Jha *et al. Genome Biology* (2022) 23:117
https://doi.org/10.1186/s13059-022-02681-3

Genome Biology

RESEARCH Open Access

Identifying common transcriptome signatures of cancer by interpreting deep learning models

Anupama Jha[1*†], Mathieu Quesnel-Vallières[2,3*†], David Wang[2], Andrei Thomas-Tikhonenko[4,5,6], Kristen W Lynch[3] and Yoseph Barash[1,2*]

†Anupama Jha and Mathieu Quesnel-Vallières contributed equally to this work.
*Correspondence:
anupamaj@seas.upenn.edu
mathieu.quesnel-vallieres@pennmedicine.upenn.edu
yosephb@pennmedicine.upenn.edu
[1]Department of Computer and Information Science, School of Engineering and Applied Science, Philadelphia, USA
[2]Department of Genetics, Philadelphia, USA
Full list of author information is available at the end of the article

Abstract

Background: Cancer is a set of diseases characterized by unchecked cell proliferation and invasion of surrounding tissues. The many genes that have been genetically associated with cancer or shown to directly contribute to oncogenesis vary widely between tumor types, but common gene signatures that relate to core cancer pathways have also been identified. It is not clear, however, whether there exist additional sets of genes or transcriptomic features that are less well known in cancer biology but that are also commonly deregulated across several cancer types.

Results: Here, we agnostically identify transcriptomic features that are commonly shared between cancer types using 13,461 RNA-seq samples from 19 normal tissue types and 18 solid tumor types to train three feed-forward neural networks, based either on protein-coding gene expression, lncRNA expression, or splice junction use, to distinguish between normal and tumor samples. All three models recognize transcriptome signatures that are consistent across tumors. Analysis of attribution values extracted from our models reveals that genes that are commonly altered in cancer by expression or splicing variations are under strong evolutionary and selective constraints. Importantly, we find that genes composing our cancer transcriptome signatures are not frequently affected by mutations or genomic alterations and that their functions differ widely from the genes genetically associated with cancer.

Conclusions: Our results highlighted that deregulation of RNA-processing genes and aberrant splicing are pervasive features on which core cancer pathways might converge across a large array of solid tumor types.

Keywords: Cancer genomics, Transcriptomics, Deep learning

Background

Cancer is a loosely defined term that designates cells that have acquired pathological properties, mainly loss of cell cycle regulation, high proliferation rate, and loss of contact inhibition leading to invasion of surrounding tissues. In time, tumor cells disrupt the normal function of tissues where they are located and can metastasize to other tissues. Oncogenes contribute to cell transformation while tumor suppressor genes stop aberrant cell proliferation. Changes in the expression, activation, or function of these genes are expected to lead to cancer-like phenotype in various cell or tissue types and many such genes are commonly affected by genomic lesions in cancer. In addition to mutations to hallmark oncogenes and tumor suppressor genes, cancer driver mutations that contribute to disease onset and progression are found in subsets of cancer types [1]. While these genetic alterations are diverse, several genes that are altered in cancer converge on a few molecular mechanisms that are commonly involved in tumorigenesis [2]. These pathways have wide-ranging effects that span the cell cycle, inflammation, and apoptosis, among others. The mechanisms through which they operate in cancer are therefore highly diverse and molecularly heterogeneous, but they are also interrelated. In addition, a recent gene network analysis identified a relatively small number of regulatory modules on which a majority of somatic mutations in cancer converge [3]. Because changes in cellular pathways and biological activity ultimately impact gene expression and post-transcriptional regulation, this leaves the possibility that tumors that arise from the disruption of different pathways share common molecular signatures in the form of transcriptomic variations.

Previous studies have attempted to leverage these projected common signatures of cancer in order train computational models to distinguish tumors from normal samples or distinguish different tumor types. Typically, these studies rely on protein-coding gene expression data combined with deep neural networks or other machine learning algorithms that classify samples into two or more categories [4–10]. These studies showed that machine learning models can successfully distinguish between normal tissues and tumors given a certain set of conditions, including the pre-selection of biological features before model training. Several automatic feature selection methods exist to lower the number of genes used as input and thus facilitate the training of such machine learning models [10–16]. However, pre-selecting genes on the basis of their functions or differential expression in cancer, or removing redundant genes identified by automatic selection prior to model training deprives the models from learning about potentially novel genes contributing to the transcriptomic signature of cancer. In addition, the application of such approaches has not been tested on large heterogeneous sets of tissues.

Recent methods for the interpretation of deep neural networks offer the opportunity to agnostically discover transcriptomic variations characterizing cancer biology from models that successfully predict biological classes [17, 18]. In particular, we recently described enhanced integrated gradients (EIG), a method for deep neural network interpretation [19] that generates attribution values as a measure of the weight or importance of each biological input feature in the model. For example, we used EIG to find splicing events that are differentially included in the brain compared to other tissues without prior knowledge of splicing variations [19].

Here, we aimed to draft a molecular profile of cancer that applies to most solid tumor types by leveraging the predictive power of deep neural networks along with the interpretation capability of enhanced integrated gradients to identify common transcriptomic

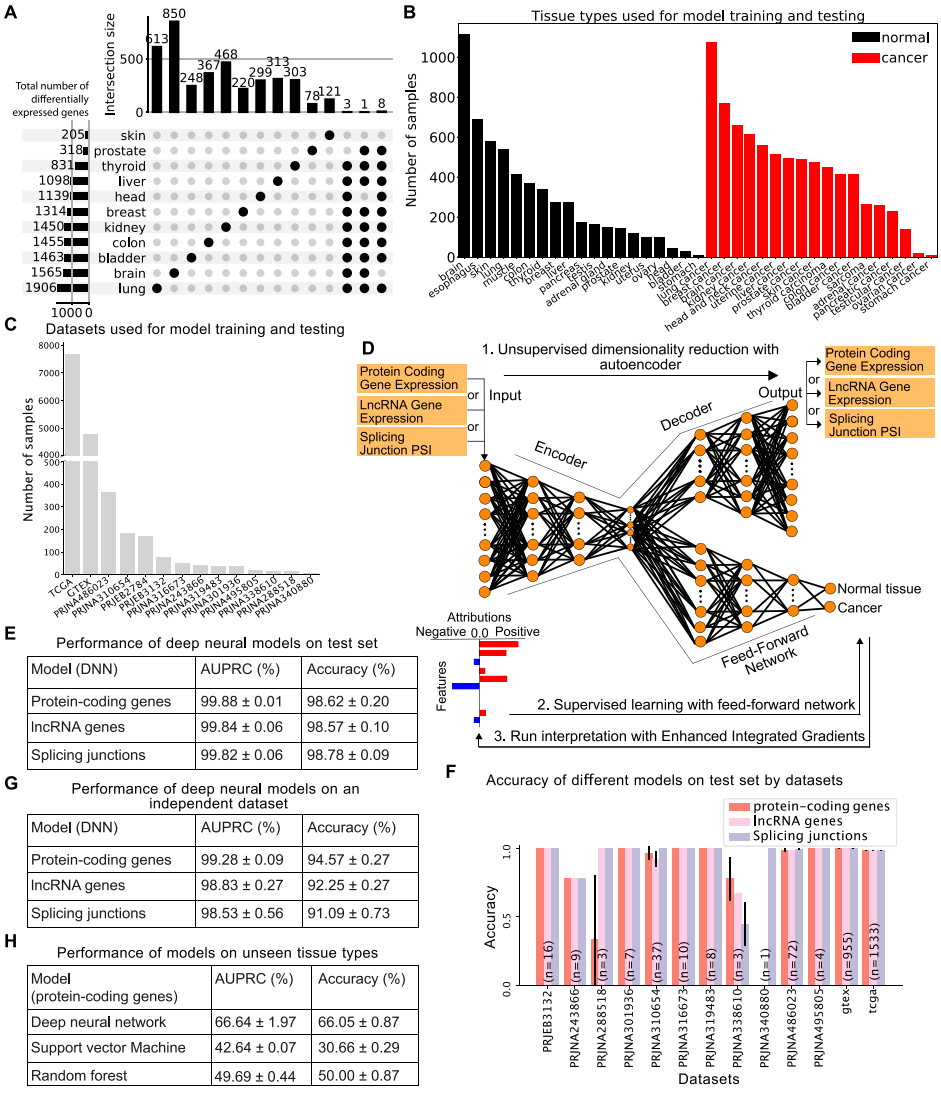
signatures across a large array of tumor types. We trained feed-forward neural networks with protein-coding gene expression, lncRNA gene expression or splice junction usage data from several normal tissue and tumor types. We then derive attribution values from these models and establish a list of high-attribution features corresponding to a common signature of cancer, which could be causally involved in cancer or result from oncogenic transformation, or both. Finally, we assess the biological functions of these transcriptomic variations.

Results

A feed forward neural network trained with protein-coding gene expression distinguishes between normal and cancer tissues

We aimed to uncover the transcriptomic features that commonly define cancer state. Performing differential gene expression analysis on 11 normal tissue-tumor pairs from GTEx and TCGA and then looking at the overlap in the genes that are deregulated between these analyses show that few protein-coding genes are consistently up- or downregulated (abs(log2FC) $>$ 2, adjusted p-value $<$ 0.01) in six or more tumor types and that none is consistently deregulated in more than nine tumor types (Fig. 1A and Additional file 1: Fig. S1A). Instead, a large fraction of cancer-deregulated genes are specific to a single tumor type (Fig. 1A and Additional file 1: Fig. S1B). In addition, while hallmark oncogenes are expected to be disrupted in many cancer types, we observe in a sampling of 11 oncogenes that these are either not significantly differentially expressed in any of the 11 tumors analyzed (e.g., *BCR*, *CTNNB1*, *DDX6*, *FUS*, *KRAS*, *MDM2*, *TPR*) or only disrupted in certain tumors (e.g., *EGFR*, *ETV4*, *JUN*, *MYC*; Additional file 1: Fig. S1C). Such apparent inconsistency between the function of oncogenes and their lack of change in expression in many tumor types can be partially explained by alternative mechanisms of activation that are independent of changes in transcript levels. Nonetheless, these results demonstrate that using a simple differential gene expression analysis fails to capture the complexity and heterogeneity of the transcriptomic variations existing across various cancer types.

In order to overcome the limitations of such naive searches for common cancer transcriptome signatures, we sought to train interpretable deep learning models capable of distinguishing between normal and cancer samples. We assembled a large RNA-Seq dataset comprising 13,461 samples from 19 normal tissue types and 18 tumor types and split the data into two classes reflecting cancer state: normal or tumor (Fig. 1B; 5622 total normal samples and 7839 total tumor samples). Samples were sourced from TCGA (https://www.cancer.gov/tcga), GTEx [20] and 12 other datasets (Fig. 1C). Because technical biases and batch effects are a major concern when using large-scale RNA-Seq datasets, especially when comparing perfectly confounded datasets like GTEx and TCGA, we included in our compendium 12 smaller datasets containing either only tumor samples or tumor and matched normal tissue samples from the same donors. These additional datasets allowed us to mitigate dataset-specific biases and focus on cancer-specific signals by performing a tissue/tumor-specific mean correction across the 14 datasets (see Additional file 1: Fig. S2A-B). We also considered alternatives to mean correction, such as the commonly used COMBAT method [21], but this approach severely limited the data and gene sets that could be used for model training (see the "Batch correction" section in the "Methods" section for details).



