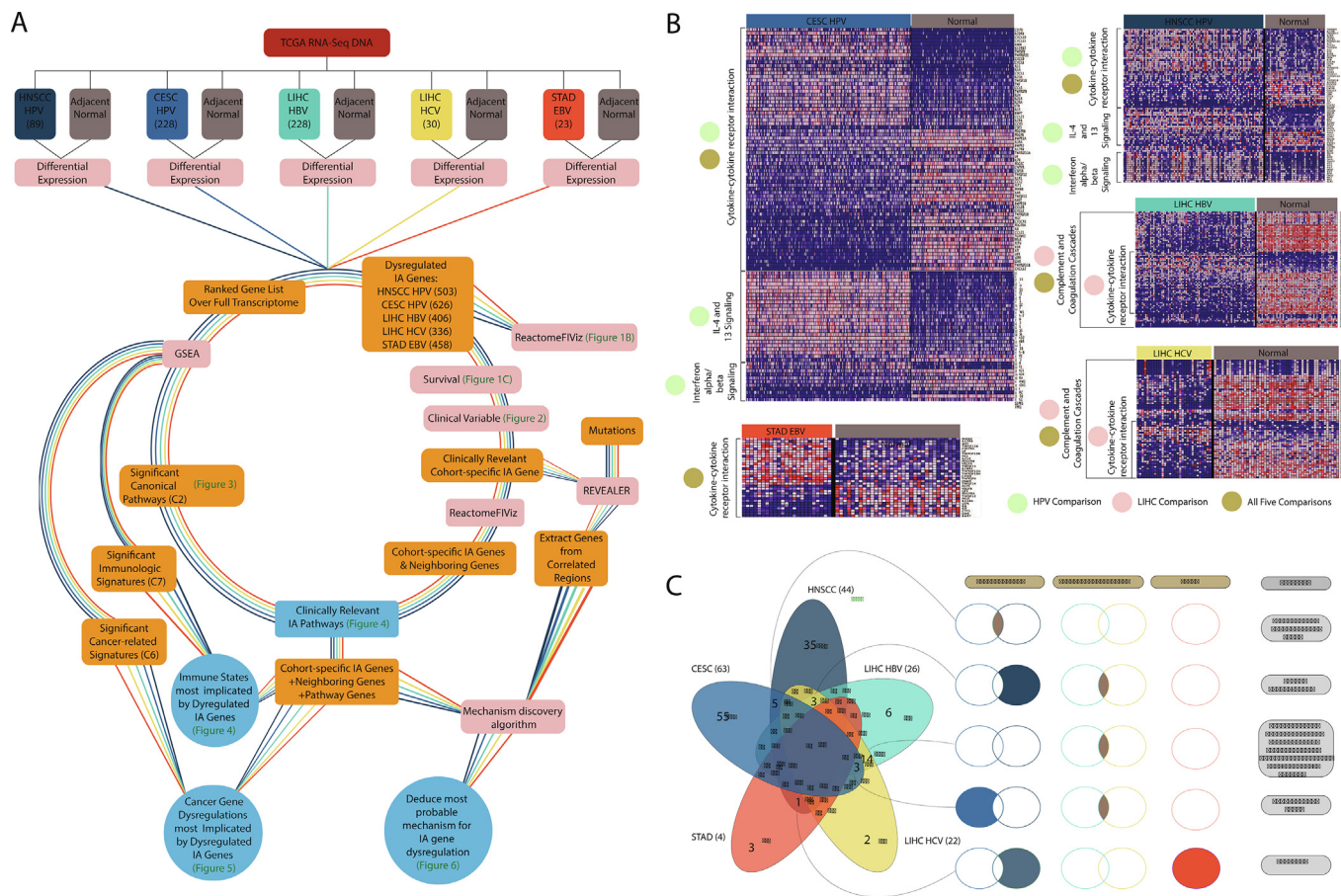
Fig. 1. Summary of study procedures, differential expression results, and survival correlation comparisons. (A) Schematic of computational analyses and data processing procedures used. The direction of workflow is always downwards and is sometimes indicated by converging colored lines. Left of the circle involves pathways-level analysis while right of the circle indicates the gene-level analysis. Orange, pink, and blue boxes indicate genes or gene sets, analyses, and results, respectively. Number of patients in each cohort are indicated within the parentheses. (B) Heatmap of significantly dysregulated IA genes visualized in pathway annotations (FDR < 1x10^{10}) from ReactomeFIViz when comparing virus samples to normal samples in each cohort. Green, pink, and gold circles indicate that they are found in HPV, LIHC, and in all five comparisons, respectively. (C) Venn diagram demonstrating unique IA genes and overlapping IA genes between cohorts after filtering differentially expressed IA genes for patient survival. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

filtered them for clinical relevance first to eliminate genes that may not be critical to the disease. Next, we expanded this list of most important dysregulated genes to include genes that are their direct interactors. This list is then searched within IA pathways to determine the most clinically relevant pathways that are dysregulated, which should be pathways that contain the greatest number of genes from our list of dysregulated genes and their direct interactors. Thus, we were able to benefit from the functional insight of pathways-level analysis while maintaining focus on the most dysregulated genes (gene-level analysis) at the same time. Furthermore, with our mechanism discovery algorithm based on a further-expanded gene list and genomic alteration correlations, we will be able to derive our own mechanistic pathway with minimal reliance on existing pathways, which is often too narrow or too broad to accommodate the specific phenotype of study.

2.2. Identification of significant dysregulations in the immune landscape of virus-induced cancers

We obtained RNA-sequencing data for HNSCC, CESC, LIHC, and STAD from TCGA for patients with documented presence of HPV, HBV, HCV, or EBV. HPV is known to cause HNSCC and CESC, HBV and HCV are known to cause LIHC, and EBV is known to cause STAD. We first identified differentially expressed IA genes in virus-induced cancer samples compared to normal samples

(Fig. 1A). The significantly dysregulated IA genes across the five cohorts of virus-induced cancers (HNSCC-HPV, CESC-HPV, LIHC-HBV, LIHC-HCV, and STAD-EBV) were clustered into pathways using the ReactomeFIViz pathway annotation. The most representative pathways to which the differentially expressed IA genes belong to are plotted in Fig. 1B (FDR < 1 × 10^{10}). Several pathways are significantly dysregulated in multiple cohorts. We discovered that a significant number of IA genes within the cytokine-cytokine receptor interaction pathway, which contains cytokines and their receptors, are dysregulated across all five cohorts. However, the direction of dysregulation is not the same for different cohorts. In virus-induced LIHC, most genes within the cytokine-cytokine receptor interaction pathway are downregulated, while the numbers of genes upregulated and downregulated in the cytokine-cytokine receptor interaction pathway are comparable in HNSCC, CESC, and STAD. The IL4/IL13 signaling and IFN-alpha/beta signaling pathways are dysregulated in both HPV-induced CESC and HNSCC, while genes involved in the complement and coagulation cascade are dysregulated in both HBV and HCV-induced LIHC.

