Identifying common transcriptome signatures of cancer by interpreting deep learning models

Anupama Jha^{1*} Mathieu Quesnel-Vallières^{2,3*} Andrei Thomas-Tikhonenko^{4,5,6} Kristen W. Lynch^{2,3}
Yoseph Barash^{1,3}

Abstract

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, signifying that cancer cases display common hallmark molecular features. 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. Here, in order to agnostically identify transcriptomic features that are commonly shared between cancer types, we used RNA-Seq datasets encompassing thousands of samples from 19 healthy 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 healthy and tumor samples. All three models achieve high precision, recall and accuracy on test sets derived from 13 datasets used during training and on an independent test dataset, indicating that our 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, suggesting that they have important cellular functions. Importantly, we found 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. Finally, our results also highlighted that deregulation of RNA-processing genes and aberrant splicing are pervasive features across a large array of solid tumor types. The transcriptomic features that we highlight here define cancer signatures that may reflect causal variations or consequences of disease state, or a combination of both.

Introduction

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 expression, activation or function of these genes are expected to lead to cancer-like behavior in various cell or tissue types and many such genes are commonly affected by genomic lesions in cancer. In addition to hallmark oncogenes and tumor suppressor genes, cancer driver mutations that contribute to disease onset and progression are found in subsets of cancer types (Haigis et al., 2019). While these genetic alterations are diverse, several genes that are altered in cancer converge on few molecular mechanisms that are commonly involved in tumorigenesis (Sanchez-Vega et al., 2018). 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

*Equal contribution ¹Department of Computer and Information Science, School of Engineering and Applied Science, ²Department of Biochemistry and Biophysics, ³Department of Genetics, ⁴Department of Pathology and Laboratory Medicine, ⁵Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. ⁶Division of Cancer Pathobiology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. Correspondence to: Anupama Jha <anupamaj@seas.upenn.edu>, Mathieu Quesnel-Vallières <mathieu.quesnel-vallieres@pennmedicine.upenn.edu>, Yoseph Barash <yosephb@pennmedicine.upenn.edu>.

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 (Paull et al., 2021). 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 to train computational models to distinguish tumor from healthy 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 in two or more categories (Dave et al., 2004; Roessler et al., 2010; Wei et al., 2014; Ahn et al., 2018; Grewal et al., 2019; Frost et al., 2019). While these studies show that machine learning models can successfully distinguish between healthy tissues and tumors given a certain set of conditions, the genes used to build these networks are pre-selected to lower the number of biological features used as input and make the training of deep neural network models easier. Although this pre-selection facilitates the training of machine learning models (Blum & Langley, 1997; Guyon & Elisseeff, 2003; Liu & Motoda, 2012), pre-selecting genes with known functions in the literature or genes already known to be differentially expressed in cancer also deprives the models from learning about potentially novel genes contributing to the transcriptomic signature of cancer.

In addition, while previous studies focused their analyses on the identification of tumor markers, 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. In particular, we recently described an enhanced integrated gradients method for deep neural network interpretation (Jha et al., 2020) that generates attribution values as a measure of the weight or importance of each biological input feature in the model. For example, we recently used enhanced integrated gradients to find splicing events that are differentially included in brain compared to other tissues without prior knowledge of splicing variations (Jha et al., 2020).

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 healthy 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 healthy and cancer tissues

We aimed to uncover transcriptomic features that commonly define cancer state. Performing differential gene expression analysis on 11 healthy tissue-tumor pairs from GTEx and TCGA and then looking at the overlap in the genes that are deregulated between these analyses shows that no protein-coding gene is consistently up- or down-regulated ($\text{abs}(\log_2\text{FC}) > 2$, adjusted p-value < 0.01) in more than 9 cancer types (Figure 1A). In addition, while hallmark oncogenes are expected to be disrupted in many cancer types, we observe in a sampling of 11 oncogenes they are often not significantly differentially expressed in any of the 11 cancer analyzed (e.g. BCR, CTNNB1, DDX6, FUS, KRAS, MDM2, TPR) or only disrupted in certain cancer types (e.g. EGFR, ETV4, JUN, MYC; Figure S1, panel A). Although the 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 independent of changes in transcript levels, these results demonstrate that using a simple differential gene expression analysis fails to grasp the complexity and heterogeneity of the transcriptomic variations existing across various cancer types.

In order to overcome these limitations and identify what transcriptomic features are important in the determination of cancer state, we sought to train interpretable deep learning models capable of distinguishing between healthy and cancer samples. We assembled a large RNA-Seq dataset comprising 13461 samples from 19 healthy tissue types and 18 cancer types (Figure 1B). Samples were sourced from TCGA (<https://www.cancer.gov/tcga>), GTEx (Lonsdale et al., 2013) and 12 other datasets (Figure 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 healthy tissue samples from the same donors. These additional datasets allow us to mitigate dataset-specific biases and focus on cancer-specific signals by performing a

Figure 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 from hypergeometric tests.

tissue/tumor-specific mean correction across the 14 datasets (see Figure S1, panel B).