Model (DNN)	AUPRC (%)	Accuracy (%)
Protein-coding genes	99.88 ± 0.01	98.62 ± 0.20
lncRNA genes	99.84 ± 0.06	98.57 ± 0.10
Splicing junctions	99.82 ± 0.06	98.78 ± 0.09

Model (DNN)	AUPRC (%)	Accuracy (%)
Protein-coding genes	99.28 ± 0.09	94.57 ± 0.27
lncRNA genes	98.83 ± 0.27	92.25 ± 0.27
Splicing junctions	98.53 ± 0.56	91.09 ± 0.73

Model (protein-coding genes)	AUPRC (%)	Accuracy (%)
Deep neural network	66.64 ± 1.97	66.05 ± 0.87
Support vector Machine	42.64 ± 0.07	30.66 ± 0.29
Random forest	49.69 ± 0.44	50.00 ± 0.87

Fig. 1 **A** Upset plot summarizing pairwise differential gene expression analyses performed on tumors and their corresponding normal tissue. No gene is significantly deregulated in more than 9 out of 11 cancer types tested. **B**, **C** Dataset assembled to train and test binary models to distinguish between normal and tumor samples shown by tissue type (**B**) or dataset (**C**). **D** Graphical representation of the computational framework used to train, test, and interpret the models. **E** Performance of models trained with protein-coding gene expression, lncRNA gene expression, or splicing variations evaluated by area under the precision-recall curve (AUPRC) and accuracy (sum of true positives and true negatives over the total population). **F** Accuracy of models trained with protein-coding gene expression, lncRNA gene expression, or splice junction usage across the 13 datasets used to assemble the training set. **G** Performance of models trained with protein-coding gene expression, lncRNA gene expression, or splice junction usage on an independent dataset consisting of normal and cancer lung samples. **H** Performance of the deep learning model, SVM, and random forest using protein-coding gene expression on unseen tissue types (blood cancers) with no batch correction. The training set consists of solid tumors only

We first used the mean corrected expression data from 19,657 protein-coding genes to train an autoencoder for dimensionality reduction, followed by a supervised deep neural network to predict cancer state (normal tissue versus tumor, Fig. 1D; see the "Methods" section for details about dimensionality reduction and model training). We divided our

dataset into training (8504 samples), validation (2127 samples), and test sets (2658 samples). We tuned model hyperparameters (learning rate, hidden layers, number of nodes, activation functions, and dropout probability) on the validation set and fixed our model architecture using the hyperparameters with the best performance on the validation set (see Additional file 2: Table S1 for the final model architecture of the deep neural network model for the protein-coding genes). To ensure that our model did not learn dataset-specific biases, we evaluated model performance on a previously unseen set of samples (test set with 2658 samples) extracted from the 13 datasets used for training as well as one independent dataset (PRJEB2784) that was not used during training but that comprises 172 samples of tissue types that were included in the training set (normal and tumor lung samples). Our protein-coding gene expression model accurately predicted whether an individual sample corresponds to a normal tissue or a tumor (accuracy 98.62% $\pm$ 0.20% and area under the precision-recall curve (AUPRC) 99.88% $\pm$ 0.01%, Fig. 1E and Additional file 1: Fig. S3), and this performance generalizes over all 13 datasets despite the dataset imbalance (Fig. 1F, see numbers in the bars for the number of samples in each dataset). The only datasets where the performance is variable have very few samples (PRJNA340880 has 1 sample and PRJNA288518 has 3 samples). Importantly, the model performs almost as well when applied on the independent dataset (Fig. 1G and Additional file 1: Fig. S4).

To evaluate how our model generalizes to cancer types not included in the training set, we assembled a group of normal (macrophages, monocytes, and lymphocytes) and malignant (acute myeloid leukemia and acute lymphoblastic leukemia) blood cells from three datasets (ArrayExpress E-MTAB-2319, Blueprint, and TARGET consortia, see Additional file 2: Table S2 for details on the samples). We omitted batch correction on these samples to assess how our model would fare on a dataset that considerably differs from the training set both at the biological (solid tissues and tumors in training set vs. hematologic cells and tumors in the test set) and technical levels (batch-corrected in training set vs. uncorrected test set). Strikingly, despite these significant differences between training and test sets, our deep neural network model successfully distinguishes normal and cancer samples from blood (Fig. 1H and Additional file 1: Fig. S5), although, as expected, we observe a reduction in accuracy.

Finally, in order to assess how our deep neural network model compares to other machine learning algorithms, we first trained support vector machine and random forest models using the same training set as for our deep neural network model and then tested them on the same independent dataset consisting of batch-corrected normal and cancer lung samples. Under these conditions, all three models perform similarly (Additional file 2: Table S3). However, in sharp contrast with the deep neural network model, both the support vector machine and random forest models completely fail when applied to the hematologic dataset (Fig. 1H). In summary, these results demonstrate that our deep neural network model can more accurately and robustly identify cancer samples compared to commonly used machine-learning methods and motivate the subsequent analysis of the associated features.

lncRNA expression or splice site usage profiles suffice to define cancer state

Other types of transcriptomic features, including lncRNA expression and RNA splicing, have been used as prognostic markers or to predict drug response in cancer [22–24]. In

addition, a small number of mutations located in lncRNA genes or disrupting splicing in protein-coding genes have been shown to drive cancer [25]. However, it is not known whether widespread changes in lncRNA expression or RNA splicing commonly characterize cancer state. We thus asked if these other types of transcriptomic features could be used to distinguish between normal and tumor samples, similar to what we found for protein-coding gene expression.

We used the same strategy as for the model trained with protein-coding gene expression above and trained models with expression data from 14,257 lncRNA genes or splice site usage data from 40,147 splice junctions. Similar to protein-coding genes, we tuned the hyperparameters for deep neural network models using lncRNA and splicing junctions on the validation set and evaluated model generalization performance on the test set (see Additional file 2: Tables S4 and S5 for the final architectures of the lncRNA and splice junction deep neural networks, respectively). Remarkably, these models achieved 98.57% $\pm$ 0.1% and 98.78% $\pm$ 0.09% accuracy, respectively, with high AUPRC (99.84% $\pm$ 0.06% for lncRNA expression and 99.82% $\pm$ 0.06% for splice junction usage, Fig. 1E). As observed with the protein-coding gene expression-trained model, the lncRNA gene expression and the splice junction usage-trained models perform consistently well across all of the test datasets on the task of predicting the cancer state, again despite the dataset imbalance (Fig. 1F), or when tested on an independent dataset (Fig. 1G), These results further support the robustness of our models as capable of identifying true biological signals rather than confounders.

As for the protein-coding genes model, we compared our deep neural network model for lncRNA gene expression with support vector machine and random forest models on the same two datasets. All three models performed well on the batch-corrected dataset consisting of normal and tumor lung samples (Additional file 2: Table S3B). However, while our deep neural network model also correctly predicted cancer state with the hematologic dataset (normal leukocyte and leukemia samples), both support vector machine and random forest models completely failed again on this dataset (Additional file 2: Table S6). On the splicing data, our deep neural network model outperforms both the support vector machine and random forest models on the independent batch-corrected normal and tumor lung dataset (Additional file 2: Table S3C). The strong performance of our lncRNA- and splicing-trained models indicates that tumor samples can be defined not only by their protein-coding gene expression profile, but also using exclusively their lncRNA gene expression or splice junction usage profile.

Interpretation of deep learning networks uncovers new transcriptomic features characterizing cancer state

Given the high performance of our models, we wanted to know what transcriptomic features are the most important in each of our models and whether these features consist mostly of the usual suspects, i.e., genes known to be genetically associated with cancer. To do this, we generated feature importance scores known as attribution values for tumor samples using enhanced integrated gradients (EIG) [19]. Briefly, EIG measures a feature's contribution, either positive or negative, to the model label predictions (normal tissue versus cancer) when comparing a cancer sample to a baseline. Following our previous work [19], we used the median of normal samples as the baseline (see the "Methods" and "Interpretation of tumor classification models" sections for details).

We selected 1768 protein-coding genes, 1763 lncRNAs and 562 splice junctions that have high median attribution values across tumor types (Fig. 2A and Additional file 2: Tables S7-S9; see the "Methods" and "Selection of feature sets" sections for the selection criteria and Additional file 1: Fig. S6). We also defined "neutral" sets with a sample size equivalent to sets of high-attribution features using features that display attribution values close to zero (Additional file 2: Tables S10-S12). When looking at the cancer type-specific attribution values across 14 tumor types for the top 100 features with positive or negative attribution, we found that protein-coding genes, lncRNAs and splice junctions with the