We next correlated the expressions of dysregulated IA genes to patient survival and identified a list of survival-correlated IA genes (Cox regression, $p < 0.05$). No IA gene is survival-associated across all cohorts, and most survival-correlated genes are unique to each cohort (Fig. 1C). However, HBV and HCV-induced LIHC possess a
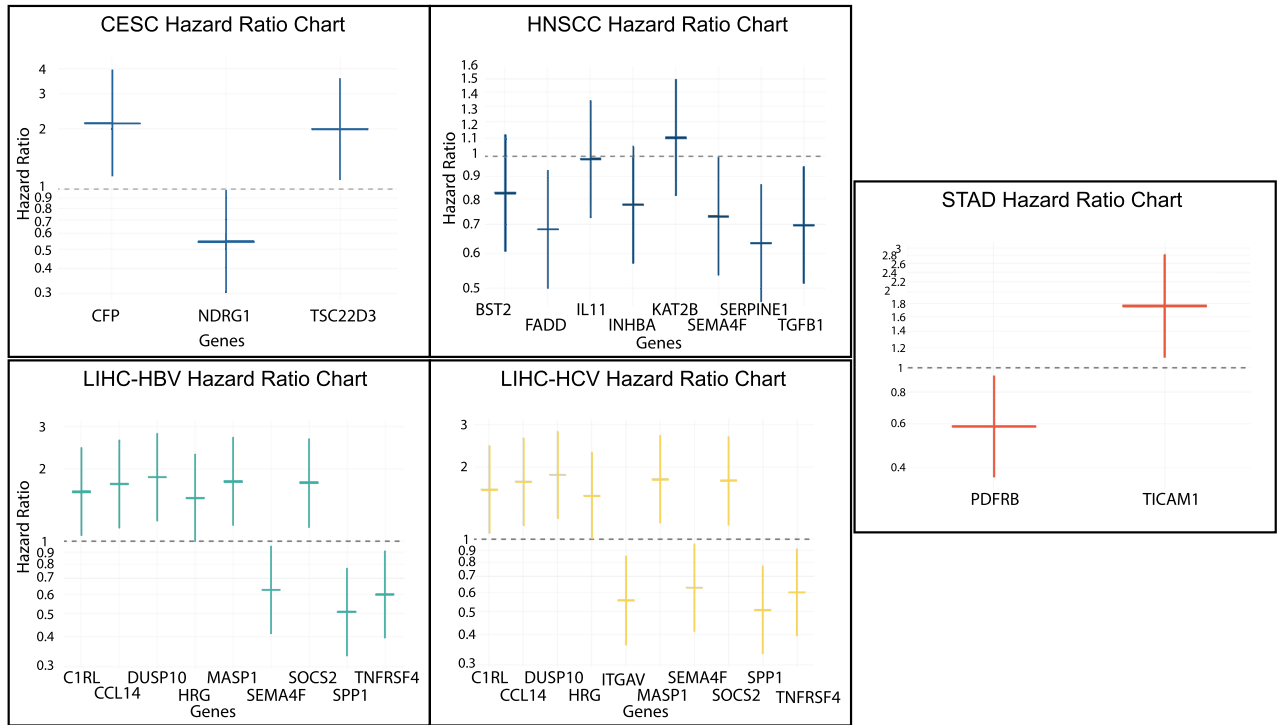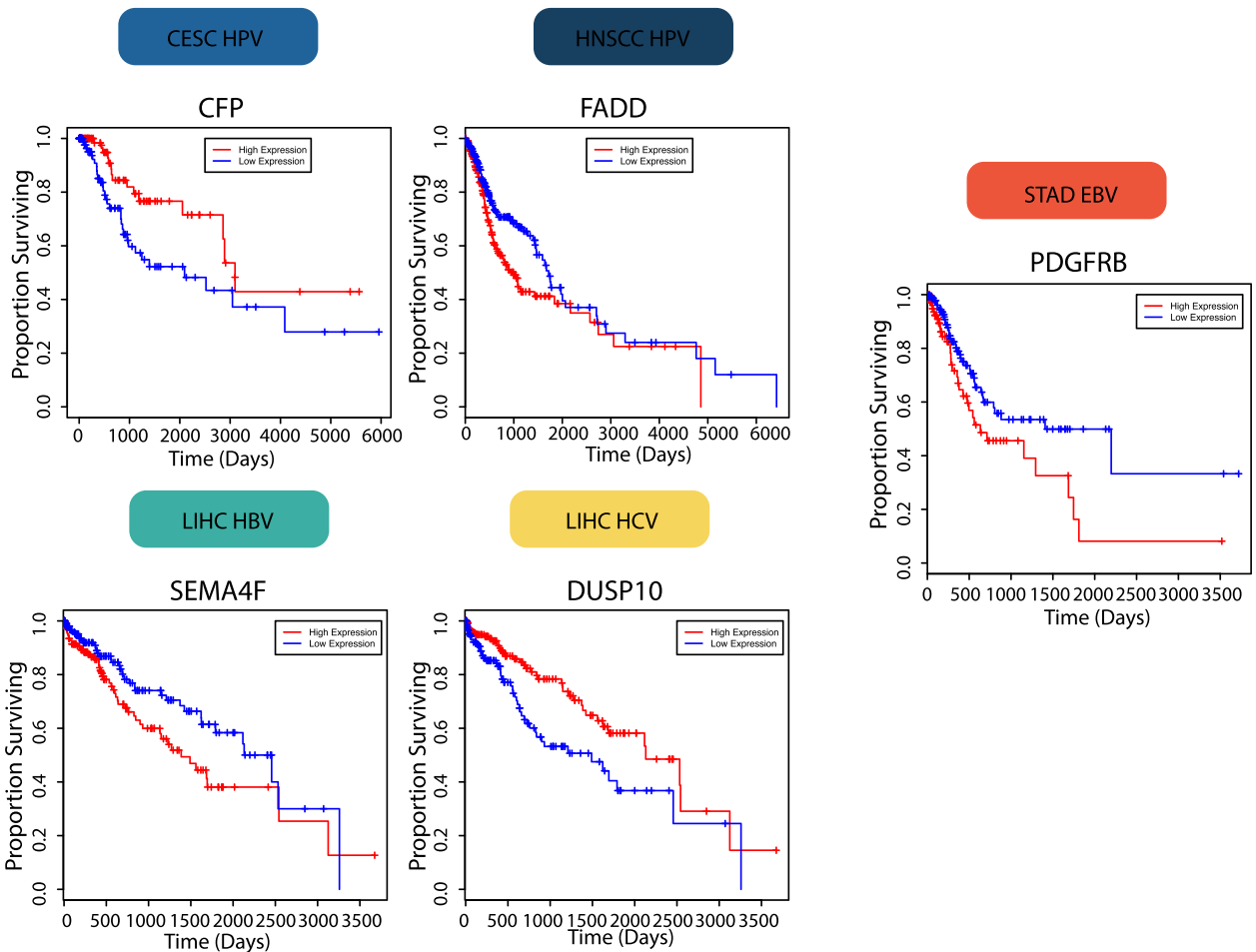
large number of survival-associated IA genes in common, suggesting that the different viral origins lead to LIHC with similar prognostically-relevant immune dysregulations. The HPV-induced cancers exhibit five common survival-associated IA genes, suggesting that HPV leads to similar prognostically-relevant immune dysregulation in HNSCC and CESC. Interestingly, very few genes correlated with survival in STAD. One potential reason for this is that there were fewer STAD patients compared to any other cancer cohort, making it harder for correlations to be significant. Additionally, we used strict logFC and FDR cutoffs to filter genes when performing differential expression analysis, which could have also resulted in fewer STAD genes being selected.

2.3. Identification of most clinically relevant dysregulated IA genes

In order to identify biological pathways whose dysregulation has the most impact on clinical phenotype and prognosis, we correlated the gene expression of survival-associated genes with clinical variables and identified a set of most clinically relevant dysregulated IA genes, which we defined as dysregulated IA genes whose expression correlates with patient survival and two or more clinical variables (Kruskal-Wallis test, $p < 0.05$). Three IA genes were retained for CESC, two were retained for STAD, while eight to ten were retained for other cohorts (Fig. 2A). Hazard ratios greater than one correspond to better patient prognosis while hazard ratios less than one correspond to decreased patient survival. The genes retained for LIHC cohorts tended to have hazard ratios greater than one, which demonstrates that the downregulation of these genes corresponds to decreased patient survival. On the other hand, the genes retained for HNSCC tended to have hazard ratios less than one, which indicate that the upregulation of these genes correspond to higher patient mortality. Both STAD and LIHC-HCV had around equal numbers of retained IA genes that were associated with lower patient survival when upregulated or downregulated. Examples of Kaplan Meier survival graphs for a selected gene for each cohort are presented in Fig. 2B.

The majority of clinical variable correlations were with respect to histologic grade, pathologic T or clinical T stage, and pathologic or clinical stage (Fig. 2C). For most of the associated clinical variables including neoplasm cancer status, clinical pathologic T, and residual tumor, it appears that the LIHC cohorts contained the most significant IA gene-clinical variable correlations in comparison to other cohorts. However, HPV-induced HNSCC contained the greatest proportion of correlations with the clinical variables clinical pathologic M and presence of perineurial invasion. Both LIHC cohorts and HPV-induced cohorts contained approximately equal significant IA gene-clinical variable pairs for the clinical variables clinical pathologic N. Finally, HPV-induced CESC contained the greatest number of IA gene-clinical variable pairs with clinical pathologic stage. The absence of more correlations with N and M stages suggests that virus-induced IA dysregulation has a much higher impact on local tumor growth than regional or distant metastasis (Supplementary Fig. 1). For all eight clinical variables,

the cohorts that tended to have the most significant p-value per clinical variable include HPV-induced HNSCC, HBV-induced LIHC, and HCV-induced LIHC (Fig. 2D). Both LIHC cohorts had identical p-values as this analysis is performed using all patients labeled as LIHC.