We first used expression data from 19657 protein-coding genes to train an autoencoder for dimensionality reduction, followed by a supervised neural network to predict cancer

state (healthy tissue versus tumor, Figure 1D; see Methods for details about dimensionality reduction and model training). To ensure that our model did not learn dataset-specific biases, we evaluated model performance on a previously unseen set of 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 (healthy and tumor lung samples). Our protein-coding gene expression model accurately predicted whether an individual sample corresponds to a healthy tissue or a tumor (accuracy: $98.6\% \pm 0.2\%$, precision: $98.8\% \pm 0.4\%$ and recall: $98.8\% \pm 0.1\%$, Figure 1E) and this performance generalizes over all 13 datasets despite the class imbalance between datasets (Figure 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 examples (PRJNA340880 has 1 sample and PRJNA288518 has 3 samples). Importantly, the model performs almost as well when applied on the independent dataset (Figure 1G).

To evaluate how our model generalizes to cancer types not included in the training set, we assembled a group of healthy (T and B lymphocytes) and malignant (acute myeloid leukemia and acute lymphoblastic leukemia) blood cells from three datasets (ArrayExpress E-MTAB-2319, Blueprint and TARGET consortia, see Supplementary Table 1 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 test set) and technical levels (batch-corrected in training set vs. uncorrected test set). Strikingly, despite these vast differences between training and test sets, our deep neural network model successfully distinguishes healthy and cancer samples from blood (Figure 1H), although at lower accuracy, as expected due to the biological and technical differences between the training and test sets.

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 healthy and cancer lung samples. Under these conditions, all three models perform similarly (Supplementary Table 7A). 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 (Figure 1H). These results demonstrate that a deep neural network trained on protein-coding gene expression profiles can robustly predict whether a sample comes from a healthy tissue or a tumor, even when tested with tissue types not previously seen, and that deep neural networks perform better than other algorithms for this task.

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 (Ching et al., 2016; Bolha et al., 2017; Brinkman, 2004). In addition, a small number of mutations located in lncRNA genes or disrupting splicing in protein-coding genes have been shown to drive cancer (Carlevaro-Fita et al., 2020). 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 healthy and tumor samples, similarly 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 14257 lncRNA genes or splice site usage data from 40147 splice junctions. Remarkably, these models achieved $98.6\% \pm 0.2\%$ and $98.8\% \pm 0.1\%$ accuracy, respectively, with high precision ($98.5\% \pm 0.2\%$ for lncRNA expression and $98.7\% \pm 0.1\%$ for splice junction usage) and recall ($99.2\% \pm 0.1\%$ for lncRNA expression and $99.2\% \pm 0.2\%$ for splice junction usage, Figure 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 class imbalance (Figure 1F), or on an independent dataset (Figure 1G), showing that our models predict classes based on biological signal and not confounding factors.

As for the protein coding gene 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 healthy and tumor lung samples (Supplementary Table 7B). However, while our deep neural network model also correctly predicted cancer state with the hematologic dataset (healthy lymphocytes and leukemia samples), both Support-Vector machine and Random Forest completely failed again on this dataset (Supplementary Table 8). On the splicing data, our deep neural network model outperforms both the Support-Vector machine and Random forest models on the independent batch-corrected healthy and tumor lung dataset (Supplementary Table 7C). 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) (Jha et al., 2020). Briefly, EIG measures a feature's contribution, either positive or negative, to the model label predictions (healthy tissue versus cancer) when comparing a cancer sample to a baseline. Following our previous work (Jha et al., 2020), we used the median of normal samples as the baseline (see Methods section "Interpretation of tumor classification models" for details).

We selected 1768 protein-coding genes, 1763 lncRNAs and 562 splice junctions that have high median attribution values across tumor types (Figure 2A; see Methods section "Selection of feature sets" for selection criteria and Figure S2). 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. 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 highest median attribution values across all tumor samples have consistently high attribution values in most if not all cancer types (Figure 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 values does not necessarily reflect the change in expression in cancer. In other words, a gene with a high positive attribution value won't necessarily be upregulated in all or most cancers, and conversely, a gene with a high negative attribution value won't 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. Since the models are not causal, 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 at-

tribution and TSGs receiving negative values (Figure 2C). However, most the known oncogenes and TSGs have moderate attribution values relative to our top-scoring features, with many having neutral attribution values (close to 0). This result was observed for both the expression of COSMIC genes or usage of splice junctions found of those genes. Compared to our criteria for selection of high-attribution features we only observe a small, although statistically significant, enrichment for COSMIC genes among our high negative attribution genes (Figure 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 (Figure 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.

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 (Figure 3A). 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 (Figure 3B). 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 loss-of-function observed/expected upper bound fraction (LOEUF) score (Karczewski et al., 2020) (Figure 3C).

Finally, we inferred the functional impact of lncRNA genes with high attributions by looking at their density of a class of DNA motifs termed pyknons. Pyknons are located in loci that were previously reported as often differentially transcribed between healthy and colorectal cancer tissues and that can affect the oncogenic functions of lncRNAs (Rigoutsos et al., 2006; 2017; Evangelista et al., 2020). We found that high-attribution lncRNA genes carry a higher density of pyknons (Figure 3D) than lncRNA genes from

Figure 3. (A) Evolutionary conservation of protein-coding genes (left panel), genes with variable splice junctions (middle panel) or lncRNA genes (right panel) in 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 (Karczewski et al., 2020), which corresponds to the frequency at which loss-of-function mutations are observed in the human population for a given gene, showing score for protein-coding genes (left panel) or genes with splice junctions whose usage we used for model training (right panel). A low LOEUF score means high selective pressure against loss of function. (D) Pyknon density in lncRNA genes whose expression we used for model training. All statistics: unpaired t-test.