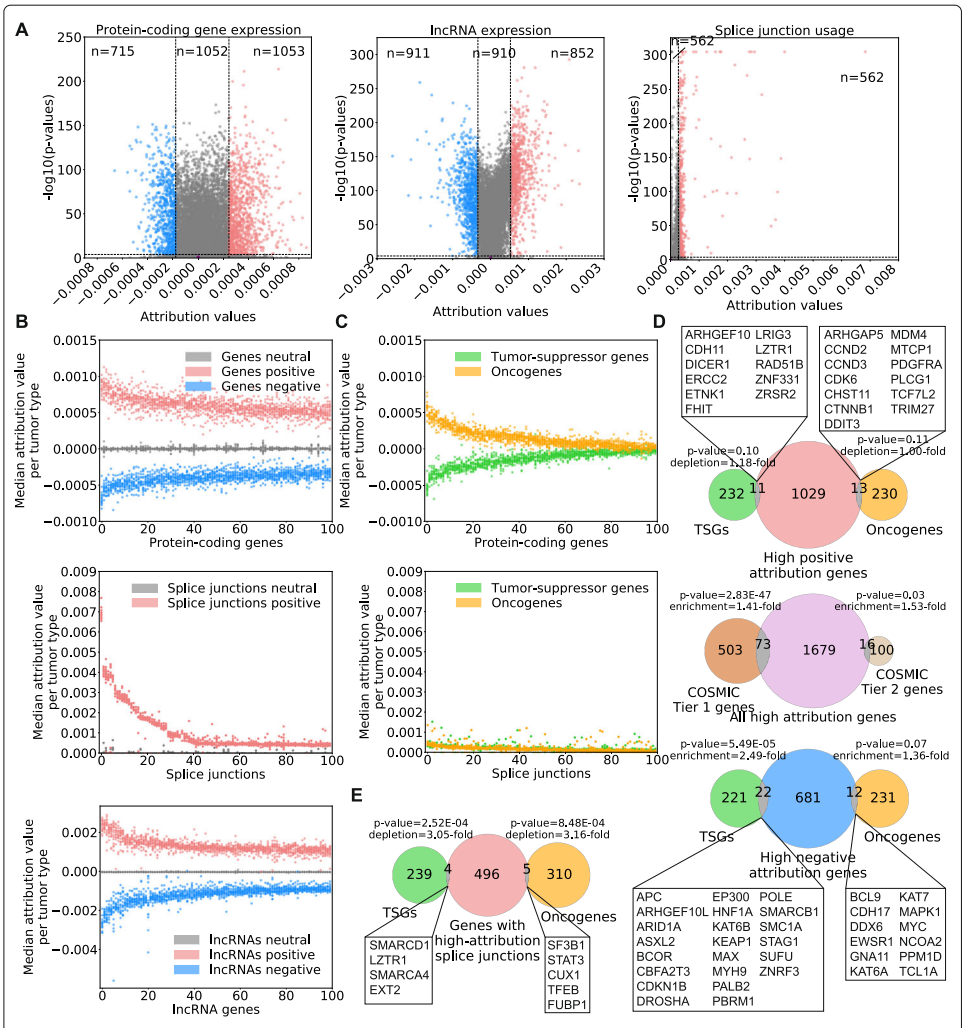


Fig. 2 **A** Selection of high-attribution features from models trained with protein-coding gene expression, lncRNA gene expression or splice junction usage. Dotted lines show cutoffs used; purple points around coordinates (0,0) show features selected in neutral sets. **B** Median attribution values of 100 protein-coding genes, lncRNAs or splice junctions with the highest positive and negative attributions across tumor tissues. **C** Median attribution values of genes associated with cancer from the COSMIC database. **D** Overlap between COSMIC oncogenes and TSGs, and genes with high positive (top panel) or negative (bottom panel) attribution values, or between COSMIC tier 1 genes (high confidence for causal role in cancer) and tier 2 genes (some evidence of causal involvement in cancer), and all high-attribution genes (central panel). **E** Overlap between genes associated with cancer and genes harboring junctions with high attribution values. In both **D** and **E**, enrichment or depletion factors were calculated from the ratio of observed vs. expected overlapping genes between sets, and *p*-values were calculated using the hypergeometric test

highest median attribution values across all tumor samples have consistently high attribution values in most if not all cancer types (Fig. 2B), highlighting that our models are not driven by outlier expression or splice junctions usage in cancer types with a large sample size, but rather rely on common transcriptomic features of cancer.

In agreement with our differential gene expression analysis that showed that no gene is significantly deregulated in the same manner across all tumor types, we find that the sign of the attribution value for a given gene does not necessarily reflect the change in expression in cancer. In other words, a gene with a high positive attribution value would not necessarily be upregulated in all or most cancers, and conversely, a gene with a high negative attribution value would not necessarily be downregulated in all or most cancers. Thus, rather than highlighting genes and splicing variations that are similarly altered in many cancer types, the interpretation of our models exposes transcriptomic variations that consistently deviate from the norm in cancer. The transcriptomic variations that they identify could reflect changes that cause or are a consequence of cancer, or a mixture of both.

We next sought to assess the relation between model attributions and known oncogenes or tumor suppressor genes (TSGs). Strikingly, we found a clear separation between the latter two groups with oncogenes receiving positive attribution and TSGs receiving negative values (Fig. 2C). However, most of the known oncogenes and TSGs have lower attribution values relative to our top-scoring features, with many having neutral attribution values (close to 0). This result was observed with attribution values from both the expression of COSMIC genes or usage of splice junctions found in those genes. We only observed a small, although statistically significant, enrichment of COSMIC genes among our high negative attribution genes (Fig. 2D). Of note, well-known oncogenes and TSGs are depleted among genes that have splice junctions with high attributions, meaning that there are fewer oncogenes and TSGs with high-attribution splice junctions than would be expected by chance (Fig. 2E). These results show that our models rely on gene expression and splicing variations in genes that mostly differ from established oncogenes and TSGs to predict tumor samples and that the transcriptomic definition of cancer that we provide here largely differs from genes harboring hallmark mutations causally implicated in cancer.

Frequency of genetic alterations in transcriptomic features characterizing cancer state

Next, we wondered if previously unreported genetic alterations in our high-attribution genes might be driving the transcriptomic variations highlighted by our models. We first postulated that high-attribution genes would rarely carry driver mutations since these genes are not known to be genetically linked to cancer, which we confirmed by investigating TCGA samples and finding that less than 2% of high-attribution genes carry a driver mutation in at least one of any of the samples in TCGA (Fig. 3A). While high-attribution genes do not carry driver mutations, our analysis shows that genes with high negative attribution values by expression display a higher frequency of passenger mutations than their reference neutral set and that the frequency of passenger mutations in high negative attribution genes is as high as in COSMIC oncogenes (Fig. 3B). The frequency of structural variants, although higher in high-attribution genes than their reference neutral sets, is lower for all sets of high-attribution genes than for COSMIC genes (Fig. 3C). Similarly, the frequency at which high-attribution genes are impacted by amplification (Fig. 3D) or deletion events (Fig. 3E) is not significantly different from the neutral sets or the COSMIC

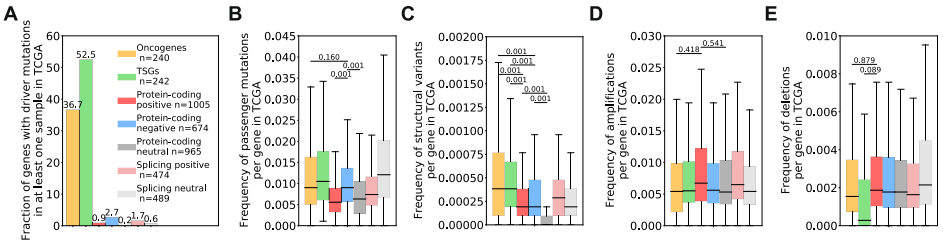


Fig. 3 Fraction of genes with driver mutations (**A**), frequency of passenger mutations (**B**), structural variants (**C**), amplification events (**D**), or deletion events (**E**) in the TCGA cohort. The analysis was performed on the following sets of genes: COSMIC oncogenes (yellow), COSMIC tumor-suppressor genes (TSGs, green), genes with high positive ("protein-coding positive," red), high negative ("protein-coding negative," blue), or neutral attribution values by expression ("protein-coding neutral," gray) or genes with splice junctions with high ("splicing high," light red) or neutral ("splicing neutral," light gray) attribution values in our models. *p*-values were calculated using a one-way ANOVA with Tukey post hoc tests

genes. Overall, we conclude that the cancer transcriptomic features we identified are not frequently affected by genetic alterations, which suggests that the cancer expression and splicing patterns obtained from our models are not driven by genetic variations in these genes.

High evolutionary and selective constraints in transcriptomic features defining tumor state

After establishing a list of genes with high attribution values by expression or splice junction usage and discovering that most of these genes do not correspond to COSMIC oncogenes or TSGs, we wondered whether transcriptomic features that carry high attribution values in our models have properties that may indicate important roles in cells. We discovered that protein-coding genes, lncRNA genes and genes with splice junctions corresponding to high-attribution features in our models are highly evolutionarily conserved relative to the neutral sets (Fig. 4A). We noted that protein-coding genes that have high negative attributions as well as lncRNA genes that have high positive or negative attributions are in general significantly longer than the reference neutral sets, but that genes with splice junctions with high attributions are significantly shorter (Fig. 4B). We also observed that protein-coding genes and genes with splice junctions with high attributions display high selective pressure against loss of function mutations, as estimated by the gnomAD LOEUF score [26] (Fig. 4C).

Finally, we inferred the functional impact of lncRNA genes with high attributions by examining the density of a class of DNA motifs termed pyknons. Pyknons are located in loci that were previously reported as often differentially transcribed between normal and colorectal cancer tissues and that can affect the oncogenic functions of lncRNAs [27–29]. We found that high-attribution lncRNA genes carry a higher density of pyknons (Fig. 4D) than lncRNA genes from the neutral set. This was true for both positive- and negative-attribution lncRNAs, but it was particularly marked in negative-attribution lncRNAs, where average pyknon density is seven times higher than neutral-attribution lncRNAs. Together, these findings show that high attribution protein-coding and lncRNA genes by expression or splicing are under strong evolutionary and selective constraints and suggest that these protein-coding genes and lncRNAs with altered expression or abnormal splicing in cancer have essential functions in cells.

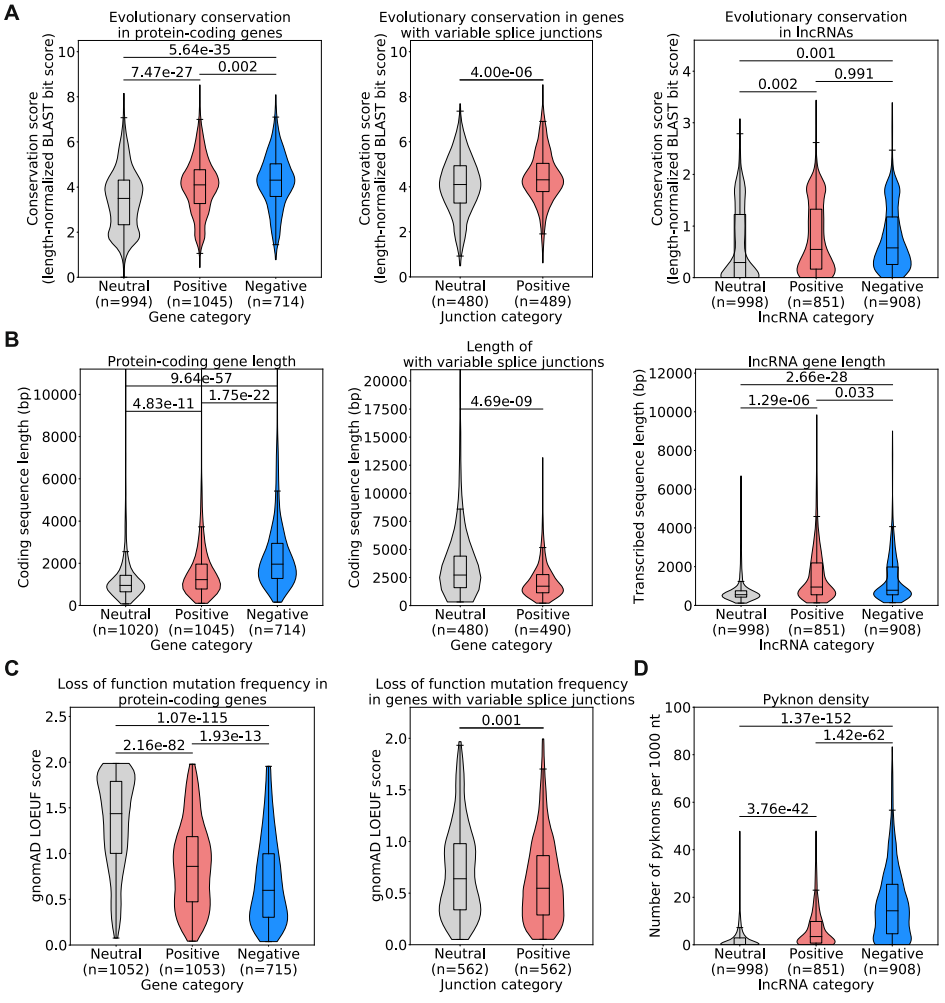


Fig. 4 **A** Evolutionary conservation of protein-coding genes (left panel), genes with variable splice junctions (middle panel) or lncRNA genes (right panel) across human, chimpanzee, mouse, cattle, xenopus, zebrafish, and chicken. **B** Gene length derived from the longest annotated transcript in Ensembl (hg38) for protein-coding genes (left panel), genes with variable splice junctions (middle panel), or lncRNA genes (right panel). **C** Selective pressure against loss-of-function mutations in the human population as assessed by gnomAD LOEUF score [26], showing score for protein-coding genes (left panel) or genes with splice junctions (right panel) with high attribution values. A low LOEUF score implies high selective pressure against loss of function. **D** Pyknon density in lncRNA genes with high attribution values. All *p*-values shown are calculated using an unpaired *t*-test