2.4. Identification of IA and Cancer-Associated (CA) pathways dysregulation through GSEA

GSEA was used to identify the pathways containing genes that are collectively dysregulated in each cohort compared to normal samples. Curated canonical pathways for cancer and immune-related processes were downloaded from the Molecular Signature Database (MSigDB). We visualized the distribution of dysregulated IA/CA pathways associated with each cohort (Fig. 3A). Cancer-associated (CA) pathways are defined as pathways classically implicated in tumor development, progression, or metastasis, as classified by MSigDB. We found that a similar distribution was found for all cohorts except for the CESC cohort, which had few pathways with general innate and general adaptive responses. The HPV-induced cancers seemed to have higher enrichment of viral infection response pathways in comparison to the other cohorts. We next separately visualized the enriched pathways into IA vs. CA pathways (Supplementary Fig. 2). A log(p-value) to the left of the 0 line corresponds to downregulation of the pathway while a log(p-value) to the right of the 0 line corresponds to the upregulation of the pathway. In HPV-induced cancers, more IA pathways are upregulated than downregulated, while in LIHC cohorts, more IA pathways are downregulated (Supplementary Fig. 2A-2D). As we expected, the majority of IA pathways dysregulated in HBV-induced LIHC is also dysregulated in HCV-induced LIHC, but several pathways are uniquely upregulated in HCV-induced LIHC, namely interferon-related pathways, antigen cross presentation pathways, and antiviral immunity pathways (i.e. RIG-I-like receptors) (Supplementary Fig. 2A, 2B) [9]. Between the HPV cohorts, cytokine signaling pathways, antigen cross presentation, HIV infection, and interferon signaling pathways are commonly upregulated (Supplementary Fig. 2C, 2D). Interestingly, many of these upregulated pathways are also upregulated in HCV-induced LIHC, but not in HBV-induced LIHC. In particular, the JAK/STAT pathway is downregulated in both LIHC cohorts, while the complement cascade pathway is downregulated in HPV+ CESC (Fig. 3B). In EBV-induced STAD, a larger number of cancer-associated pathways are downregulated, and an even larger number of IA pathways are upregulated compared to all other cohorts (Supplementary Fig. 2E).

2.5. Identification of most clinically relevant IA pathways in virus-induced cancers

We identified pathways most implicated with clinically relevant IA gene dysregulations through a mechanistic approach inte-

Fig. 2. IA gene correlations with patient survival and clinical variables in virus versus normal samples for all five comparisons. (A) Hazard ratio plots of significantly dysregulated differentially expressed IA genes for each of the five comparisons. Only genes with two or more significant clinical variable correlations were displayed. Individual IA genes with center lines greater or less than the hazard ratio cutoff of one indicate that downregulation or upregulation of the gene corresponds to worse patient survival, respectively. Whiskers extending from the center lines denote the confidence interval. (B) Kaplan Meier plots of select significantly dysregulated differentially expressed IA genes for each of the five comparisons. (C) Pie charts demonstrating the proportion of significantly dysregulated IA genes that are correlated with their respective cohort across all five cohorts for the clinical variables neoplasm histologic grade, cancer neoplasm status, perineural invasion presence, residual tumor, clinical/pathologic stage, clinical/pathologic T, clinical/pathologic N, and clinical/pathologic M (Kruskal-Wallis, $p < 0.05$). (D) Boxplot examples of the most significant clinical variable-IA gene correlations for each of the eight clinical variables shown in Fig. 2C.

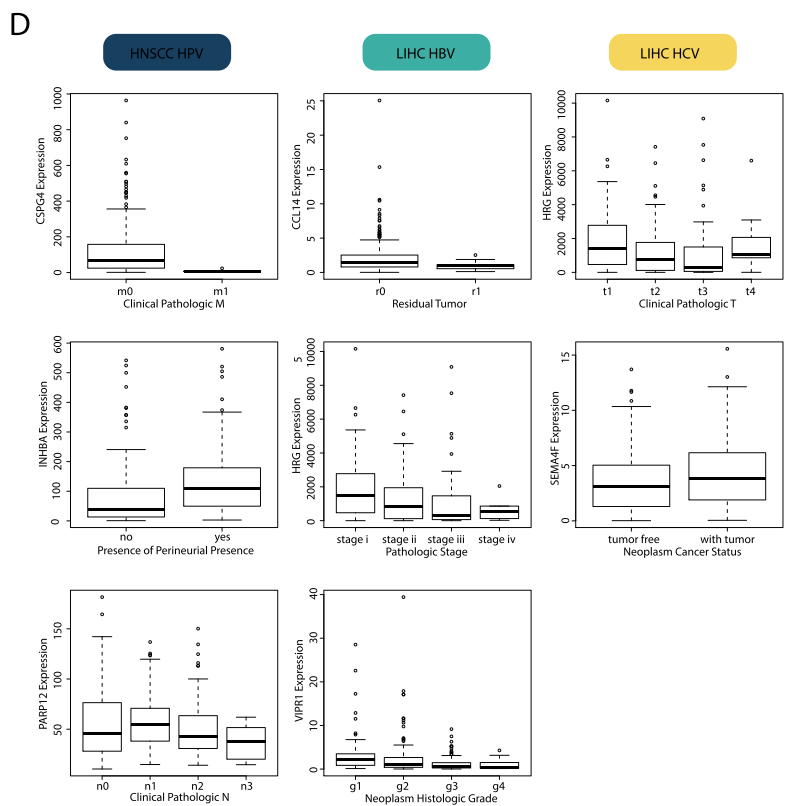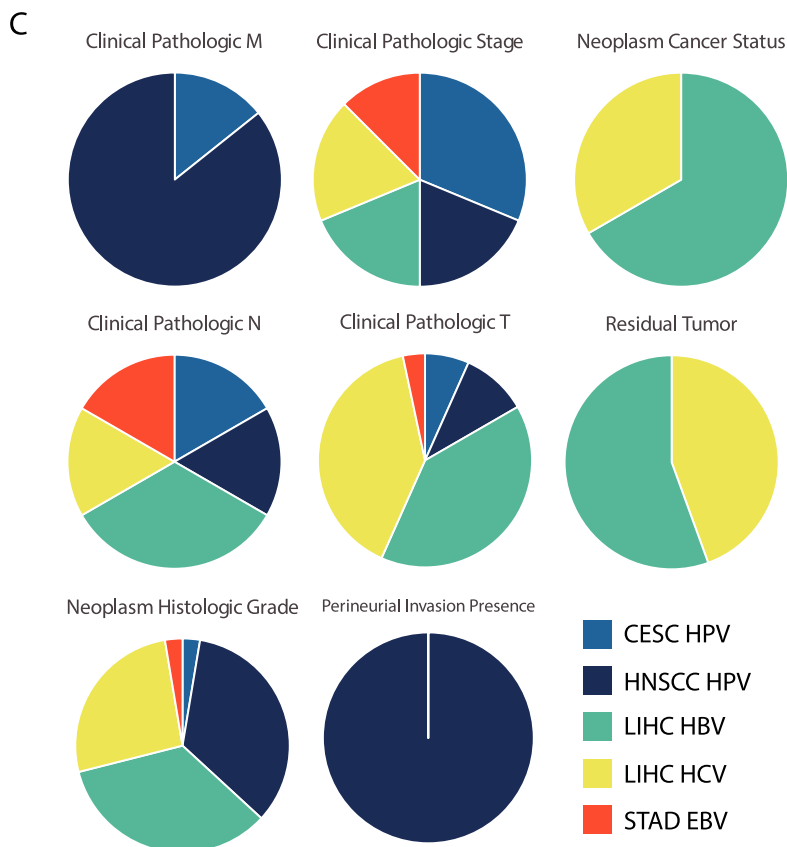**Fig. 2** (*continued*)