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.

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 (Figure S3, panel A). We also looked at predicted disorderiness of the peptide sequence corresponding to the two exons immediately upstream and downstream of variable splice junctions and found that the predicted peptide disorderiness level is no different in high-attribution junctions from what we observe in the neutral set (Figure S3, panel B).

We then verified 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 Converged 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 (Figure 4). While genes contributing to PKC signaling have been implicated in cancer as oncogenes or tumor suppressors (Garg et al., 2013), 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 alterations of cancer signaling protein domains.

Frequency of genetic alterations in transcriptomic features characterizing cancer state

Next, we wondered if 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 looking at 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 (Figure 5A). 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 (Figure 5B). 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 (Figure 5C). Similarly, the frequency at which high-attribution genes are involved in amplification (Figure 5D) or deletion events (Figure 5E) is not significantly different from the neutral sets or the COSMIC 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.

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 (Figure 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 (Figure S4, panel A). 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 (Figure S4, panel B). Similar to protein-coding genes with high positive attribution values, genes with splice junctions with high 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 (Figure S4, panel C). Of note, the set of genes

Figure 4. Protein domains that are affected by at least two splice junctions with high attributions.)

Figure 5. Fraction of genes with driver mutations (A), or frequency of passenger mutations (B), structural variants (C), amplification events (D) or deletion events (E) in the TCGA cohort for COSMIC oncogenes, COSMIC tumor-suppressor genes (TSGs), genes with high positive ("Protein-coding positive"), high negative ("Protein-coding negative") or neutral attribution values by expression ("Protein-coding neutral"), or genes with splice junctions with high ("Splicing high") or neutral ("Splicing neutral") attribution values in our models. One-way ANOVA with Tukey post-hoc tests.

with neutral attribution values by expression are enriched for heterogeneous and unrelated terms (Figure S4, panel D) 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 (Figure 6B). The core network 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 (Figure 6B). Ingenuity Pathway Analysis for molecular and cellular functions associated with high-attribution genes confirmed that functions of high negative attribution genes are often distinct from those of high positive attribution genes, with transcription and RNA processing being predominant, respectively (Figure 6C). Pathway analysis also revealed how high negative attribution genes by expression and genes with high attribution by splice junction usage 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

Figure 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).

negative attribution genes for KRAS signaling (Figure 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 healthy 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

Figure 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; Supplementary Table 9 lists the molecules found in each network (node).

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, genes with high negative attribution values comprised 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 (Hanahan & Weinberg, 2000; 2011). 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 (Goodall & Wickramasinghe, 2020). 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 (Anczuków & Krainer, 2016; Marabti & Younis, 2018; Sveen et al., 2015), 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 (Figure 7, Supplementary Table 9).

Our transcriptomic definition of cancer includes several ele-

ments 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 *DYNCH1* (Gong et al., 2019), *WSB1* (Kim et al., 2016; Cao et al., 2015; Kim et al., 2015), *RUFY3* (Xie et al., 2017; Wang et al., 2015), *DOCK5* (Frank et al., 2016), *MYSM1* (Li et al., 2017), *DSE* (Liao et al., 2018), *DCUNID5* (Bommeljé et al., 2013; Guo et al., 2012), *SARNP* (Kang et al., 2019) and *FNTA* (Tian et al., 2019; Chen et al., 2020a), 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 (Garg et al., 2013), DEAD-like (Fuller-Pace, 2013) and RING (Lipkowitz & Weissman, 2011) 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 (Zhang et al., 2019; Hong et al., 2018; Gray et al., 2014; Wöss et al., 2019; Ye & Field, 2012; Cheng et al., 2018). However, splicing variations related to cancer have not been reported for any of these genes except *PAK1*, for which a JMJD6-regulated exon inclusion altering the PKC domain enhances MAPK signaling in melanoma (Liu et al., 2017). While the splice junction we identified differs from the one reported before, it also affect the PKC

domain of PAK1. *IRAK1* has two well-characterized alternative splicing events (Jain et al., 2014), but there exists no evidence that these events are directly involved in tumorigenesis, and they also differ from our splice junctions with high attributions.

Interestingly, while 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, variant frequency is much lower for genes with high positive attributions by expression or splicing, which could explain why many of these transcriptomic features have previously been overlooked in genomic studies.

Future work should be directed at assessing whether these and additional novel transcriptomic features of cancer that we highlight here are causally involved in tumorigenesis, thereby highlighting a core transcriptomic component of cancer development, or whether they represent the downstream consequences of genetic alterations driving cancer, and how these variations relate to core transcriptional circuitries of cancer (Chen et al., 2020b).

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 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 healthy 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 $\text{attr} \in \mathbb{R}^p$, with each value representing how each of the p features contributes to the prediction $F(x)$.