Characterization of splice junctions with high attributions

While it is easy to conceive how changes in the expression level of a gene can drive tumorigenesis, interpreting the impact of splicing changes in disease is not as straightforward. We thus wanted to assess how variable splice junctions with high attributions are predicted to impact protein sequence and function. We first noted that high-attribution junctions are predicted to disrupt the reading frame of the gene as often as our reference neutral junction set (Additional file 1: Fig. S7A). Previous studies have shown that alternative splicing can modulate protein-protein interactions by targeting disordered regions [30–32]. Therefore, we looked at predicted disorderness of the peptide sequence corresponding to the two exons immediately upstream and downstream of variable splice

junctions but found that the predicted peptide disorderness level is no different in high-attribution junctions from what we observe in the neutral set (Additional file 1: Fig. S7B).

We then assessed whether high-attribution splice junctions affect known protein domains by predicting the protein domains encoded from the two exons immediately upstream and downstream of high-attribution junctions using the NCBI Conserved Domain Database. Interestingly, we discovered that 11 splice junctions in 10 genes (*CSNK2A2*, *MAPK9*, *RIOK1*, *PRKDC*, *TYK2*, *PAK1*, *IRAK1*, *CSNK2A1*, *VRK1*, *MARK3*) affect a part of the transcript matching sequences of protein kinase C (PKC)-like superfamily domains (Fig. 5). Genes contributing to PKC signaling have been implicated in cancer as oncogenes or tumor suppressors [33], but little is known about the impact of splicing variations altering PKC-like superfamily domains in cancer. We also found additional high-attribution splice junctions that affect other domains that are linked to cancer signaling, in particular DEAD-like, RING, and C2 domains. Thus, it is possible that some of the high-attribution splice junctions that we uncovered regulate cancer through the alteration of cancer signaling protein domains.

Contrasting functions of genes with high positive or negative attributions by expression or splicing in cancer

Finally, given that a majority of the protein-coding genes or genes with splice junctions with high attribution values in our models were not previously associated with cancer, we sought to understand the functions of those genes. We first checked whether genes that have high attribution values by expression differ from the genes that have high attributions by splice junction usage and confirmed that a large majority of the genes with high attributions by expression differ from the genes with high attributions by splice junction usage (Fig. 6A). We performed a Gene Ontology analysis for protein-coding genes with high attribution values and found that protein-coding genes with high negative attribution values are enriched for functions related to transcription, mitosis, histone modification, chromatin regulation, and localization to the centrosome, in line with the traditional view of cancer (Additional file 1: Fig. S8A). In sharp contrast, protein-coding genes with high positive attribution values are enriched for post-transcriptional and post-translational modifications, in particular tRNA modification, RNA splicing and protein neddylation, as well as membrane-bound organelles (Additional file 1: Fig. S8B). Similar to protein-coding genes with high positive attribution values, genes with splice junctions with high

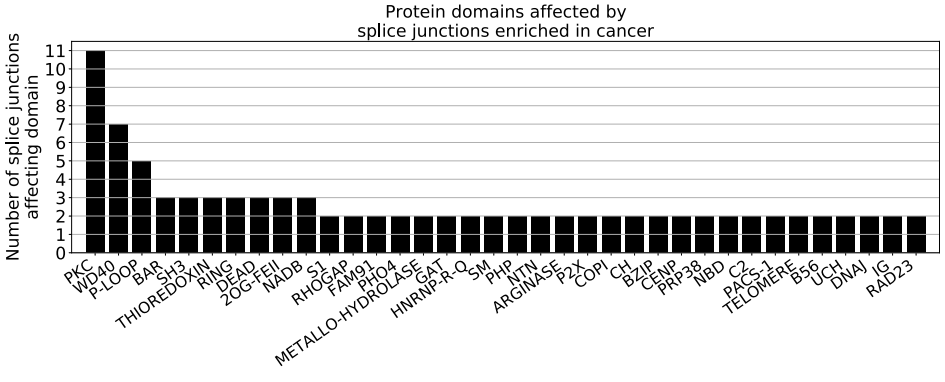


Fig. 5 Protein domains that are affected by at least two splice junctions with high attributions

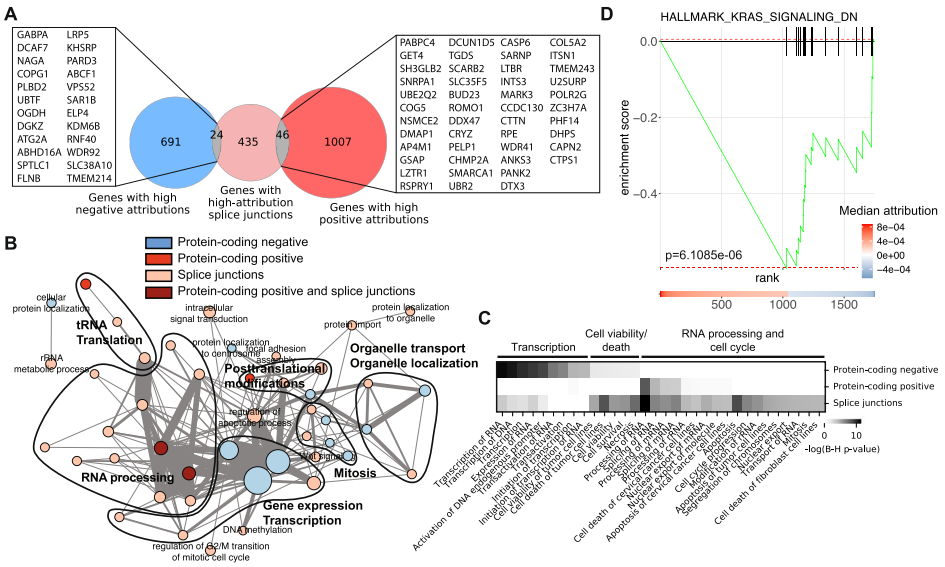


Fig. 6 **A** Overlap between genes with high positive (red set) or negative (blue set) attribution values and genes with splice junctions that have high attribution values (light red set) with the list of genes overlapping between sets. **B** Enrichment map showing GO terms related to biological processes that are enriched among protein-coding genes with high negative (protein-coding negative, blue) or positive (protein-coding positive, red) attributions by expression, genes with high-attribution splice junctions (splice junctions, light red), or enriched both in protein-coding positive and splice junction genes (protein-coding positive and splice junctions, dark red). Each node is a GO term and the color of the nodes corresponds to gene sets in which they are enriched. The thickness of edges corresponds to the number of genes in common between GO terms. **C** Heatmap of terms obtained from an Ingenuity Pathway Analysis (QIAGEN) for molecular and cellular functions. **D** Gene set enrichment analysis on high-attribution protein-coding genes by expression showing an enrichment in KRAS signaling among negative attribution genes. High-attribution genes were ranked on the *x*-axis from high positive (left, red) to high negative attribution (right, blue)

attribution values are also enriched for functions related to RNA processing, in particular, splicing and transport, but also carry a component of terms related to chromatin (Additional file 1: Fig. S8C). Of note, the set of genes with neutral attribution values by expression are enriched for heterogeneous and unrelated terms (Additional file 1: Fig. S8D) and genes with neutral attribution values by splice junction usage failed to return any terms with an adjusted p-value < 0.05, indicating that our high-attribution genes consist of sets of biologically related and consistent functions.

An enrichment map built from GO terms related to biological processes shows that high positive attribution genes by expression or splicing form a highly interconnected network whose core relates to functions associated with RNA biology (Fig. 6B). The network core from high positive attribution genes by expression or splicing differs from the highly interconnected network core drawn from genes with high negative attributions by expression, which are associated with gene expression and transcription (Fig. 6B). Ingenuity Pathway Analysis for molecular and cellular functions associated with high-attribution genes confirmed that functions of high negative attribution genes are distinct from those of high positive attribution genes, with transcription and RNA processing being predominant in each group, respectively (Fig. 6C). Pathway analysis also revealed how high negative attribution genes by expression and genes with high attribution by splice junction usage partially overlap in functions related to cell survival and how high attribution genes by splice junction usage are involved in the cell cycle. Finally, gene set enrichment analysis

unveiled an enrichment of high negative attribution genes for KRAS signaling (Fig. 6D), while no significant enrichment was found for genes with high positive attributions by expression or splicing.

Thus, while genes that have high negative attribution values in cancer share functions of known oncogenes and TSGs, including in how they are implicated in genome maintenance and transcription, genes that have high positive attributions by expression or splicing have distinct functions, several of which are related to RNA regulation and RNA processing.

Discussion

Our results demonstrate that feed-forward neural networks can be used to distinguish between normal and tumor samples using transcriptomic features. Importantly, we show that models trained with lncRNA expression or splice junction usage perform as well as, if not better than, a model trained with protein-coding expression data. This observation highlights how various elements of the transcriptome can inform on disease state and emphasizes the importance of pursuing molecular markers beyond variations in protein-coding gene expression, in particular, by assessing variations in lncRNA gene expression and splice junction usage in cancer. Our approach uncovered common transcriptomic profiles consisting of a number of gene expression and splicing variation markers that are not altered in the same way across all solid tumor types, making up a novel molecular definition of cancer that would be impossible to establish using traditional approaches such as differential gene expression analysis.

The interpretation of our models revealed known and novel molecular features of cancer. Known cancer drivers were moderately enriched among genes that we find to have high attribution values, which can be expected for genes with driver mutations resulting in loss of function or lower protein expression. In addition, among genes with high negative attribution values were genes with functions typically associated with genome integrity maintenance, such as histone modification and chromatin regulation, as well as transcription, two long-known aspects of cancer development [34, 35]. On the other hand, many of the protein-coding genes that have high positive attribution values have roles in RNA regulation or RNA processing. Interestingly, RNA deregulation has become a recurrent theme in cancer research [36]. Driver mutations have been found in several RNA-binding proteins (e.g., *SF3B1*, *U2AF1*, *SRSF2*, *HNRNPA2B1*, *SRRM2*) in cancers ranging from blood malignancies to glioblastoma [37–39], but several questions remain regarding how widely this group of proteins and their targets are involved in cancer. Our results suggest that RNA deregulation might be a central component of cancer, upon which many cellular pathways involved in cancer may converge. Indeed, network analysis shows that genes with high attribution values by expression or splice junction usage and that have functions related to RNA regulation are tightly connected to the canonical pathways of cancer (Fig. 7 and Additional file 2: Table S13).