grating individual gene-level analyses with GSEA pathway enrichment. Briefly, clinically relevant IA genes and differentially expressed IA genes adjacent to them in a ReactomeFIViz interaction map, which is compiled from publicly available network regulations databases, such as KEGG, Reactome, and Pathways Interactions Database, are searched within the canonical pathways
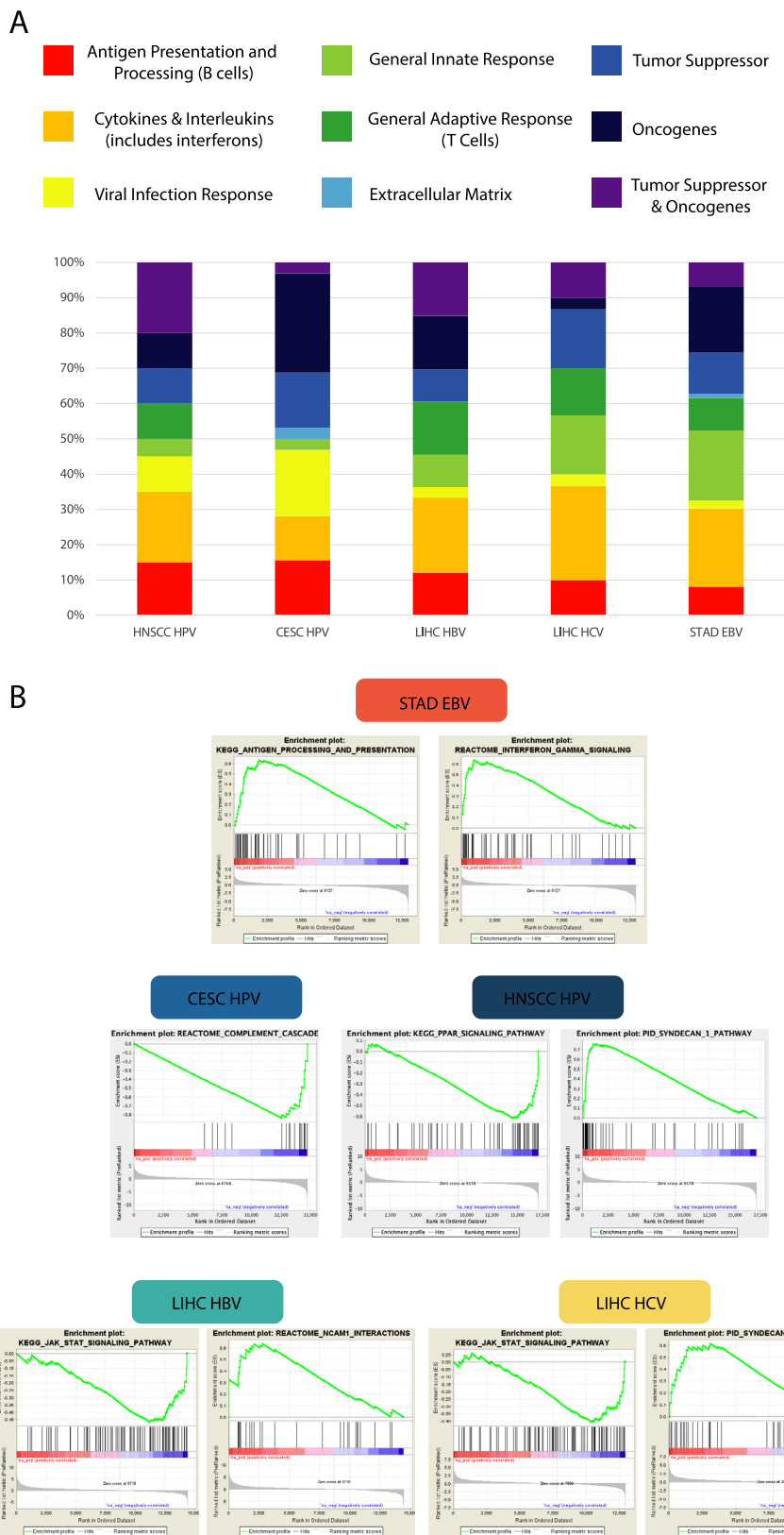
Fig. 3. Canonical (C2) IA and CA pathway enrichment using GSEA. (A) Stacked bar plot demonstrating the proportions of pathways that fall into nine categories (Antigen Presentation and Processing (B cells), Cytokines & Interleukins (includes interferons), Viral Infection Response, General Innate Response, General Adaptive Response (T cells), Extracellular Matrix, Tumor Suppressor, Oncogenes, Tumor Suppressor & Oncogenes) for each cohort (FDR < 0.01). (B) Select GSEA plot examples of pathways categorized in Fig. 3A (FDR < 0.01). The green peak or valley in the GSEA plot corresponds to the upregulation or downregulation of the pathway listed in the plot title, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
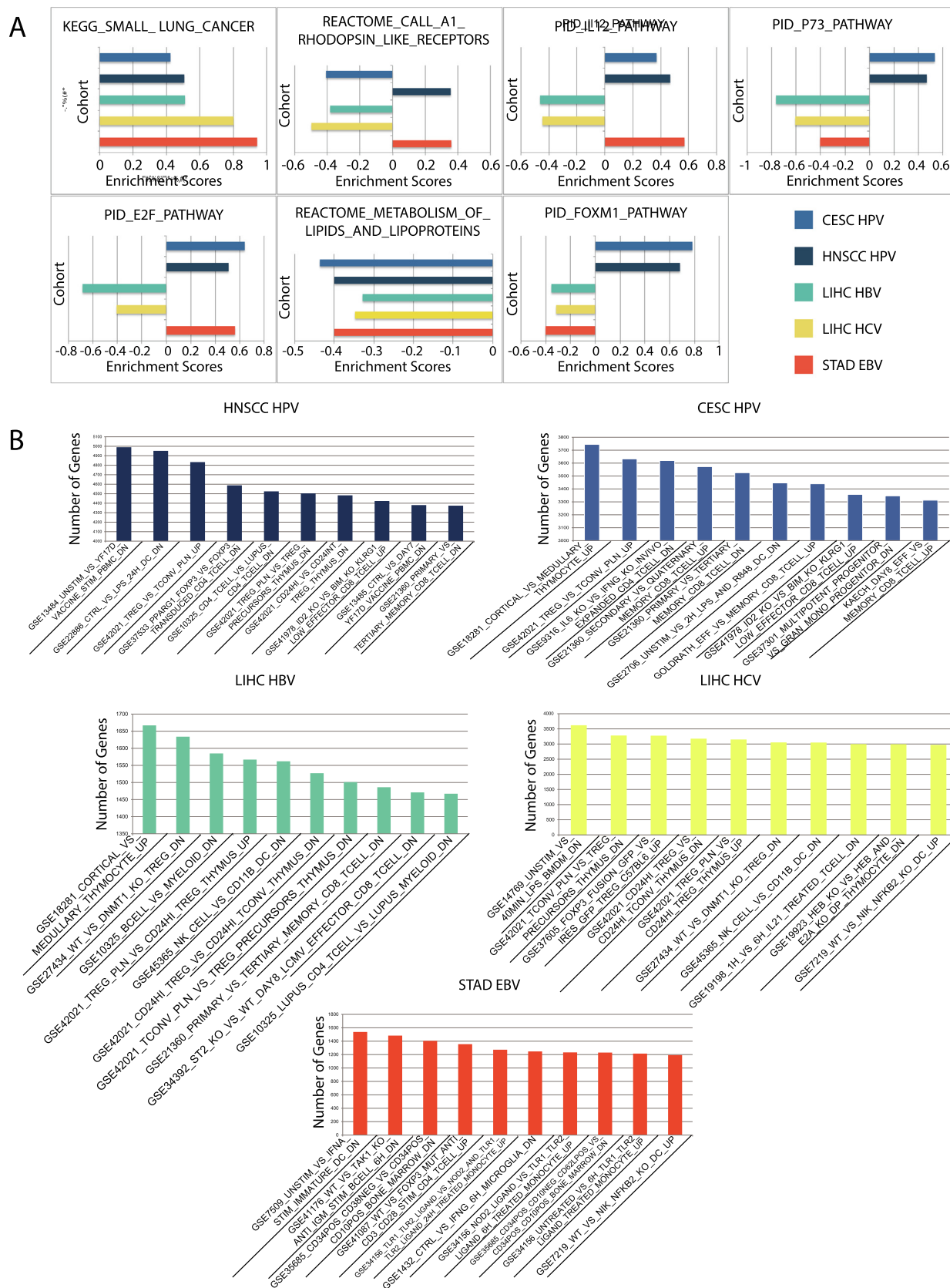
Fig. 4. Canonical IA pathway and immunologic signatures most implicated in each cohort following integration with gene-level analysis. (A) Horizontal bar graphs comparing enrichment scores of all five cohorts for canonical pathways (C2) with the greatest number of cohort-specific IA genes and neighboring genes, including Kegg small cell lung cancer, Reactome A1 Rhodopsin-like receptors, PID IL12 pathway, PID p73 pathway, PID E2F pathway, Reactome metabolism of lipids and lipoproteins, and PID FOXM1 pathway. (B) Vertical bar graphs demonstrating the top ten immunologic signatures (C7) with the greatest number of cohort-specific IA genes and neighboring genes.