Datasets

In this work, we process RNA-seq samples from healthy human tissues and tumors from 14 datasets (see Supplementary Table 1 for a list of all samples and their tissue/cancer identity and figure 1A and 1B for the number of samples in each dataset and their cancer state). The two largest datasets among these are from the Genotype-Tissue Expression (GTEx) consortium and the Cancer Genome Atlas (TCGA) program. We processed 7665 samples from 19

healthy tissues from GTEx and 4777 tumor samples from 18 cancer types from TCGA. Since large datasets often suffer from batch effects, we included 12 other datasets in our analysis (Ju et al., 2012; Seo et al., 2012; Lee et al., 2016; Kirby et al., 2016; Ooi et al., 2016; Yang et al., 2017; Yun et al., 2017; Lin et al., 2019; Corona et al., 2020; Qin et al., 2020). Two of these datasets included lung tumor samples with matched healthy lung samples, two included liver tumor samples with matched healthy liver samples, one included stomach tumor samples with matched healthy stomach samples, one included breast tumor samples with matched healthy breast samples, one included brain tumor samples with matched healthy brain samples, one included colon tumor samples with matched healthy colon samples, one included only head and neck cancer samples, one included only pancreatic tumor samples, one included only ovarian tumor samples and one included only prostate tumor samples. 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 (Fernández et al., 2016), 35 CD4⁺ and 14 CD8⁺ lymphocyte samples from dataset E-MTAB-2319 (Ranzani et al., 2015), and 40 B-cell acute lymphoblastic leukemia samples and 40 acute myeloid leukemia from the pediatric TARGET cohort (Ma et al., 2018).

RNA-Seq data processing

In order to minimize the introduction of technical biases, all RNA-Seq samples were processed from fastq files (legacy TCGA) in the same manner. The raw reads from RNA-Seq experiments are passed through quality control using FastQC (Andrews et al., 2010). Sequencing adapters were trimmed with TrimGalore (v0.6.6) (Krueger, 2018), reads were aligned with STAR (v2.5.2a) (Dobin et al., 2013) against the hg38 human genome assembly (Hubbard et al., 2002), and mapped reads were sorted and indexed using samtools (v1.11). Gene expression quantification was carried out using Salmon (v0.14.0) in quasi-mapping mode using an index generated with Ensembl GRCh38 transcriptome release 94. Gene expression estimates (transcripts per million) are generated from the STAR output using Salmon (Patro et al., 2017). Transcripts per million (TPM) is a normalization method for RNA-Seq data described as:

$$TPM_i = 10^6 \times Z \times \frac{c_i}{\ell_i N} \quad (1)$$

$$Z = \sum_i \tau_i \ell_i$$

where TPM_i is transcript per million for transcript i , c_i is number of reads mapped to transcript i , N is the total number of mappable reads, ℓ_i is the length of transcript i , τ_i is the expression level (relative abundance) of transcript i , and Z is the normalization factor (mean length of expressed

transcripts) (Dewey et al., 2019). Most genes have more than one transcript and we summed all of the transcript TPMs into a single TPM value per gene (Love et al., 2017). Splice junctions were quantified using MAJIQ (v2.1) (Vaquero-Garcia et al., 2016). MAJIQ defines alternative splicing in terms of local splicing variations (LSVs). LSVs can be binary or complex. Binary LSVs comprise of 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 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 (Vaquero-Garcia et al., 2016).

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 healthy tissues and tumor samples. This step is necessary as our signal of healthy tissues versus tumor is confounded with whether the sample came from GTEx or TCGA. GTEx consists of healthy tissues and TCGA consists of mostly tumor samples, therefore, we added 12 other small datasets to ensure that we learn cancer-specific signal and are not confounded by dataset-specific technical biases. Next, we correct dataset bias for gene expression data by mean-centering each tissue/tumor-type separately across datasets. Batch correction step ensures our deep learning model using protein-coding and lncRNA gene expression generalizes across multiple datasets.

Tumor classification models

We train three deep learning models for defining the transcriptomic signature of solid tumor samples. Each of these models use 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 19657 protein coding genes. We extracted the protein-coding genes from Ensembl BioMart (see Supplementary Table 2 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 healthy 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 ReLU activation function and are trained using the adam optimizer (see Supplementary Table 3 for the detailed architecture of both models).

lncRNA gene expression based model

For the next model, the input features are the gene expression of 14257 lncRNA genes. We extract the lncRNA genes from Ensembl BioMart, see Supplementary Table 4 for the list of lncRNA genes. The model architecture and training process of the lncRNA genes based model is similar to the protein-coding gene expression model described in the previous section (see Supplementary Table 5 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 40147 alternative splice junctions from 11219 genes. We generated the splicing quantification for these junctions in the healthy tissues and tumors using MAJIQ (see 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 both these models is similar to the previous gene expression (protein-coding or lncRNA) based models (see Supplementary Table 6 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 method for interpretation of a deep learning model. Here interpretation means attributing the prediction of a deep learning model to its input features. Enhanced integrated gradients 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 model's proxy to human counterfactual intuition. This implies that humans assign blame for difference in two entities on attributes that are present in one entity but absent in the other. Enhanced integrated gradients offers multiple baselines and paths. Since we want to find features that distinguish between tumor samples from healthy tissue samples, we use healthy tissues as the baseline class. Using median baseline from the healthy tissues class, we compute the attributions for the tumor samples by aggregating the gradients between the baseline and each tumor sample along a linear path in the original feature space. We assess 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 healthy 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 (Jha et al., 2020).