Our transcriptomic definition of cancer includes several elements that were not previously genetically associated with cancer but that display strong sequence constraints, which suggests that these genes or splice junctions play essential roles in cells. Interestingly, several of these genes that are not listed as COSMIC oncogenes still display tumorigenic characteristics. A few examples include *DYNC1H1* [40], *WSB1* [41–43], *RUFY3* [44, 45], *DOCK5* [46], *MYSM1* [47], *DSE* [48], *DCUN1D5* [49, 50], *SARNP* [51],

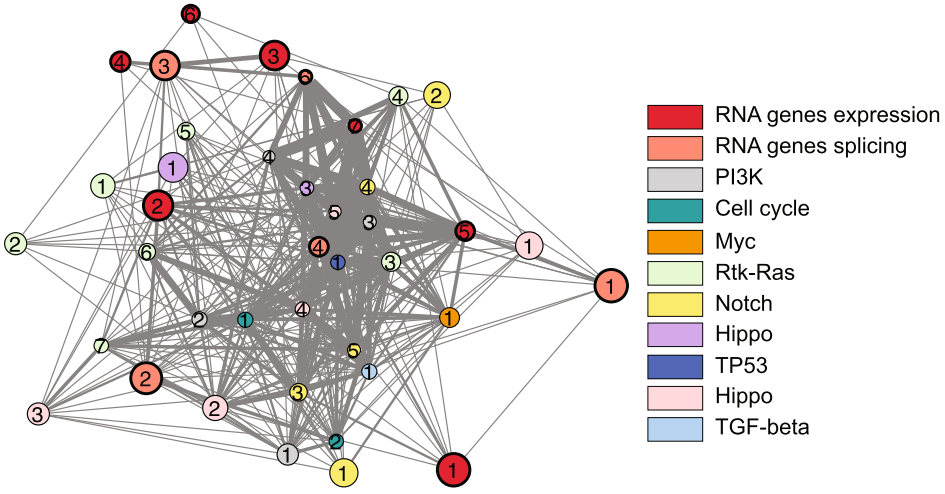


Fig. 7 Network analysis of common cancer pathways (PI3K, cell cycle, Myc, Rtk-Ras, Notch, Hippo, TP53, Hippo, TGF-beta) together with genes in GO terms related to RNA regulation or RNA processing that are enriched in our protein-coding or splicing models. Each node is a network as predicted with Ingenuity Pathway Analysis; the size of the nodes represents the number of molecules in each network and the thickness of the edges represents the number of molecules in common between two networks. Networks formed by high-attribution genes in our protein-coding and splicing models are highlighted with a thicker node border. Only networks comprising at least three molecules and connected by at least two shared molecules are shown. Numbers identify networks; Additional file 2: Table S13 lists the molecules found in each network (node)

and *FNTA* [52, 53], which can all promote cell proliferation or transformation, at least in some conditions. Likewise, we have identified splice junctions in cancer that deviate from normal tissues in functional domains implicated in cancer signaling, such as PKC-like [33], DEAD-like [54] and RING [55] domains, in genes associated with cancer, such as *CSNK2A1*, *CSNK2A2*, *RIOK1*, *PRKDC*, *TYK2*, *PAK1*, and *IRAK1*, for which gene expression and posttranslational modifications act as mechanisms for cancer progression [56–61]. However, splicing variations related to cancer have not been reported for any of these genes except *PAK1*, for which a JMJD6-regulated exon inclusion event altering the PKC domain enhances MAPK signaling in melanoma [62]. While the splice junction we identified differs from the one reported before, it also affects the PKC domain of PAK1. *IRAK1* has two well-characterized alternative splicing events [63], but there exists no evidence that these events are directly involved in tumorigenesis, and they also differ from our splice junctions with high attributions. Overall, keeping in mind that the models we developed were not designed for clinical application, their robustness across tumor types and the functional properties of their most informative features hint that our signatures could be leveraged to design markers for cancer detection.

Interestingly, our analysis of variant frequency shows that high negative attribution genes in our protein-coding gene expression model are more frequently mutated than the neutral set, and almost as frequently as COSMIC oncogenes and TSGs. In contrast, variant frequency is much lower for genes with high positive attributions by expression or splicing. This observation could explain why many of these transcriptomic features have previously been overlooked in genomic studies. In addition, while the directionality of the attribution value does not directly reflect the difference in expression across all cancer types (e.g., a gene with a high positive attribution would not necessarily have higher expression in all cancer types), we noticed that known oncogenes generally have

positive attribution values and known tumor suppressors generally have negative attribution values (Fig. 2C). This observation suggests that positive attribution genes could be considered "oncogene-like" while negative attribution genes could be considered "tumor-suppressor-like" in the way they are altered in cancer and, perhaps, in how they contribute to cancer biology.

Conclusions

Altogether, our results show that alteration of RNA processing pathways is a hallmark of several types of cancer. Future work should be directed at assessing whether the transcriptomic features of cancer that we highlight here are causally involved in tumorigenesis or in tumor suppression, thereby highlighting a core transcriptomic component of cancer development, or whether they represent the downstream consequences of genetic alterations in core transcriptional circuitries of cancer [64].

Methods

Notation

We are interested in defining the transcriptomic signature of solid tumors. We achieve this by first predicting the cancer state of an RNA-seq sample using a deep learning model with gene expression (protein-coding or lncRNA) or splicing quantification as input. Subsequently, we interpret the prediction made by the deep learning model given our input observation by assigning attributions to each feature of the observation. Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ be the output or label space. Input $\boldsymbol{x}$ is in a p-dimensional feature space $\mathcal{X} = \mathbb{R}^p$. Since we only consider a binary classification task for defining the cancer state, $\mathcal{Y} = \{0, 1\}$, where 0 represents normal tissue and 1 represents tumor. Predictions are obtained by a prediction function on the feature space $F : \mathcal{X} \rightarrow \mathcal{Y}$. The goal of the interpretation step is to obtain a p-dimensional vector of attributions called $\mathbf{attr} \in \mathbb{R}^p$, with each value representing how each of the p features contributes to the prediction $F(\boldsymbol{x})$.

Datasets

In this work, we process RNA-seq samples from normal human tissues and tumors from 14 datasets (see Additional file 2: Table S2 for a list of all samples and their tissue/cancer identity and Fig. 1B, C for the number of samples representing each tissue or tumor type and source dataset, respectively). The two largest datasets among these are from the Genotype-Tissue Expression (GTEx) consortium and the Cancer Genome Atlas (TCGA) program. We processed 5622 samples from 19 normal tissues and 7839 samples from 18 cancer types. Since large datasets often suffer from batch effects, we included 12 other datasets in our analysis. These datasets included lung [65, 66], liver [67], stomach [68], breast [69–71], brain [72], and colon [73] tumor samples with matched normal samples, as well as head and neck [74–76], pancreatic [77], ovarian [78], and prostrate [79] tumor samples without matched normal samples (see Additional file 2: Table S2 for the number of samples and dataset labels). For testing on independent, unseen tissue and tumor types, we processed 16 macrophage samples, 8 monocyte samples and 9 $CD4^+$ lymphocyte samples from the Blueprint project [80], 35 $CD4^+$ and 14 $CD8^+$ lymphocyte samples from dataset E-MTAB-2319 [81], and 40 B-cell acute lymphoblastic leukemia samples and 40 acute myeloid leukemia from the pediatric TARGET cohort [82].

RNA-Seq data processing

In order to minimize the introduction of technical biases, all RNA-Seq samples were processed from fastq files in the same manner. The raw reads from RNA-Seq experiments are passed through quality control using FastQC [83]. Sequencing adapters were trimmed with TrimGalore (v0.6.6) [84], reads were aligned with STAR (v2.5.2a) [85] against the hg38 human genome assembly [86], and mapped reads were sorted and indexed using samtools (v1.11). Gene expression quantification was carried out using Salmon (v0.14.0) [87] in quasi-mapping mode using an index generated with Ensembl GRCh38 transcriptome release 94. Splice junctions were quantified using MAJIQ (v2.1) [88]. MAJIQ defines alternative splicing in terms of local splicing variations (LSVs). LSVs can be binary or complex. Binary LSVs comprise only two junctions and complex LSVs have more than two junctions. We only use binary splicing variations that were quantified in at least 80% of samples of a given tissue or tumor type and select junction usage value randomly from one of the two junctions composing the splicing variation. The splicing quantification of a junction in a condition is called percent spliced-in (PSI). For binary LSVs, PSI measures the ratio of the number of reads supporting the inclusion of a junction in a condition over the total number of reads supporting its inclusion or exclusion. MAJIQ uses a beta-binomial distribution over the reads to quantify PSI. For more details on the statistical model, please refer to [88].

Batch correction

To mitigate batch effects from our RNA-seq data, we take two steps. First, in addition to GTEx and TCGA, we searched for other datasets with normal tissues and tumor samples. This step is necessary as our signal of normal tissues versus tumor is confounded with whether the sample came from GTEx or TCGA. GTEx consists of normal tissues and TCGA consists of mostly tumor samples. Therefore, we added 12 other small datasets to ensure that we learn cancer-specific signals and are not confounded by dataset-specific technical biases. We found additional data sources for 13/19 normal tissues and 8/18 tumor types. Next, we correct dataset bias for gene expression data by mean-centering each tissue/tumor-type separately across datasets. The batch correction step ensures our deep learning models using protein-coding and lncRNA gene expression generalize across multiple datasets.

We opted for tissue-specific mean correction for mitigating batch effects instead of standard batch correction methods such as COMBAT because our dataset did not meet the criteria required for traditional batch correction methods, such as the requirement for having at least two data sources for each tissue/tumor type [21]. However, to ensure the robustness of attributions generated by mean-corrected data, we compared it with attributions generated by COMBAT-corrected data. We ran COMBAT batch correction on 11/19 normal tissues and 7/18 tumor types, for which data was available from multiple sources. Two additional normal tissues and one tumor type with multiple data sources had high imbalance, which led COMBAT to filter approximately 14,000–15,000 genes, thus making the corrections unusable. As expected, the application of COMBAT on our gene expression data results in the loss of approximately 5500 genes. However, it is worth noting that despite these limitations we observe a very high correlation between the attribution values identified by both types of data pre-processing (mean correction vs. COMBAT correction, see Additional file 1: Fig. S2C, right panel; Pearson's R 0.90,

Spearman's *R* 0.85), indicating the stability of our top feature attributions regardless of the batch correction method.

Tumor classification models

We train three deep learning models for defining the transcriptomic signature of solid tumor samples. Each of these models uses a different set of transcriptomic features to predict the cancer state of an RNA-seq sample: protein-coding gene expression, lncRNA gene expression and quantification of splice junctions. Since we have thousands of noisy transcriptomic features, we first train an autoencoder model for dimensionality reduction followed by a supervised feed-forward neural network for the cancer state prediction. In the next three sections, we describe these models in detail.