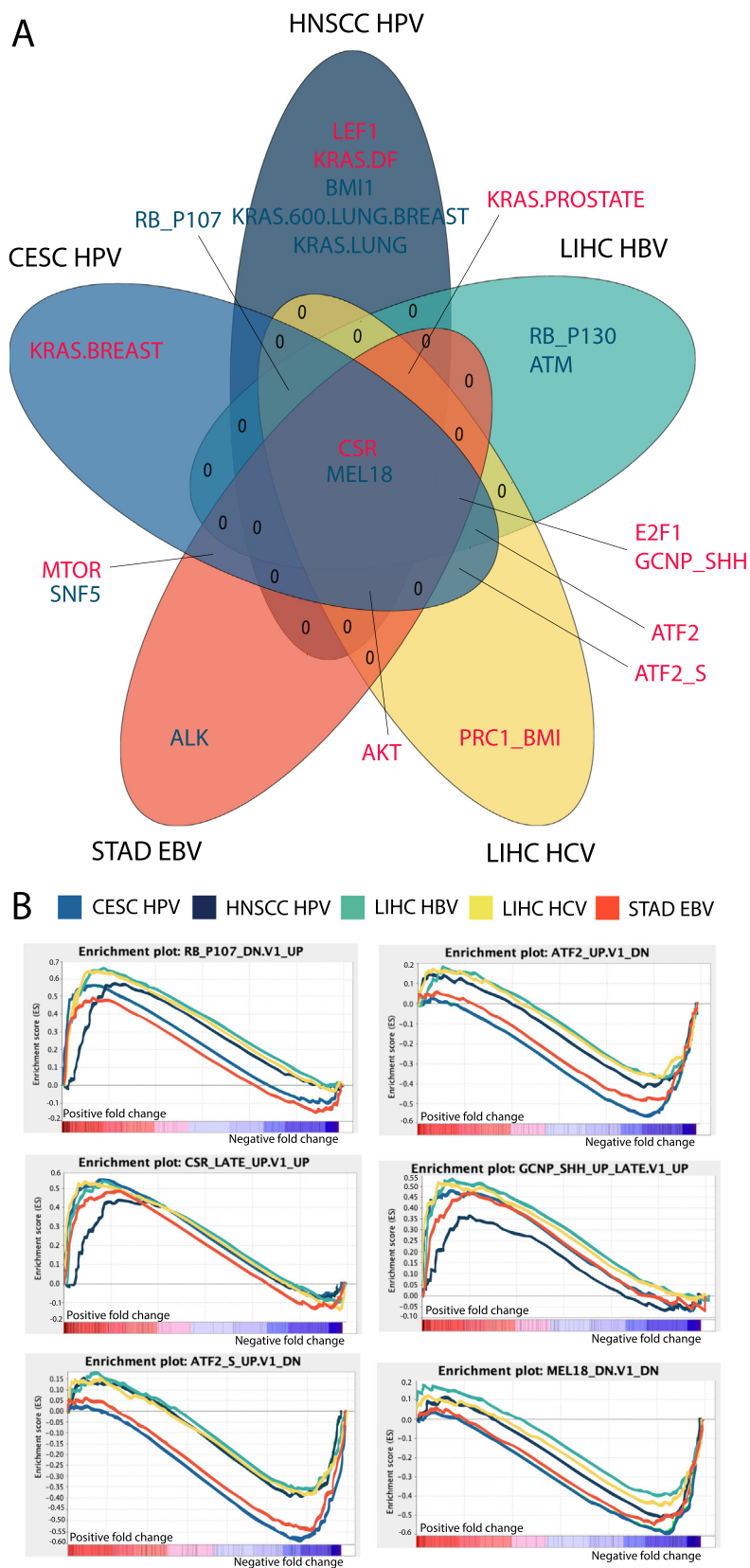**Fig. 5.** Oncogenic signature comparisons. (A) Five-way Venn Diagram comparing oncogenic signatures (C6) most implicated in each cohort following integration with gene-level analysis. (B) Superimposed GSEA plots of six oncogenic signatures that are most similarly enriched across all 5 cohorts.

found to be most dysregulated by GSEA. Canonical pathways found to contain the greatest number of clinically-relevant and related genes are most likely representative of the biological processes implicated in the dysregulation of the clinically-relevant IA genes, and therefore the most clinically-relevant pathways dysregulated in relation to immunologic processes.

Through this analysis, we sought to identify a set of clinically relevant pathways implicated in all five virus-induced cancer cohorts to investigate whether viruses dysregulate similar pathways in different cancers. We found the small cell lung cancer pathway to be the most IA-implicated, clinically important pathway, as evidenced by clinical variable correlation to the expression of individual genes, in all five cohorts, with genes within the pathway being consistently upregulated across the cohorts (Fig. 4A). In other words, the small cell lung cancer pathway is the dysregulated pathway containing the largest number of clinically relevant IA genes and neighboring IA genes. Interestingly, clinically important pathways with IA functions are dysregulated in opposite directions for different cancers. The class A1 rhodopsin-like receptor pathway, which contains chemokine receptors, is strongly downregulated in CESC and LIHC but upregulated in HNSCC and STAD, while the IL12 pathway is upregulated in HNSCC, CESC, and STAD but downregulated in LIHC. Well-known cancer-associated pathways comprise the rest of the top seven most IA-implicated, clinically important pathways that are commonly dysregulated across the cohorts (Fig. 4A). All pathways, except for KEGG small cell lung cancer and REACTOME metabolism of lipids and lipoproteins, are all dysregulated in a similar trend as the IL12 pathway for the different cohorts, especially HNSCC HPV and the two LIHC cohorts (Fig. 4A). Our results suggest that different viruses may dysregulate the same biological processes in opposite directions by mediating IA gene dysregulation.

2.6. Identification of immunologic states mediated by clinically relevant IA dysregulations

We examined the potential effects of IA gene dysregulations on the states of immune cells through immunologic signatures provided by MSigDB. A signature contains differentially expressed genes between different immune cell states or cell types. GSEA was first employed to identify immunologic signatures most significantly implicated in each cohort, then a list of genes reflecting the most clinically-relevant IA dysregulations, including the clinically-relevant IA genes, IA genes surrounding them according to the ReactomeFIViz interaction map, and genes within the most implicated canonical pathways, were searched within the significant immunologic signatures. Signatures with the highest number of genes implicated in clinically relevant IA dysregulations represent immunologic states that viral-induced IA dysregulations may induce.

From the top 10 signatures implicated in each cohort, we noticed that signatures involving Treg cells are heavily implicated with clinically relevant IA dysregulations in all cohorts except for EBV-induced STAD (Fig. 4B). Signatures involving other subtypes of T-cells are mostly implicated in HNSCC, CESC, and HBV-induced LIHC, while signatures involving monocytes and dendritic cells are implicated in EBV-induced STAD (Fig. 4B). Our results suggest that different viruses likely influence different immune processes, but commonalities exist in the immune processes and pathways affected in HPV-induced cancers.

2.7. Identification of cancer-related gene signatures implicated in clinically relevant IA dysregulations

We next examined signatures of cancer-related processes in relation to IA gene dysregulations. Using the same method as we

employed for immunologic signatures, we sought to uncover cancer-related genes most implicated with clinically relevant IA dysregulations. Each signature we examined is a set of gene expression changes from the dysregulation of a CA gene. We discovered that the majority of the top 10 CA genes most strongly associated with IA dysregulation for HPV-induced CESC, HCV-induced LIHC, and EBV-induced STAD tended to be upregulated, while the remaining cohorts had a roughly even split between the number of downregulated and upregulated CA genes (Supplementary Fig. 3A). The downregulated genes are mostly tumor suppressors, while the upregulated genes are mostly oncogenes. Overall, the signatures most in common between all five cohorts included CSR (serum starvation) upregulation and MEL18 downregulation (Fig. 5A). Each cohort had at least one signature that it was uniquely associated with. Out of all five cohorts, HPV-induced HNSCC contained the most uniquely implicated signatures in its top 10 signatures. Other signatures were shared with at least two or more cohorts. Almost all signature correlations are consistent with the role of the cancer-associated genes. For example, signatures associated with oncogenes, such as KRAS, E2F1, and ATF2, are upregulated, while those associated with tumor suppressors, such as RB, are downregulated. However, an exception includes ALK in STAD, which is expected to be upregulated since it is an oncogene.

We also identified the top 6 signatures that are associated with IA dysregulation in all five cohorts (Supplementary Fig. 3B). These signatures include MEL18 and RB downregulation, ATF2 and SHH upregulation, and serum starvation, suggesting that viruses may engage in similar cancer pathways in different cancers to dysregulate IA genes. These cancer-related processes were dysregulated in the same direction across all five cohorts according to the superimposed GSEA plots of all cohorts (Fig. 5B).