Selection of feature sets

High-attribution features were selected on the basis of their FDR-corrected (1%) p-value < 0.0001 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 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 (Chen et al., 2013) 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 (Mootha et al., 2003; Subramanian et al., 2005)

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 .

Functional characterizations

As for splice junction characterization, protein sequence used was translated from the sequence of the longest coding transcript in RefSeq. The conservation score was calculated from the summation of BLAST bit score from 6 species: human, chimpanzee, mouse, cattle, xenopus, zebrafish and chicken (taxids: 9606, 9598, 10090, 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 (Karczewski et al., 2020). Pyknon density was calculated using the list of human pyknons available from the pyknon database (Tsirigos & Rigoutsos, 2008) and shown as the number of pyknons found per 1,000 nt in the longest RefSeq annotated transcript.

Acknowledgements

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

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 figures will be available at <https://bitbucket.org/biociaphers/pan-cancer> and be deposited in Zenodo before publication. RNA-Seq samples used for this study are listed in Supplementary Table 1.

Author's 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.

Figure S1. (A) Differential gene expression of four oncogenes in 11 healthy tissue-cancer pairs. NA: not differentially expressed. (B) UMAP showing the gene expression of protein coding genes in healthy lung samples in four datasets (GTEx, TCGA, PRJEB2784 and PRJEB3132) before (left) and after (right) batch correction. (C) Left: UMAP embedding of the protein-coding autoencoder's latent layer representation of the cancer and normal samples along with the median baseline points from the normal class. Right: UMAP embedding of protein-coding DNN's layer layer representation of the cancer and normal samples along with the median baseline points from the normal class.

A

Figure S2. Distribution of transcriptomic features ranked from highest positive values to lowest negative values from models trained on protein-coding gene expression (top), lncRNA expression (middle) or splice junction usage data (bottom). Knee-point of the curves were determined using KneeLocator of the kneed package and are shown with broken horizontal lines.

Figure S3. (A) Fraction of splicing variations where reading frame is preserved in both isoforms. (B) Predicted level of peptide disorderiness around variable splice junctions.

Figure S4. Gene ontology analysis of genes with high negative (A) or high positive (B) attribution values by expression, genes with splice junctions that have high attribution values (C), or genes with neutral attribution values by expression (D). Circle size and values indicate the fraction of genes corresponding to a GO term present in the gene set analyzed. Only top 20 terms are shown if more terms meet our cutoff of adjusted p-value < 0.05.

References

- Ahn, T., Goo, T., hee Lee, C., Kim, S., Han, K., Park, S., and Park, T. Deep learning-based identification of cancer or normal tissue using gene expression data. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, dec 2018. doi: 10.1109/bibm.2018.8621108.
- Anczuków, O. and Krainer, A. R. Splicing-factor alterations in cancers. *RNA*, 22(9):1285–1301, aug 2016. doi: 10.1261/rna.057919.116.
- Andrews, S. et al. Fastqc: a quality control tool for high throughput sequence data, 2010.
- Blum, A. L. and Langley, P. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2):245–271, 1997.
- Bolha, L., Ravnik-Glavač, M., and Glavač, D. Long non-coding RNAs as biomarkers in cancer. 2017:1–14, 2017. doi: 10.1155/2017/7243968.
- Bommeljé, C. C., Weeda, V. B., Huang, G., Shah, K., Bains, S., Buss, E., Shaha, M., Gönen, M., Ghossein, R., Ramanathan, S. Y., and Singh, B. Oncogenic function of SCCRO5/DCUN1d5 requires its neddylation e3 activity and nuclear localization. *Clinical Cancer Research*, 20(2):372–381, nov 2013. doi: 10.1158/1078-0432.ccr-13-1252.
- Brinkman, B. M. Splice variants as cancer biomarkers. 37 (7):584–594, jul 2004. doi: 10.1016/j.clinbiochem.2004.05.015.
- Cao, J., Wang, Y., Dong, R., Lin, G., Zhang, N., Wang, J., Lin, N., Gu, Y., Ding, L., Ying, M., He, Q., and Yang, B. Hypoxia-induced WSB1 promotes the metastatic potential of osteosarcoma cells. *Cancer Research*, 75(22):4839–4851, sep 2015. doi: 10.1158/0008-5472.can-15-0711.
- Carlevaro-Fita, J., Lanzós, A., Feuerbach, L., Hong, C., Mas-Ponte, D., Pedersen, J. S., and and, R. J. Cancer LncRNA census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. *Communications Biology*, 3(1), feb 2020. doi: 10.1038/s42003-019-0741-7.
- Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., Clark, N. R., and Ma’ayan, A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):128, 2013. doi: 10.1186/1471-2105-14-128.
- Chen, J., Sun, Y., Ou, Z., Yeh, S., Huang, C.-P., You, B., Tsai, Y.-C., jen Sheu, T., Zu, X., and Chang, C. Androgen receptor-regulated circ FNTA activates KRAS signaling to promote bladder cancer invasion. *EMBO reports*, 21 (4), feb 2020a. doi: 10.15252/embr.201948467.
- Chen, Y., Xu, L., Lin, R. Y.-T., Müschen, M., and Koeffler, H. P. Core transcriptional regulatory circuitries in cancer. *Oncogene*, 39(43):6633–6646, sep 2020b. doi: 10.1038/s41388-020-01459-w.
- Cheng, B. Y., Lau, E. Y., Leung, H.-W., Leung, C. O.-N., Ho, N. P., Gurung, S., Cheng, L. K., Lin, C. H., Lo, R. C.-L., Ma, S., Ng, I. O.-L., and Lee, T. K. IRAK1 augments cancer stemness and drug resistance via the AP-1/AKR1b10 signaling cascade in hepatocellular carcinoma. *Cancer Research*, 78(9):2332–2342, feb 2018. doi: 10.1158/0008-5472.can-17-2445.
- Ching, T., Peplowska, K., Huang, S., Zhu, X., Shen, Y., Molnar, J., Yu, H., Tiirikainen, M., Fogelgren, B., Fan, R., and Garmire, L. X. Pan-cancer analyses reveal long intergenic non-coding RNAs relevant to tumor diagnosis, subtyping and prognosis. *EBioMedicine*, 7:62–72, may 2016. doi: 10.1016/j.ebiom.2016.03.023.
- Corona, R. I., Seo, J.-H., Lin, X., Hazelett, D. J., Reddy, J., Fonseca, M. A., Abassi, F., Lin, Y. G., Mhawech-Fauceglia, P. Y., Shah, S. P., et al. Non-coding somatic mutations converge on the pax8 pathway in ovarian cancer. *Nature communications*, 11(1):1–11, 2020.
- Dave, S. S., Wright, G., Tan, B., Rosenwald, A., Gascoyne, R. D., Chan, W. C., Fisher, R. I., Braziel, R. M., Rimsza, L. M., Grogan, T. M., et al. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *New England Journal of Medicine*, 351(21):2159–2169, 2004.
- Dewey, C., Craven, M., and Gitter, A. Measuring transcriptomes with rna-seq, 2019. URL <https://www.biostat.wisc.edu/bmi776/lectures/rnaseq.pdf>.
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.
- Evangelista, A. F., de Menezes, W. P., Berardinelli, G. N., Santos, W. D., Scapulatempo-Neto, C., Guimarães, D. P., Calin, G. A., and Reis, R. M. Pyknon-containing transcripts are downregulated in colorectal cancer tumors, and loss of PYK44 is associated with worse patient outcome. *Frontiers in Genetics*, 11, nov 2020. doi: 10.3389/fgene.2020.581454.
- Fernández, J. M., 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, E. C., and