Protein-coding gene expression based model

In our first model, the input features are the expression values of 19,657 protein-coding genes. We extracted the protein-coding genes from Ensembl BioMart (see Additional file 2: Table S14 for the list of protein-coding genes). Using these features, we first train an autoencoder model and then, using the reduced feature set from the latent space of the autoencoder, we train a supervised feed-forward neural network that predicts normal tissue versus tumor for each RNA-seq sample. The encoder in our autoencoder model has two latent layers followed by a third latent layer that produces the feature set with reduced dimensionality. The decoder mirrors the encoder. It takes in the features from the third latent layer of the encoder and reconstructs the gene expression values of the protein-coding genes. Using the latent features from the encoder as input features, we then train a discriminator network with three latent layers followed by the output layer. Both the autoencoder and the discriminator networks use the ReLU activation function and are trained using the adam optimizer (see Additional file 2: Table S1 for the detailed architecture of both models).

LncRNA gene expression based model

For the next model, the input features are the gene expression of 14,257 lncRNA genes. We extract the lncRNA genes from Ensembl BioMart. See Additional file 2: Table S15 for the list of lncRNA genes. The model architecture and training process of the lncRNA-based model is similar to the protein-coding gene expression model described in the previous section (see Additional file 2: Table S4 for the detailed architecture of the lncRNA autoencoder and discriminator networks).

Splicing junctions based model

Finally, the input features for our third model are the splicing quantification for 40,147 alternative splice junctions from 11,219 genes. See Additional file 2: Table S16 for the list of genes. We generated the splicing quantification for these junctions in the normal tissues and tumors using MAJIQ (see the "RNA-Seq data processing" section for details on quantification of these splicing junctions from the RNA-seq data). As we have thousands of splicing junctions as features, we again train an autoencoder for dimensionality reduction followed by a supervised neural network for tumor classification. The model architecture and training process of this model is similar to the previous gene expression

(protein-coding or lncRNA) based models (see Additional file 2: Table S5 for the detailed architecture of the splicing junctions based models).

Interpretation of tumor classification models

In order to find the features responsible for classifying an RNA-seq sample as tumor, we employ the enhanced integrated gradients (EIG) method for interpretation of a deep learning model [19]. Here, interpretation means attributing the prediction of a deep learning model to its input features. Briefly, EIG computes feature attribution by aggregating gradients along a linear/non-linear path between a sample and a class-agnostic/specific baseline. Sample here refers to an input sample to the deep learning model. Baseline refers to a model's proxy to human counterfactual intuition. This implies that humans assign blame for the difference in two entities on attributes that are present in one entity but absent in the other. EIG offers multiple baselines and paths. In this work, since we want to find features that distinguish between tumor samples from normal tissue samples, we use normal tissues as the baseline class. Specifically, we use the median in the latent space over the normal tissue samples. Then, we compute the attributions for the given tumor samples by aggregating the gradients between the baseline and each tumor sample along a linear path in the original feature space. We assess the class-wide significance of each feature by computing p-values by comparing the attribution distribution of a feature for the tumor class versus a random mixture of normal tissues and tumors using a one-sided t-test with FDR correction for multiple hypothesis testing. For further details on enhanced integrated gradients, please refer to [19].

Selection of feature sets

High-attribution feature sets were selected from features with an adjusted p-value $<$ 0.0001 (Benjamini-Hochberg FDR correction $<$ 0.01) and ranking above the knee-point of the curve in the case of positive attribution values or below the knee-point of the curve in the case of negative attribution values. A neutral set of a size equivalent to the number of high-attribution features was selected among genes that had FDR-corrected p-value $>$ 0.05 and ranking in the middle of the distribution of attribution values (attribution values close to 0). The list of COSMIC oncogenes and tumor suppressor genes (TSGs) was established from the COSMIC genes census that had a role in cancer, comprising the annotation "oncogene" but not "TSG" for oncogenes and comprising the annotation "TSG" but not "oncogene" for TSGs.

Gene ontology analysis, gene set enrichment analysis and Ingenuity Pathway Analysis

Gene ontology analysis was performed with EnrichR [89] v.1.0 using a 2018 release of the GO Consortium annotations and including terms from the molecular function, cellular component and biological process categories. Gene set enrichment analysis was performed using GSEA4.1.0 [90, 91] against the msigdb.v7.4 gene set library and filtered for enrichment corresponding to hallmark, reactome and GO gene sets with a normalized p-value $<$ 0.01.

Splice junction characterization

Disorderness was predicted with IUPred2 [92] from the two exons immediately upstream and downstream of the most distant splice site corresponding to a variable junction.

Protein domains were predicted from the same transcript region with the NCBI WEB CD-search tool [93].

Functional characterizations

The conservation score was calculated from the summation of BLAST bit scores from 6 species: human, chimpanzee, mouse, cattle, xenopus, zebrafish, and chicken (taxids: 9606, 9598, 10,090, 8364, 7955, 9031, 9913), normalized to the length of the human transcript. Loss of function mutation frequency is expressed as the gnomAD LOEUF score only for genes for which a LOEUF score was reported [26]. Pyknon density was calculated using the list of human pyknons available from the pyknon database [94] and shown as the number of pyknons found per 1000 nt in the longest RefSeq annotated transcript.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-022-02681-3.

Additional file 1: Supplementary Figs. S1–8.
Additional file 2: Supplementary Tables 1–16.
Additional file 3: Review history.

Acknowledgements
We would like to thank Joseph K. Aicher and Paul Jewell for their assistance in processing the gene expression and splicing data.

Peer review information
Stephanie McClelland was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
Review history is available as Additional file 3.

Authors' contributions
A.J. and M.Q.V. designed the study, analyzed and interpreted the data, and wrote the manuscript. A.T.T, K.W.L., and Y.B. interpreted the data and wrote the manuscript. A.J. and D.W. did the batch correction analysis. All authors have read the manuscript and agreed to its publication.

Funding
This research was supported by NIH grant R01 LM013437 to Y.B., R01 GM128096 to Y.B. and U01 CA232563 to Y.B., K.W.L. and A.T.T.

Availability of data and materials
All processed data and code to reproduce the figures are available from a Bitbucket repository [95]. Source code is also available from Zenodo [96] under a BSD 3-clause license. RNA-Seq samples used for this study are publicly available and are listed in Additional file 2: Table S2.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Computer and Information Science, School of Engineering and Applied Science, Philadelphia, USA. [2]Department of Genetics, Philadelphia, USA. [3]Department of Biochemistry and Biophysics, Philadelphia, USA. [4]Department of Pathology and Laboratory Medicine, Philadelphia, USA. [5]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA. [6]Division of Cancer Pathobiology, Children's Hospital of Philadelphia, Philadelphia, USA.

Received: 19 November 2021 Accepted: 27 April 2022
Published online: 17 May 2022

References

1. Haigis KM, Cichowski K, Elledge SJ. Tissue-specificity in cancer: the rule, not the exception. Science. 2019;363(6432):1150–1. https://doi.org/10.1126/science.aaw3472.
2. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, et al. Oncogenic signaling pathways in the cancer genome atlas. Cell. 2018;173(2):321–37.
3. Paull EO, Aytes A, Jones SJ, Subramaniam PS, Giorgi FM, Douglass EF, Tagore S, Chu B, Vasciaveo A, Zheng S, Verhaak R, Abate-Shen C, Alvarez MJ, Califano A. A modular master regulator landscape controls cancer transcriptional identity. Cell. 2021;184(2):334–35120. https://doi.org/10.1016/j.cell.2020.11.045.
4. Dave SS, Wright G, Tan B, Rosenwald A, Gascoyne RD, Chan WC, Fisher RI, Braziel RM, Rimsza LM, Grogan TM, et al. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. N Engl J Med. 2004;351(21):2159–69.
5. Roessler S, Jia H-L, Budhu A, Forgues M, Ye Q-H, Lee J-S, Thorgeirsson SS, Sun Z, Tang Z-Y, Qin L-X, et al. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. Cancer Res. 2010;70(24):10202–12.
6. Wei IH, Shi Y, Jiang H, Kumar-Sinha C, Chinnaiyan AM. RNA-seq accurately identifies cancer biomarker signatures to distinguish tissue of origin. Neoplasia. 2014;16(11):918–27. https://doi.org/10.1016/j.neo.2014.09.007.
7. Ahn T, Goo T, Lee C-H, Kim S, Han K, Park S, Park T. Deep learning-based identification of cancer or normal tissue using gene expression data. IEEE Int Conf Bioinforma Biomed (BIBM). 2018. https://doi.org/10.1109/bibm.2018.8621108.
8. Grewal JK, Tessier-Cloutier B, Jones M, Gakkhar S, Ma Y, Moore R, Mungall AJ, Zhao Y, Taylor MD, Gelmon K, Lim H, Renouf D, Laskin J, Marra M, Yip S, Jones SJM. Application of a neural network whole transcriptome–based pan-cancer method for diagnosis of primary and metastatic cancers. JAMA Netw Open. 2019;2(4):192597. https://doi.org/10.1001/jamanetworkopen.2019.2597.
9. Frost FG, Cherukuri PF, Milanovich S, Boerkoel CF. Pan-cancer RNA-seq data stratifies tumours by some hallmarks of cancer. J Cell Mol Med. 2019;24(1):418–30. https://doi.org/10.1111/jcmm.14746.
10. Liu S, Xu C, Zhang Y, Liu J, Yu B, Liu X, Dehmer M. Feature selection of gene expression data for cancer classification using double RBF-kernels. BMC Bioinformatics. 2018;19(1). https://doi.org/10.1186/s12859-018-2400-2.
11. Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artif Intell. 1997;97(1-2):245–71.
12. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3(Mar):1157–82.
13. Huang H-H, Liu X-Y, Liang Y. Feature selection and cancer classification via sparse logistic regression with the hybrid l1/2 + 2 regularization. PLoS ONE. 2016;11(5):0149675. https://doi.org/10.1371/journal.pone.0149675.
14. Al-Rajab M, Lu J, Xu Q. A framework model using multifilter feature selection to enhance colon cancer classification. PLoS ONE. 2021;16(4):0249094. https://doi.org/10.1371/journal.pone.0249094.
15. Jiang Q, Jin M. Feature selection for breast cancer classification by integrating somatic mutation and gene expression. Front Genet. 2021;12. https://doi.org/10.3389/fgene.2021.629946.
16. Xi M, Sun J, Liu L, Fan F, Wu X. Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine. Comput Math Methods Med. 2016;2016:1–9. https://doi.org/10.1155/2016/3572705.
17. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digit Signal Process. 2018;73:1–15. https://doi.org/10.1016/j.dsp.2017.10.011.
18. Hanczar B, Zehraoui F, Issa T, Arles M. Biological interpretation of deep neural network for phenotype prediction based on gene expression. BMC Bioinformatics. 2020;21(1). https://doi.org/10.1186/s12859-020-03836-4.
19. Jha A, Aicher JK, Gazzara MR, Singh D, Barash Y. Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study. Genome Biol. 2020;21(1):1–22.
20. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. The genotype-tissue expression (gtex) project. Nat Genet. 2013;45(6):580.
21. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics. 2007;8(1):118–27.
22. Ching T, Peplowska K, Huang S, Zhu X, Shen Y, Molnar J, Yu H, Tiirikainen M, Fogelgren B, Fan R, Garmire LX. Pan-cancer analyses reveal long intergenic non-coding RNAs relevant to tumor diagnosis, subtyping and prognosis. EBioMedicine. 2016;7:62–72. https://doi.org/10.1016/j.ebiom.2016.03.023.
23. Bolha L, Ravnik-Glavač M, Glavač D. Long noncoding RNAs as biomarkers in cancer. 2017;2017:1–14. https://doi.org/10.1155/2017/7243968.
24. Brinkman BMN. Splice variants as cancer biomarkers. 2004;37(7):584–94. https://doi.org/10.1016/j.clinbiochem.2004.05.015.
25. Carlevaro-Fita J, Lanzós A, Feuerbach L, Hong C, Mas-Ponte D, Pedersen JS, and RJ. Cancer LncRNA census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. Commun Biol. 2020;3(1). https://doi.org/10.1038/s42003-019-0741-7.
26. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Neale BM, Daly MJ, MacArthur DG. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43. https://doi.org/10.1038/s41586-020-2308-7.
27. Rigoutsos I, Huynh T, Miranda K, Tsirigos A, McHardy A, Platt D. Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. Proc Natl Acad Sci. 2006;103(17):6605–10. https://doi.org/10.1073/pnas.0601688103.