2.8. Discovery of mechanisms leading to IA gene dysregulation from genomic alterations

We explored possible mechanisms for genomic alterations in virus-induced cancers to cause dysregulation of clinically relevant IA genes by writing an algorithm that integrates genomic alterations, gene co-expression relationships, and functional mapping of dysregulated IA genes. As input, the algorithm takes clinically relevant IA genes, as identified through IA gene expression correlation with clinical variables. The algorithm then outputs the pathway by which a mutation/CNA can lead to the dysregulation of these IA genes. The pathways are built from genes within clinically relevant IA pathways or genes close to dysregulated IA genes on a functional interaction map (Fig. 4A). Briefly, we correlated IA gene expression to copy number alterations (CNA) and mutations using the REVEALER algorithm ($|CIC| > 0.3$). All genes within significant CNA regions identified by REVEALER as well as significantly mutated genes were examined for coexpression with clinically relevant IA genes, their neighboring genes on a pathway map, and genes within clinically relevant IA pathways (Spearman correlation, $p < 0.05$). Genes most coexpressed with the IA genes are then placed on a functionally annotated map with other dysregulated IA genes, and the shortest path to trace a coexpressed gene to the IA gene through the interaction map is determined to be the most probable mechanism of IA gene dysregulation from genomic alterations. Each pathway begins with a mutation or gene within a CNA region, proceed to genes co-expressed with dysregulated IA genes, and ends with a dysregulated IA gene. Details of the algorithm can be found in the Methods section.

We were able to discover point mutations or genes within CNA regions that have high correlation coefficient (R^2) after Spearman's correlation test for each clinically relevant IA gene (Fig. 6A). Surprisingly, despite the large number of point mutations correlating
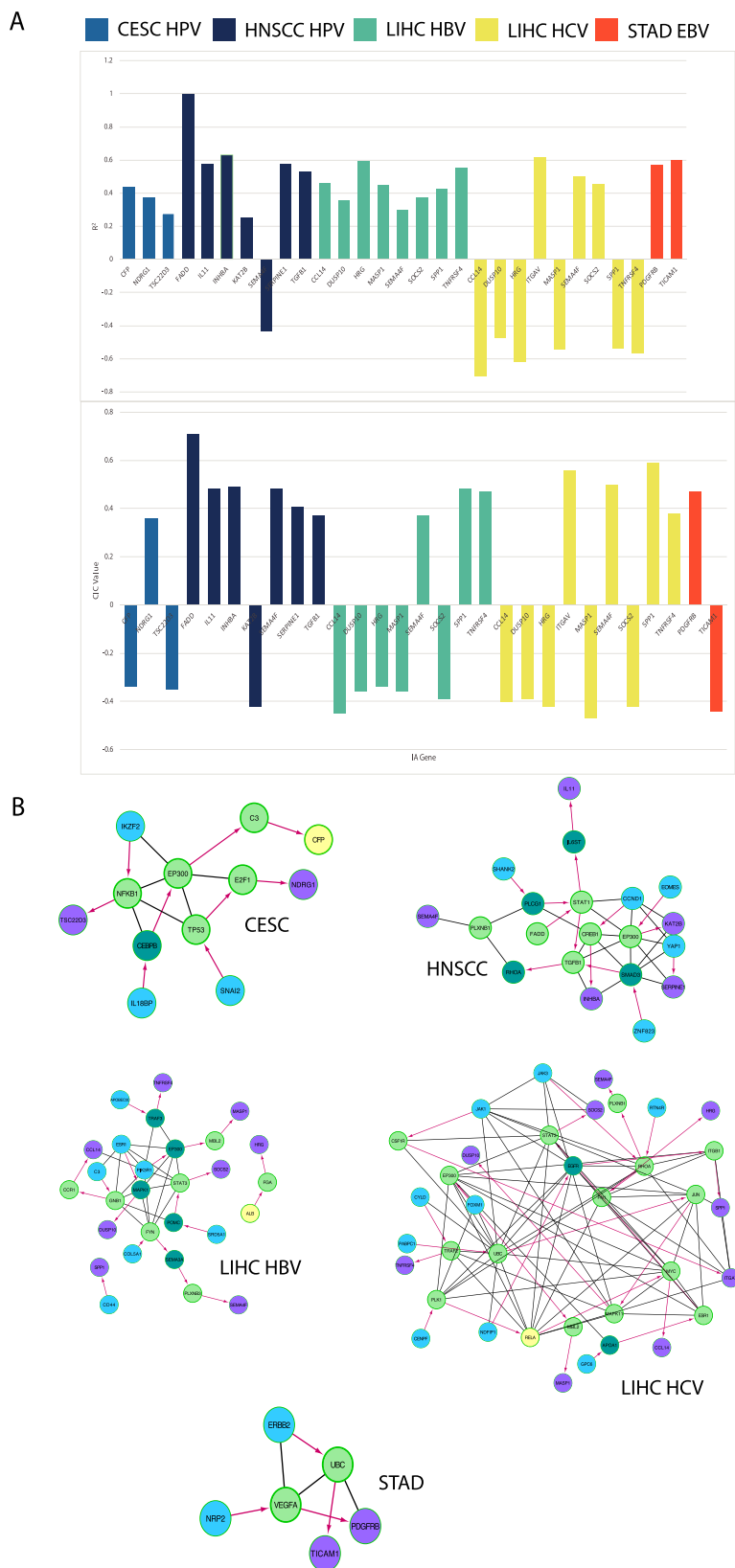
Fig. 6. Genomic alterations correlated with IA gene dysregulation and inferred mechanistic explanations. (A) Bar graph comparison of R^2 and CIC values of genomic alterations associated with cohort-specific IA genes. Only the best R^2 value is displayed if the genomic alteration locus contains multiple genes. (B) Interaction map depicting possible mechanisms of the effects of genomic alterations on the respective IA gene. Green circles indicate the genes related to the cohort-specific IA genes identified by our algorithm, dark cyan circles indicate linker genes, blue circles indicate the starting gene (genomic alteration), yellow circles indicate genes that fall under the same categories as both green and blue circles, and purple circles indicate the unique IA gene. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with IA gene expression, the expression of genes with point mutations generally does not exhibit as high of a correlation with IA gene expression as genes within the CNA regions do. After we mapped the most likely mechanism of IA gene dysregulation using our algorithm, we discovered that the great majority of mechanisms for every IA gene involves a well-known cancer-related gene (Fig. 6B). This observation is consistent with the assumption that cancer-related genomic alterations would lead to dysregulation of a cancer-associated gene. Thus, our mechanisms are plausible. Interestingly, we also observed the same genes participating in the mechanistic pathways in different cohorts. EP300 participates in a mechanism for all cohorts except for STAD, although it is not dysregulated in HBV-induced LIHC. STAT1 or STAT3 participates in a mechanism for HNSCC and both LIHC cohorts.

3. Discussion

In this study, we sought to discover common features between five different types of epithelial cancers that can be induced by viruses, as well as profile differences between them, from an immunologic angle. The pathogenesis mechanism of viruses with respect to cancer have been well-documented and can be grouped into two main categories: indirect and direct mechanisms. Direct mechanisms entail the integration of viral genomes into the host genome or expression of viral oncogenes, while indirect mechanisms entail the induction of chronic inflammation or immunosuppression by the virus, which could both result in cancer development [3].

The different viruses we examined are known to have different mechanism of inducing cancer. EBV produces the viral protein LMP1 and LMP2A, which activates PI3K, STAT, and MAPK [10]. The primary target of infection is B-cells, which could transfer the virus to stomach epithelial cells [11]. There have been arguments that EBV can promote chronic inflammation, but the results are not conclusive [12]. HPV can cause cancer by producing the oncoproteins E6 and E7 to dysregulate key cancer-related genes [13]. However, the vast majority of people with HPV infection do not develop cancer because of immune clearance or immune recognition of cancerous cells [14]. Therefore, how viruses evade immune detection in certain populations is an important question. HCV is known to exert its oncogenic properties mainly through the oncoprotein Core, which dysregulates a complex web of oncogenes, tumor suppressors, ROS production, NF-kB, and many other pathways [3]. Its activities can lead to high levels of inflammation, indirectly leading to cancer [3]. Like HCV, the mechanism of HBV-induced carcinogenesis is also extremely complicated, ranging from inflammation induction, oxidative stress increase, methylation, and the oncoprotein HBx [15]. EBV, HBV, and HPV can all integrate into the host genome and cause insertional mutagenesis, but HCV cannot [16,17]. The complicated mechanisms of most viruses for causing cancer lend a strong case for an in-depth investigation using large sample sizes.

To identify the potential impact of viruses on the immune environment of the tumor, we focused on examining IA genes that are dysregulated in patients with a particular virus, compared to normal samples. The dysregulated IA genes were profiled across different cohorts, and we then examined the mechanistic impact of IA gene dysregulation on biological pathways and immune cells' status. Finally, we examined cancer-associated signatures implicated in IA gene dysregulation and also created an algorithm to identify possible mechanisms for genomic alterations to initiate IA gene dysregulations.