- Valencia, A. The BLUEPRINT data analysis portal. 3(5): 491–495.e5, nov 2016. doi: 10.1016/j.cels.2016.10.021.
- Frank, S. R., Köllmann, C. P., van Lidth de Jeude, J. F., Thiagarajah, J. R., Engelholm, L. H., Frödin, M., and Hansen, S. H. The focal adhesion-associated proteins DOCK5 and GIT2 comprise a rheostat in control of epithelial invasion. *Oncogene*, 36(13):1816–1828, sep 2016. doi: 10.1038/onc.2016.345.
- Frost, F. G., Cherukuri, P. F., Milanovich, S., and Boerkoel, C. F. Pan-cancer RNA-seq data stratifies tumours by some hallmarks of cancer. *Journal of Cellular and Molecular Medicine*, 24(1):418–430, nov 2019. doi: 10.1111/jcmm.14746.
- Fuller-Pace, F. V. DEAD box RNA helicase functions in cancer. *RNA Biology*, 10(1):121–132, jan 2013. doi: 10.4161/rna.23312.
- Garg, R., Benedetti, L. G., Abera, M. B., Wang, H., Abba, M., and Kazanietz, M. G. Protein kinase c and cancer: what we know and what we do not. *Oncogene*, 33(45): 5225–5237, dec 2013. doi: 10.1038/onc.2013.524.
- Gong, L.-B., Wen, T., Li, Z., Xin, X., Che, X.-F., Wang, J., Liu, Y.-P., and Qu, X.-J. DYNC1i1 promotes the proliferation and migration of gastric cancer by up-regulating IL-6 expression. *Frontiers in Oncology*, 9, jun 2019. doi: 10.3389/fonc.2019.00491.
- Goodall, G. J. and Wickramasinghe, V. O. RNA in cancer. *Nature Reviews Cancer*, 21(1):22–36, oct 2020. doi: 10.1038/s41568-020-00306-0.
- Gray, G. K., McFarland, B. C., Rowse, A. L., Gibson, S. A., and Benveniste, E. N. Therapeutic CK2 inhibition attenuates diverse prosurvival signaling cascades and decreases cell viability in human breast cancer cells. *Oncotarget*, 5(15):6484–6496, jul 2014. doi: 10.18632/oncotarget.2248.
- Grewal, J. K., Tessier-Cloutier, B., Jones, M., Gakkhar, S., Ma, Y., Moore, R., Mungall, A. J., Zhao, Y., Taylor, M. D., Gelmon, K., Lim, H., Renouf, D., Laskin, J., Marra, M., Yip, S., and Jones, S. J. M. Application of a neural network whole transcriptome-based pan-cancer method for diagnosis of primary and metastatic cancers. *JAMA Network Open*, 2(4):e192597, apr 2019. doi: 10.1001/jamanetworkopen.2019.2597.
- Guo, W., Li, G.-J., Xu, H.-B., Xie, J.-S., Shi, T.-P., Zhang, S.-Z., Chen, X.-H., and Huang, Z.-G. In vitro biological characterization of DCUN1d5 in DNA damage response. *Asian Pacific Journal of Cancer Prevention*, 13(8):4157–4162, aug 2012. doi: 10.7314/apjcp.2012.13.8.4157.
- Guyon, I. and Elisseeff, A. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.
- Haigis, K. M., Cichowski, K., and Elledge, S. J. Tissue-specificity in cancer: The rule, not the exception. *Science*, 363(6432):1150–1151, mar 2019. doi: 10.1126/science.aaw3472.
- Hanahan, D. and Weinberg, R. A. The hallmarks of cancer. *Cell*, 100(1):57–70, jan 2000. doi: 10.1016/s0092-8674(00)81683-9.
- Hanahan, D. and Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, mar 2011. doi: 10.1016/j.cell.2011.02.013.
- 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., and Zhang, Z. Targeting posttranslational modifications of RIOK1 inhibits the progression of colorectal and gastric cancers. *eLife*, 7, jan 2018. doi: 10.7554/elife.29511.
- 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 research*, 30(1):38–41, 2002.
- Jain, A., Kaczanowska, S., and Davila, E. IL-1 receptor-associated kinase signaling and its role in inflammation, cancer progression, and therapy resistance. *Frontiers in Immunology*, 5, nov 2014. doi: 10.3389/fimmu.2014.00553.
- Jha, A., Aicher, J. K., Gazzara, M. R., Singh, D., and Barash, Y. Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study. *Genome biology*, 21(1):1–22, 2020.
- Ju, Y. S., Lee, W.-C., Shin, J.-Y., Lee, S., Bleazard, T., Won, J.-K., Kim, Y. T., Kim, J.-I., Kang, J.-H., and Seo, J.-S. A transforming kif5b and ret gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. *Genome research*, 22(3):436–445, 2012.
- Kang, G. J., Park, M. K., Byun, H. J., Kim, H. J., Kim, E. J., Yu, L., Kim, B., Shim, J. G., Lee, H., and Lee, C. H. SARNP, a participant in mRNA splicing and export, negatively regulates e-cadherin expression via interaction with pinin. *Journal of Cellular Physiology*, 235(2):1543–1555, jul 2019. doi: 10.1002/jcp.29073.
- Karczewski, K. J., , Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D.,

- Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zapala, Z., O'Donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., 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, M. E., Neale, B. M., Daly, M. J., and MacArthur, D. G. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, may 2020. doi: 10.1038/s41586-020-2308-7.
- Kim, J. J., Lee, S. B., Jang, J., Yi, S.-Y., Kim, S.-H., Han, S.-A., Lee, J.-M., Tong, S.-Y., Vincelette, N. D., Gao, B., Yin, P., Evans, D., Choi, D. W., Qin, B., Liu, T., Zhang, H., Deng, M., Jen, J., Zhang, J., Wang, L., and Lou, Z. WSB1 promotes tumor metastasis by inducing pVHL degradation. *Genes & Development*, 29(21):2244–2257, nov 2015. doi: 10.1101/gad.268128.115.
- Kim, J. J., Lee, S. B., Yi, S.-Y., Han, S.-A., Kim, S.-H., Lee, J.-M., Tong, S.-Y., Yin, P., Gao, B., Zhang, J., and Lou, Z. WSB1 overcomes oncogene-induced senescence by targeting ATM for degradation. *Cell Research*, 27(2): 274–293, dec 2016. doi: 10.1038/cr.2016.148.
- Kirby, M. K., Ramaker, R. C., Gertz, J., Davis, N. S., Johnston, B. E., Oliver, P. G., Sexton, K. C., Greeno, E. W., Christein, J. D., Heslin, M. J., et al. Rna sequencing of pancreatic adenocarcinoma tumors yields novel expression patterns associated with long-term survival and reveals a role for angptl4. *Molecular oncology*, 10(8): 1169–1182, 2016.
- Krueger, F. Trim galore! babraham bioinformatics, 2018.
- Lee, J.-R., Kwon, C. H., Choi, Y., Park, H. J., Kim, H. S., 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*, 16 (1):1–11, 2016.
- Li, Y., Li, J., Liu, H., Liu, Y., and Cui, B. Expression of MYSM1 is associated with tumor progression in colorectal cancer. *PLOS ONE*, 12(5):e0177235, may 2017. doi: 10.1371/journal.pone.0177235.
- Liao, W.-C., Liao, C.-K., Tsai, Y.-H., Tseng, T.-J., Chuang, L.-C., Lan, C.-T., Chang, H.-M., and Liu, C.-H. DSE promotes aggressive glioma cell phenotypes by enhancing HB-EGF/ErbB signaling. *PLOS ONE*, 13(6):e0198364, jun 2018. doi: 10.1371/journal.pone.0198364.
- Lin, X., Spindler, T. J., de Souza Fonseca, M. A., Corona, R. I., Seo, J.-H., Dezem, F. S., Li, L., Lee, J. M., Long, H. W., Sellers, T. A., et al. Super-enhancer-associated lncrna ucal interacts directly with amot to activate yap target genes in epithelial ovarian cancer. *IScience*, 17: 242–255, 2019.
- Lipkowitz, S. and Weissman, A. M. RINGs of good and evil: RING finger ubiquitin ligases at the crossroads of tumour suppression and oncogenesis. *Nature Reviews Cancer*, 11(9):629–643, aug 2011. doi: 10.1038/nrc3120.
- Liu, H. and Motoda, H. *Feature selection for knowledge discovery and data mining*, volume 454. Springer Science & Business Media, 2012.
- Liu, X., Si, W., Liu, X., He, L., Ren, J., Yang, Z., Yang, J., Li, W., Liu, S., Pei, F., Yang, X., and Sun, L. JMJD6 promotes melanoma carcinogenesis through regulation of the alternative splicing of PAK1, a key MAPK signaling component. *Molecular Cancer*, 16(1), nov 2017. doi: 10.1186/s12943-017-0744-2.
- 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. *Nature genetics*, 45(6):580, 2013.
- Love, M. I., Soneson, C., and Robinson, M. D. Importing transcript abundance datasets with tximport. 2017.
- Ma, X., Liu, Y., Liu, Y., Alexandrov, L. B., Edmonson, M. N., Gawad, C., Zhou, X., Li, Y., Rusch, M. C., Easton, J., Huether, R., Gonzalez-Pena, V., Wilkinson, M. R., Hermida, L. C., Davis, S., Sioson, E., Pounds, S., Cao, X., Ries, R. E., Wang, Z., Chen, X., Dong, L., Diskin, S. J., Smith, M. A., Auvil, J. M. G., Meltzer, P. S., Lau, C. C., Perlman, E. J., Maris, J. M., Meshinchi, S., Hunger, S. P., Gerhard, D. S., and Zhang, J. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. 555(7696):371–376, feb 2018. doi: 10.1038/nature25795.
- Marabti, E. E. and Younis, I. The cancer spliceome: Reprogramming of alternative splicing in cancer. *Frontiers in Molecular Biosciences*, 5, sep 2018. doi: 10.3389/fmolb.2018.00080.
- Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. PGC-1-responsive genes