28. Rigoutsos I, Lee SK, Nam SY, Anfossi S, Pasculli B, Pichler M, Jing Y, Rodriguez-Aguayo C, Telonis AG, Rossi S, Ivan C, Ivkovic TC, Fabris L, Clark PM, Ling H, Shimizu M, Redis RS, Shah MY, Zhang X, Okugawa Y, Jung EJ, Tsirigos A, Huang L, Ferdin J, Gafà R, Spizzo R, Nicoloso MS, Paranjape AN, Shariati M, Tiron A, Yeh JJ, Teruel-Montoya R, Xiao L, Melo SA, Menter D, Jiang Z-Q, Flores ER, Negrini M, Goel A, Bar-Eli M, Mani SA, Liu CG, Lopez-Berestein G, Berindan-Neagoe I, Esteller M, Kopetz S, Lanza G, Calin GA. N-BLR, a primate-specific non-coding transcript leads to colorectal cancer invasion and migration. Genome Biol. 2017;18(1). https://doi.org/10.1186/s13059-017-1224-0.
29. Evangelista AF, de Menezes WP, Berardinelli GN, Santos WD, Scapulatempo-Neto C, Guimarães DP, Calin GA, Reis RM. Pyknon-containing transcripts are downregulated in colorectal cancer tumors, and loss of PYK44 is associated with worse patient outcome. Front Genet. 2020;11. https://doi.org/10.3389/fgene.2020.581454.
30. Gueroussov S, Weatheritt RJ, O'Hanlon D, Lin Z-Y, Narula A, Gingras A-C, Blencowe BJ. Regulatory expansion in mammals of multivalent hnRNP assemblies that globally control alternative splicing. Cell. 2017;170(2):324–33923. https://doi.org/10.1016/j.cell.2017.06.037.
31. Midic U, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. Protein disorder in the human diseasome: unfoldomics of human genetic diseases. BMC Genomics. 2009;10(S1). https://doi.org/10.1186/1471-2164-10-s1-s12.
32. Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol Cell. 2012;46(6):871–83. https://doi.org/10.1016/j.molcel.2012.05.039.
33. Garg R, Benedetti LG, Abera MB, Wang H, Abba M, Kazanietz MG. Protein kinase c and cancer: what we know and what we do not. Oncogene. 2013;33(45):5225–37. https://doi.org/10.1038/onc.2013.524.
34. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57–70. https://doi.org/10.1016/s0092-8674(00)81683-9.
35. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–674. https://doi.org/10.1016/j.cell.2011.02.013.
36. Goodall GJ, Wickramasinghe VO. RNA in cancer. Nat Rev Cancer. 2020;21(1):22–36. https://doi.org/10.1038/s41568-020-00306-0.
37. Anczuków O, Krainer AR. Splicing-factor alterations in cancers. RNA. 2016;22(9):1285–301. https://doi.org/10.1261/rna.057919.116.
38. Marabti EE, Younis I. The cancer spliceome: reprograming of alternative splicing in cancer. Front Mol Biosci. 2018;5. https://doi.org/10.3389/fmolb.2018.00080.
39. Sveen A, Kilpinen S, Ruusulehto A, Lothe RA, Skotheim RI. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene. 2015;35(19):2413–27. https://doi.org/10.1038/onc.2015.318.
40. Gong L-B, Wen T, Li Z, Xin X, Che X-F, Wang J, Liu Y-P, Qu X-J. DYNC1i1 promotes the proliferation and migration of gastric cancer by up-regulating IL-6 expression. Front Oncol. 2019;9. https://doi.org/10.3389/fonc.2019.00491.
41. Kim JJ, Lee SB, Yi S-Y, Han S-A, Kim S-H, Lee J-M, Tong S-Y, Yin P, Gao B, Zhang J, Lou Z. WSB1 overcomes oncogene-induced senescence by targeting ATM for degradation. Cell Res. 2016;27(2):274–93. https://doi.org/10.1038/cr.2016.148.
42. Cao J, Wang Y, Dong R, Lin G, Zhang N, Wang J, Lin N, Gu Y, Ding L, Ying M, He Q, Yang B. Hypoxia-induced WSB1 promotes the metastatic potential of osteosarcoma cells. Cancer Res. 2015;75(22):4839–51. https://doi.org/10.1158/0008-5472.can-15-0711.
43. Kim JJ, Lee SB, Jang J, Yi S-Y, Kim S-H, Han S-A, Lee J-M, Tong S-Y, Vincelette ND, Gao B, Yin P, Evans D, Choi DW, Qin B, Liu T, Zhang H, Deng M, Jen J, Zhang J, Wang L, Lou Z. WSB1 promotes tumor metastasis by inducing pVHL degradation. Genes Dev. 2015;29(21):2244–57. https://doi.org/10.1101/gad.268128.115.
44. Xie R, Wang J, Liu X, Wu L, Zhang H, Tang W, Li Y, Xiang L, Peng Y, Huang X, Bai Y, Liu G, Li A, Wang Y, Chen Y, Ren Y, Li G, Gong W, Liu S, Wang J. RUFY3 interaction with FOXK1 promotes invasion and metastasis in colorectal cancer. Sci Rep. 2017;7(1). https://doi.org/10.1038/s41598-017-04011-1.
45. Wang G, Zhang Q, Song Y, Wang X, Guo Q, Zhang J, Li J, Han Y, Miao Z, Li F. PAK1 regulates RUFY3-mediated gastric cancer cell migration and invasion. Cell Death Dis. 2015;6(3):1682. https://doi.org/10.1038/cddis.2015.50.
46. Frank SR, Köllmann CP, van Lidth de Jeude JF, Thiagarajah JR, Engelholm LH, Frödin M, Hansen SH. The focal adhesion-associated proteins DOCK5 and GIT2 comprise a rheostat in control of epithelial invasion. Oncogene. 2016;36(13):1816–28. https://doi.org/10.1038/onc.2016.345.
47. Li Y, Li J, Liu H, Liu Y, Cui B. Expression of MYSM1 is associated with tumor progression in colorectal cancer. PLoS ONE. 2017;12(5):0177235. https://doi.org/10.1371/journal.pone.0177235.
48. Liao W-C, Liao C-K, Tsai Y-H, Tseng T-J, Chuang L-C, Lan C-T, Chang H-M, Liu C-H. DSE promotes aggressive glioma cell phenotypes by enhancing HB-EGF/ErbB signaling. PLoS ONE. 2018;13(6):0198364. https://doi.org/10.1371/journal.pone.0198364.
49. Bommeljé CC, Weeda VB, Huang G, Shah K, Bains S, Buss E, Shaha M, Gönen M, Ghossein R, Ramanathan SY, Singh B. Oncogenic function of SCCRO5/DCUN1d5 requires its neddylation e3 activity and nuclear localization. Clin Cancer Res. 2013;20(2):372–81. https://doi.org/10.1158/1078-0432.ccr-13-1252.
50. Guo W, Li G-J, Xu H-B, Xie J-S, Shi T-P, Zhang S-Z, Chen X-H, Huang Z-G. In vitro biological characterization of DCUN1d5 in DNA damage response. Asian Pac J Cancer Prev. 2012;13(8):4157–62. https://doi.org/10.7314/apjcp.2012.13.8.4157.
51. Kang GJ, Park MK, Byun HJ, Kim HJ, Kim EJ, Yu L, Kim B, Shim JG, Lee H, Lee CH. SARNP, a participant in mRNA splicing and export, negatively regulates e-cadherin expression via interaction with pinin. J Cell Physiol. 2019;235(2):1543–55. https://doi.org/10.1002/jcp.29073.
52. Tian J, Fan J, Xu J, Ren T, Guo H, Zhou L. circ-FNTA accelerates proliferation and invasion of bladder cancer. Oncol Lett. 2019. https://doi.org/10.3892/ol.2019.11150.
53. Chen J, Sun Y, Ou Z, Yeh S, Huang C-P, You B, Tsai Y-C, Sheu T.-j., Zu X, Chang C. Androgen receptor-regulated circ FNTA activates KRAS signaling to promote bladder cancer invasion. EMBO Rep. 2020;21(4). https://doi.org/10.15252/embr.201948467.
54. Fuller-Pace FV. DEAD box RNA helicase functions in cancer. RNA Biol. 2013;10(1):121–32. https://doi.org/10.4161/rna.23312.