Collectively, our results suggest that virus-induced cancers cause dysregulation of the cancer immune landscape in a variety of ways, as evidenced by the diverse IA pathways and genes found

to be dysregulated. However, we discovered that some of the most prominent dysregulations are within cytokine-cytokine signaling pathways. Genes within cytokine and cytokine receptor signaling are found to be significantly enriched functionally within the list of differentially expressed IA genes in all 5 cohorts. Despite their prevalent dysregulation, the cytokine pathways are not similarly dysregulated in all cancers. In both LIHC cohorts, the majority of genes in the cytokine-cytokine receptor interaction pathway is downregulated. Since this is the case in both HBV and HCV associated LIHC, the downregulation of genes in this pathway could be important to general viral-induced LIHC. The number of genes upregulated and downregulated in this pathway are similar in HNSCC, CESC, and STAD. This pattern of dysregulation could therefore be characteristic of HPV and EBV -induced cancer. What contributes to differences in the dysregulation of this pathway remains to be investigated.

We found that the IL4/IL13 signaling pathway was dysregulated in both HPV-associated CESC and HNSCC. This indicates that the dysregulation of this pathway could be important to HPV-induced cancers. The IL4/IL13 signaling pathway is a significant cytokine pathway known to be associated with inflammation and human cancers [18]. For most genes in this pathway, a high expression of genes was associated with cancer, while a lower expression was associated with normal tissue. This suggests that worse HPV-induced cancer outcomes could results from the upregulation of IL4 /IL13 pathway.

Additionally, the IFN (interferon)-alpha/beta signaling pathways were also dysregulated in HPV-induced cancers. Like IL4/IL13, most genes in the IFN-alpha/beta pathway with high expression were associated with cancer tissue, while a lower expression was associated with normal tissue. IFN alpha and beta are interferons significantly associated with viral infections [19], and are usually seen at elevated levels during some viral infections [20]. Thus, the upregulation of the IFN-alpha/beta pathways can be expected in virus-induced cancer tissue.

Lastly, class A1-rhodopsin receptors, which include chemokine receptors, and the IL12 signaling pathways are some of the top pathways that are clinically relevant across all five cohorts. The A1-rhodopsin receptors are downregulated in CESC and LIHC, but upregulated in HNSCC and STAD. This is interesting as the pathway is oppositely dysregulated in the CESC and HNSCC cohorts despite both being HPV-induced. This could therefore represent a difference between the impact of HPV in the two cancers. The IL12 pathway is upregulated in HNSCC, CESC, and STAD but downregulated in LIHC. This could represent a difference between viral-induced LIHC and other viral-induced cancers.

We also found that while differences exist between the different viral-induced cancers, important similarities exist, especially between HPV-induced cancers and LIHC cancers. These two groups of cohorts exhibit similar dysregulation trends that are opposite of each other for clinically significant IA pathways. There are also similarities between all cohorts. All or most groups consistently correlate with the cancer processes of CSR upregulation, KRAS upregulation, AKT upregulation, RB downregulation, ATM downregulation, and MEL18 downregulation. Furthermore, the pathways inferred by our algorithm frequently contain the same genes, such as EP300.

Using our newly developed framework, we were able to leverage large-scale transcriptomic and genomic data to trace the pathway of how virus-induced cancers interact with the immune system. Starting from genomic alterations, we identified possible mechanisms for dysregulation of IA genes, which we then implicated to be involved in the dysregulation of key IA pathways, processes, and cells. We believe that our mechanistic analyses identified key pathways to test *in vitro* and *in vivo* for investigation of virus-induced immune dysfunction, which may lead to the dis-

covery of novel intervention points for immunotherapy treatments of virus-induced cancers.

4. Conclusion

In summary, we incorporated gene-level analysis with pathways analysis to establish a novel gene dysregulation process for analyzing immune-associated dysregulation in five virus-induced cancers. We used TCGA data from HNSCC HPV+, CESC HPV+, LIHC HBV+, LIHC HCV+ and STAD EBV+ patients to uncover significantly dysregulated genes and correlated them to patient survival and clinical variables. In the next step, we used these sets of genes to find clinically relevant immune and cancer pathways within and across cohorts. Lastly, we elucidated potential mechanisms through which viruses induce gene dysregulation via genomic alterations using our own algorithm.

Cytokines are immunomodulating proteins, and our analyses found many pathways associated with cytokines to be significant across the cohorts. Interferon-alpha/beta pathways and IL4/IL13 pathways are dysregulated in HPV-associated cancers. The cytokine-cytokine receptor interaction pathway and interleukin pathways were upregulated in HPV-associated cancers and EBV-associated stomach adenocarcinoma, but not in hepatitis-induced liver cancers. Across all five cancers, class A1-rhodopsin receptors, which contain chemokine receptors, were among the most clinically significant pathways, although it is dysregulated differently in different cancers. Additionally, our algorithm found a potential mechanism of virus-associated gene dysregulation through the EP300 oncogene.

Our results suggest that viruses dysregulate the cancer immune landscape most prominently through cytokine signaling, although different viruses dysregulate this pathway differently. By integrating functional annotations and large-scale sequencing data using a novel algorithm, we provide a new approach for discovery of cancer pathogenesis mechanisms.

5. Methods

5.1. The cancer genome Atlas (TCGA) RNA-sequencing datasets and cohort designation

Level 3-normalized mRNA expression read counts for tumor samples from 89 HNSCC HPV+ patients, 228 CESC HPV+ patients, 74 LIHC HBV+ patients, 30 LIHC HCV+ patients and 23 STAD EBV + patients along with their adjacent normal tissue were downloaded from TCGA (https://tcga-data.nci.nih.gov/tcga). Data on patient disease progression, staging, and vital status recorded over this period were used in later analysis.

5.2. Differential expression analysis to identify dysregulated immune-associated genes

Using edgeR v3.5 (http://www.bioconductor.org/packages/release/bioc/html/edgeR.html), mRNA read count inputs were filtered, resulting in the removal of lowly expressed mRNAs (counts-per-million < 1 when comparing samples from the larger group to those of the smaller group in a cohort) from the analysis. Trimmed mean of M-values (TMM) were normalized and pairs of mRNAs were designated to identify those that were significantly differentially expressed when comparing one cohort to another. mRNAs considered to be significantly dysregulated were those with fold change> 2 or <−2 and false discovery rate (FDR) < 0.05, output by the edgeR analysis. After filtering for dysregulated genes, potential candidates were retained if they were considered to be

immune or cancer associated. A list of immune associated genes was obtained from ImmPort [21] and InnateDB [22].

5.3. Reactome FIViz clustering of genes by function

Reactome FIViz, a plugin for the Cytoscape gene pathways visualization software [23], was used to draw gene-gene interaction diagrams based on pathways interaction databases. The software was also used to find genes next to the IA gene of interest on the interaction diagram, termed neighboring genes.

5.4. Correlating gene expression with survival

Survival analyses were performed using the Kaplan–Meier Model, with gene expression designated as a binary variable based on expression above or below the median expression of all samples. Univariate Cox regression analysis was used to identify candidates that were significantly associated with patient survival ($p < 0.05$).

5.5. Correlating gene expression with clinical variables

Clinical significance of immune-associated genes that were correlated with patient survival was determined by employing the Kruskal-Wallis test. Gene expression values were correlated with variables including pathologic stage, pathologic TNM stage, residual tumor, neoplasm cancer status, neoplasm histologic grade and presence of perineurial invasion. All genes, separated by viral cohorts, with two or more clinical variable correlations and correlation with survival are retained for further analysis and termed clinically relevant cohort-specific IA genes.

5.6. Identification of pathways and signatures dysregulated in each cohort using GSEA

For each cancer cohort, Gene Set Enrichment Analysis (GSEA) [6] was used to find canonical pathways, immunologic signatures, and cancer-associated (CA) gene activation signatures dysregulated in cancer vs. normal samples. The above gene sets are derived from the Molecular Signatures Database (MSigDB) [24], under the gene set IDs C2, C7, and C6, respectively. The GSEA was run using the pre-ranked setting, which takes a ranked gene list over the entire transcriptome as input. The ranked gene list is a list of all genes ordered by differential expression fold change, from the gene with the most positive fold change to that with the most negative fold change (cancer vs. normal samples). This analysis allows us to examine pathways dysregulation by taking all expression data into account, which would avoid the shortcomings of single-gene analysis, such as arbitrary significance cutoffs leading to significant genes being overlooked.

5.7. Identification of clinically relevant IA pathways dysregulated within each cohort

Within each of our 5 virus-induced cancer cohorts, we separately identified clinically relevant IA pathways by searching for canonical pathways (C2) that are immune-associated and contain the greatest number of clinically relevant cohort-specific IA genes and genes that surround them in a ReactomeFIViz graph (neighboring genes). The ReactomeFIViz graph was produced with only significantly dysregulated IA genes (exact test from differential expression, p < 0.05). Only canonical pathways that are dysregulated after GSEA analysis are included in our analysis (p < 0.05). This procedure was used to effectively filter for the most clinically relevant IA pathways and pathways that are most relevant to the mechanism of viral-induced cancer.

5.8. Identification of immune states most implicated by dysregulated IA genes

Immune states, or immunologic phenotypes defined by pathways activation or immune cell population changes, most implicated by the clinically relevant dysregulated IA genes were identified by searching for significant immunologic signatures (C7) dysregulated in each cohort (from GSEA, $p < 0.05$) containing the greatest number of relevant genes. These relevant genes include the original clinically relevant cohort-specific IA genes, their neighboring genes on the ReactomeFIViz plot, and genes within the clinically relevant IA pathways identified above.

5.9. Identification of cancer gene dysregulations most implicated by dysregulated IA genes

Cancer gene dysregulation phenotypes are captured by oncogenic signatures (C6) in MSigDB. These signatures contain the top 200 genes that are upregulated or downregulated following the overexpression of knockdown of a cancer-related gene. We identified the cancer gene activation/suppression most implicated by the dysregulation of IA genes in each cohort by searching for significantly dysregulated oncogenic signatures containing the greatest number of relevant genes, which include the same genes as those used to identify immune states that were most implicated.

5.10. Correlating gene expression with genomic alterations

Copy number alteration (CNA) and mutation data were obtained from annotation files generated by the BROAD Institute GDAC Firehose on March 31, 2018. Quantification of mutation presence were analyzed by calculating the percentage of patients with each mutation, indicated by a binary value per mutation. The GDAC files were compiled into input files for the REVEALER (repeated evaluation of variables conditional entropy and redundancy) algorithm, which identifies sets of specific CNAs and mutations that are most likely implicated in changes to the target expression profile. The target profile was identified as the expression of a single CA or IA gene. The REVEALER algorithm runs in multiple iterations in order to identify the most prominent genomic alterations. For our study, we set the maximum number of iterations to three. The algorithm also allows for the use of a seed, or a particular mutation of CNA gain or loss event that may account for target activity. However, because we did not know the individual genetic alterations that were responsible for each genes' dysregulation, the seed was set to null. Significant association between genomic alteration and gene expression was determined by conditional information coefficient (CIC) > 0.25 and p-value <0.05.

5.11. Development of an algorithm to identify the mechanism of IA gene dysregulation through genomic alterations

We attempted to trace the most plausible pathway through which genomic alterations can lead to IA gene dysregulation. The algorithm integrates REVEALER output, whole transcriptome data, and a list of genes of interest, including clinically relevant IA genes, neighboring genes, and genes within clinically relevant IA pathways. The REVEALER output contains the genomic regions or mutations most strongly associated with the dysregulation of IA genes. The algorithm first takes significant CNA loci and procures a list of all genes within the reported loci. These genes were then correlated to the clinically relevant IA gene(s) associated with the CNA (Spearman). We hypothesize that genes within the CNA loci were amplified or deleted to cause the upregulation or downregulation of IA genes. Therefore, only loci genes with correlations consistent with our hypothesis are kept for the next step. After this cause-effect hypothesis was established, we proceeded to find a pathway between the potentially causal genes within the CNA locus and the IA gene by identifying genes in between these two genes within a gene-gene interaction network. This interaction network was constructed using Cytoscape, and the shortest distance between these two genes is found by an unweighted breadth-first search, using the R package igraph (igraph.org). For this study, we found the top 10 potential causal genes with the most significant correlation with the expression of each target IA gene, given that $p < 0.05$ (Spearman), and then calculated the shortest path between each of these 10 genes and the target IA gene. The gene out of these top 10 with the shortest path overall was determined to be the most likely causal gene. For mutations, the same procedure was performed, except that unlike CNAs, no correlation direction matching will need to be performed between a mutation and an IA gene. The code for this algorithm is available at https://github.com/har-li/IntegrativePathwayDiscoveryTool/.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the current study are available in the GDC Data Portal (https://portal.gdc.cancer.gov/).

CRediT authorship contribution statement

Lindsay M. Wong: Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Wei Tse Li:** Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Neil Shende:** Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Joseph C. Tsai:** Validation, Formal analysis, Investigation, Data curation, Writing – review & editing, Visualization. **Jiayan Ma:** Software, Validation, Formal analysis, Investigation, Data curation, Writing – review & editing. **Jaideep Chakladar:** Validation, Formal analysis, Investigation, Data curation, Writing – review & editing. **Aditi Gnanasekar:** Formal analysis, Investigation, Data curation. **Yuanhao Qu:** Formal analysis, Investigation, Data curation. **Kypros Dereschuk:** Investigation, Visualization. **Jessica Wang-Rodriguez:** Writing – review & editing, Funding acquisition. **Weg M. Ongkeko:** Conceptualization, Methodology, Investigation, Resources, Data curation, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.11.013.

References

[1] McLaughlin-Drubin ME, Munger K. Viruses associated with human cancer. Biochim Biophys Acta 2008;1782(3):127–50.
[2] zur Hausen H. Viruses in human cancers. Science 1991;254(5035):1167–73.
[3] Morales-Sanchez A, Fuentes-Panana EM. Human viruses and cancer. Viruses 2014;6(10):4047–79.
[4] Simillion C, Liechti R, Lischer HEL, Ioannidis V, Bruggmann R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinf 2017;18 (1):151.
[5] Reimand Jüri, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc 2019;14 (2):482–517.
[6] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102(43):15545–50.
[7] Tamayo P, Steinhardt G, Liberzon A, Mesirov JP. The limitations of simple gene set enrichment analysis assuming gene independence. Stat Methods Med Res 2016;25(1):472–87.
[8] Gatti DM et al. Heading down the wrong pathway: on the influence of correlation within gene sets. BMC Genomics 2010;11:574.
[9] Loo Y-M, Gale M. Immune signaling by RIG-I-like receptors. Immunity 2011;34 (5):680–92.
[10] Shair KHY, Schnegg CI, Raab-Traub N. EBV latent membrane protein 1 effects on plakoglobin, cell growth, and migration. Cancer Res 2008;68 (17):6997–7005.
[11] Iizasa H, Nanbo A, Nishikawa J, Jinushi M, Yoshiyama H. Epstein-Barr Virus (EBV)-associated gastric carcinoma. Viruses 2012;4(12):3420–39.
[12] Morales-Sanchez A, Fuentes-Panana EM. Epstein-Barr virus-associated gastric cancer and potential mechanisms of oncogenesis. Curr Cancer Drug Targets 2017;17(6):534–54.
[13] Ghittoni R, Accardi R, Hasan U, Gheit T, Sylla B, Tommasino M. The biological properties of E6 and E7 oncoproteins from human papillomaviruses. Virus Genes 2010;40(1):1–13.
[14] Conesa-Zamora P. Immune responses against virus and tumor in cervical carcinogenesis: treatment strategies for avoiding the HPV-induced immune escape. Gynecol Oncol 2013;131(2):480–8.
[15] Tarocchi M et al. Molecular mechanism of hepatitis B virus-induced hepatocarcinogenesis. World J Gastroenterol 2014;20(33):11630–40.
[16] Pistello M, Antonelli G. Integration of the viral genome into the host cell genome: a double-edged sword. Clin Microbiol Infect 2016;22(4):296–8.
[17] Xiao K, Yu Z, Li X, Li X, Tang K, Tu C, et al. Genome-wide analysis of Epstein-Barr virus (EBV) integration and strain in C666-1 and Raji cells. J Cancer 2016;7 (2):214–24.
[18] McCormick SM, Heller NM. Commentary: IL-4 and IL-13 receptors and signaling. Cytokine 2015;75(1):38–50.
[19] Samuel CE, *Antiviral actions of interferons.* Clin Microbiol Rev, 2001: 14(4); 778-809, table of contents.
[20] Biron CA. Interferons alpha and beta as immune regulators–a new look. Immunity 2001;14(6):661–4.
[21] Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, et al. ImmPort: disseminating data to the public for the future of immunology. Immunol Res 2014;58(2-3):234–9.
[22] Lynn DJ et al. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Mol Syst Biol 2008;4:218.
[23] Wu G et al. ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. F1000Research 2014;3:146.
[24] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27 (12):1739–40.