55. Lipkowitz S, Weissman AM. RINGs of good and evil: RING finger ubiquitin ligases at the crossroads of tumour suppression and oncogenesis. Nat Rev Cancer. 2011;11(9):629–43. https://doi.org/10.1038/nrc3120.
56. Zhang Y, Yang W-K, Wen G-M, Tang H, Wu C-A, Wu Y-X, Jing Z-L, Tang M-S, Liu G-L, Li D-Z, Li Y-H, Deng Y-J. High expression of PRKDC promotes breast cancer cell growth via p38 MAPK signaling and is associated with poor survival. Mol Genet Genomic Med. 2019;7(11). https://doi.org/10.1002/mgg3.908.
57. Hong X, Huang H, Qiu X, Ding Z, Feng X, Zhu Y, Zhuo H, Hou J, Zhao J, Cai W, Sha R, Hong X, Li Y, Song H, Zhang Z. Targeting posttranslational modifications of RIOK1 inhibits the progression of colorectal and gastric cancers. eLife. 2018;7. https://doi.org/10.7554/elife.29511.
58. Gray GK, McFarland BC, Rowse AL, Gibson SA, Benveniste EN. Therapeutic CK2 inhibition attenuates diverse prosurvival signaling cascades and decreases cell viability in human breast cancer cells. Oncotarget. 2014;5(15): 6484–96. https://doi.org/10.18632/oncotarget.2248.
59. Wöss K, Simonović N, Strobl B, Macho-Maschler S, Müller M. TYK2: an upstream kinase of STATs in cancer. Cancers. 2019;11(11):1728. https://doi.org/10.3390/cancers11111728.
60. Ye DZ, Field J. PAK signaling in cancer. Cell Logist. 2012;2(2):105–16. https://doi.org/10.4161/cl.21882.
61. Cheng BY, Lau EY, Leung H-W, Leung CO-N, Ho NP, Gurung S, Cheng LK, Lin CH, Lo RC-L, Ma S, Ng IO-L, Lee TK. IRAK1 augments cancer stemness and drug resistance via the AP-1/AKR1b10 signaling cascade in hepatocellular carcinoma. Cancer Res. 2018;78(9):2332–42. https://doi.org/10.1158/0008-5472.can-17-2445.
62. Liu X, Si W, Liu X, He L, Ren J, Yang Z, Yang J, Li W, Liu S, Pei F, Yang X, Sun L. JMJD6 promotes melanoma carcinogenesis through regulation of the alternative splicing of PAK1, a key MAPK signaling component. Mol Cancer. 2017;16(1). https://doi.org/10.1186/s12943-017-0744-2.
63. Jain A, Kaczanowska S, Davila E. IL-1 receptor-associated kinase signaling and its role in inflammation, cancer progression, and therapy resistance. Front Immunol. 2014;5 https://doi.org/10.3389/fimmu.2014.00553.
64. Chen Y, Xu L, Lin RY-T, Müschen M, Koeffler HP. Core transcriptional regulatory circuitries in cancer. Oncogene. 2020;39(43):6633–46. https://doi.org/10.1038/s41388-020-01459-w.
65. Ju YS, Lee W-C, Shin J-Y, Lee S, Bleazard T, Won J-K, Kim YT, Kim J-I, Kang J-H, Seo J-S. A transforming kif5b and ret gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. Genome Res. 2012;22(3):436–45.
66. Seo J-S, Ju YS, Lee W-C, Shin J-Y, Lee JK, Bleazard T, Lee J, Jung YJ, Kim J-O, Shin J-Y, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res. 2012;22(11):2109–19.
67. Yang Y, Chen L, Gu J, Zhang H, Yuan J, Lian Q, Lv G, Wang S, Wu Y, Yang Y-CT, et al. Recurrently deregulated lncrnas in hepatocellular carcinoma. Nat Commun. 2017;8(1):1–13.
68. Ooi WF, Xing M, Xu C, Yao X, Ramlee MK, Lim MC, Cao F, Lim K, Babu D, Poon L-F, et al. Epigenomic profiling of primary gastric adenocarcinoma reveals super-enhancer heterogeneity. Nat Commun. 2016;7(1):1–17.
69. Jiang Y-Z, Ma D, Suo C, Shi J, Xue M, Hu X, Xiao Y, Yu K-D, Liu Y-R, Yu Y, Zheng Y, Li X, Zhang C, Hu P, Zhang J, Hua Q, Zhang J, Hou W, Ren L, Bao D, Li B, Yang J, Yao L, Zuo W-J, Zhao S, Gong Y, Ren Y-X, Zhao Y-X, Yang Y-S, Niu Z, Cao Z-G, Stover DG, Verschraegen C, Kaklamani V, Daemen A, Benson JR, Takabe K, Bai F, Li D-Q, Wang P, Shi L, Huang W, Shao Z-M. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. Cancer Cell. 2019;35(3):428–4405. https://doi.org/10.1016/j.ccell.2019.02.001.
70. Zhang Y, Asad S, Weber Z, Tallman D, Nock W, Wyse M, Bey JF, Dean KL, Adams EJ, Stockard S, Singh J, Winer EP, Lin NU, Jiang Y-Z, Ma D, Wang P, Shi L, Huang W, Shao Z-M, Cherian M, Lustberg MB, Ramaswamy B, Sardesai S, VanDeusen J, Williams N, Wesolowski R, Obeng-Gyasi S, Sizemore GM, Sizemore ST, Verschraegen C, Stover DG. Genomic features of rapid versus late relapse in triple negative breast cancer. BMC Cancer. 2021;21(1). https://doi.org/10.1186/s12885-021-08320-7.
71. Zhou Y-F, Xiao Y, Jin X, Di G-H, Jiang Y-Z, Shao Z-M. Integrated analysis reveals prognostic value of HLA-i LOH in triple-negative breast cancer. J Immunother Cancer. 2021;9(10):003371. https://doi.org/10.1136/jitc-2021-003371.
72. Yuan Y, Jiaoming L, Xiang W, Yanhui L, Shu J, Maling G, Qing M. Analyzing the interactions of mRNAs, miRNAs, lncRNAs and circRNAs to predict competing endogenous RNA networks in glioblastoma. J Neuro-Oncol. 2018;137(3):493–502. https://doi.org/10.1007/s11060-018-2757-0.
73. Lee J-R, Kwon CH, Choi Y, Park HJ, Kim HS, Jo H-J, Oh N, et al. Transcriptome analysis of paired primary colorectal carcinoma and liver metastases reveals fusion transcripts and similar gene expression profiles in primary carcinoma and liver metastases. BMC Cancer. 2016;16(1):1–11.
74. Qin T, Zhang Y, Zarins KR, Jones TR, Virani S, Peterson LA, McHugh JB, Chepeha D, Wolf GT, Rozek LS, Sartor MA. Expressed HNSCC variants by HPV-status in a well-characterized michigan cohort. Sci Rep. 2018;8(1). https://doi.org/10.1038/s41598-018-29599-w.
75. Qin T, Koneva LA, Liu Y, Zhang Y, Arthur AE, Zarins KR, Carey TE, Chepeha D, Wolf GT, Rozek LS, et al. Significant association between host transcriptome-derived hpv oncogene e6* influence score and carcinogenic pathways, tumor size, and survival in head and neck cancer. Head Neck. 2020;42(9):2375–89.
76. Zhang Y, Koneva LA, Virani S, Arthur AE, Virani A, Hall PB, Warden CD, Carey TE, Chepeha DB, Prince ME, McHugh JB, Wolf GT, Rozek LS, Sartor MA. Subtypes of HPV-positive head and neck cancers are associated with HPV characteristics, copy number alterations, PIK3ca mutation, and pathway signatures. Clin Cancer Res. 2016;22(18):4735–45. https://doi.org/10.1158/1078-0432.ccr-16-0323.
77. Kirby MK, Ramaker RC, Gertz J, Davis NS, Johnston BE, Oliver PG, Sexton KC, Greeno EW, Christein JD, Heslin MJ, et al. Rna sequencing of pancreatic adenocarcinoma tumors yields novel expression patterns associated with long-term survival and reveals a role for angptl4. Mol Oncol. 2016;10(8):1169–82.
78. Lin X, Spindler TJ, de Souza Fonseca MA, Corona RI, Seo J-H, Dezem FS, Li L, Lee JM, Long HW, Sellers TA, et al. Super-enhancer-associated lncrna uca1 interacts directly with amot to activate yap target genes in epithelial ovarian cancer. IScience. 2019;17:242–55.
79. Yun SJ, Kim S-K, Kim J, Cha E-J, Kim J-S, Kim S-J, Ha Y-S, Kim Y-H, Jeong P, Kang HW, et al. Transcriptomic features of primary prostate cancer and their prognostic relevance to castration-resistant prostate cancer. Oncotarget. 2017;8(70):114845.

80. Fernández JM, de la Torre V, Richardson D, Royo R, Puiggròs M, Moncunill V, Fragkogianni S, Clarke L, Flicek P, Rico D, Torrents D, de Santa Pau EC, Valencia A. The BLUEPRINT data analysis portal. 2016;3(5):491–4955. https://doi.org/10.1016/j.cels.2016.10.021.
81. Ranzani V, Rossetti G, Panzeri I, Arrigoni A, Bonnal RJP, Curti S, Gruarin P, Provasi E, Sugliano E, Marconi M, Francesco RD, Geginat J, Bodega B, Abrignani S, Pagani M. The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of t cell differentiation by linc-MAF-4. 2015;16(3):318–25. https://doi.org/10.1038/ni.3093.
82. Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, Zhou X, Li Y, Rusch MC, Easton J, Huether R, Gonzalez-Pena V, Wilkinson MR, Hermida LC, Davis S, Sioson E, Pounds S, Cao X, Ries RE, Wang Z, Chen X, Dong L, Diskin SJ, Smith MA, Auvil JMG, Meltzer PS, Lau CC, Perlman EJ, Maris JM, Meshinchi S, Hunger SP, Gerhard DS, Zhang J. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. 2018;555(7696):371–6. https://doi.org/10.1038/nature25795.
83. Andrews S, et al. FastQC: a quality control tool for high throughput sequence data. Cambridge: Babraham Bioinformatics, Babraham Institute; 2010.
84. Krueger F. Trim Galore! Babraham Bioinformatics. 2018.
85. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. Star: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
86. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, et al. The ensembl genome database project. Nucleic Acids Res. 2002;30(1):38–41.
87. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417.
88. Vaquero-Garcia J, Barrera A, Gazzara MR, Gonzalez-Vallinas J, Lahens NF, Hogenesch JB, Lynch KW, Barash Y. A new view of transcriptome complexity and regulation through the lens of local splicing variations. elife. 2016;5:11752.
89. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles G, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14(1):128. https://doi.org/10.1186/1471-2105-14-128.
90. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. 2003;34(3):267–73. https://doi.org/10.1038/ng1180.
91. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. 2005;102(43):15545–50. https://doi.org/10.1073/pnas.0506580102.
92. Mészáros B, Erdős G, Dosztányi Z. IUPred2a: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 2018;46(W1):329–37. https://doi.org/10.1093/nar/gky384.
93. Marchler-Bauer A, Bryant SH. CD-search: protein domain annotations on the fly. Nucleic Acids Res. 2004;32(Web Server):327–31. https://doi.org/10.1093/nar/gkh454.
94. Tsirigos A, Rigoutsos I. Human and mouse introns are linked to the same processes and functions through each genome's most frequent non-conserved motifs. Nucleic Acids Res. 2008;36(10):3484–93. https://doi.org/10.1093/nar/gkn155.
95. Jha A, Quesnel-Vallières M, Wang D, Thomas-Tikhonenko A, Lynch K, Barash Y. Jha and Quesnel-Vallières et al. Genome Biology 2022. 2022. https://bitbucket.org/biociphers/pan-cancer/. Accessed 26 Apr 2022.
96. Jha A, Quesnel-Vallières M, Wang D, Thomas-Tikhonenko A, Lynch K, Barash Y. Jha and Quesnel-Vallières et al. Genome Biology 2022. Zenodo. 2022. https://doi.org/10.5281/ZENODO.6478482.

Publisher's Note