- p involved in oxidative phosphorylation are coordinately downregulated in human diabetes.
- 34*
- (3):267–273, jun 2003. doi: 10.1038/ng1180.
- Ooi, W. F., Xing, M., Xu, C., Yao, X., Ramlee, M. K., Lim, M. C., Cao, F., Lim, K., Babu, D., Poon, L.-F., et al. Epigenomic profiling of primary gastric adenocarcinoma reveals super-enhancer heterogeneity. *Nature communications*, 7(1):1–17, 2016.
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4): 417, 2017.
- Paull, E. O., Aytes, A., Jones, S. J., Subramaniam, P. S., Giorgi, F. M., Douglass, E. F., Tagore, S., Chu, B., Vasciaveo, A., Zheng, S., Verhaak, R., Abate-Shen, C., Alvarez, M. J., and Califano, A. A modular master regulator landscape controls cancer transcriptional identity. *Cell*, 184(2):334–351.e20, jan 2021. doi: 10.1016/j.cell.2020.11.045.
- Qin, T., Koneva, L. A., Liu, Y., Zhang, Y., Arthur, A. E., Zarins, K. R., Carey, T. E., Chepeha, D., Wolf, G. T., Rozek, L. S., 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*, 42(9):2375–2389, 2020.
- Ranzani, V., Rossetti, G., Panzeri, I., Arrigoni, A., Bonnal, R. J. P., Curti, S., Gruarin, P., Provati, E., Sugliano, E., Marconi, M., Francesco, R. D., Geginat, J., Bodega, B., Abrignani, S., and Pagani, M. The long intergenic non-coding RNA landscape of human lymphocytes highlights the regulation of t cell differentiation by linc-MAF-4. *16* (3):318–325, jan 2015. doi: 10.1038/ni.3093.
- Rigoutsos, I., Huynh, T., Miranda, K., Tsirigos, A., McHardy, A., and Platt, D. Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. *Proceedings of the National Academy of Sciences*, 103(17):6605–6610, apr 2006. doi: 10.1073/pnas.0601688103.
- Rigoutsos, I., Lee, S. K., Nam, S. Y., Anfossi, S., Pasculli, B., Pichler, M., Jing, Y., Rodriguez-Aguayo, C., Telonis, A. G., Rossi, S., Ivan, C., Ivkovic, T. C., Fabris, L., Clark, P. M., Ling, H., Shimizu, M., Redis, R. S., Shah, M. Y., Zhang, X., Okugawa, Y., Jung, E. J., Tsirigos, A., Huang, L., Ferdin, J., Gafà, R., Spizzo, R., Nicoloso, M. S., Paranjape, A. N., Shariati, M., Tiron, A., Yeh, J. J., Teruel-Montoya, R., Xiao, L., Melo, S. A., Menter, D., Jiang, Z.-Q., Flores, E. R., Negrini, M., Goel, A., Bar-Eli, M., Mani, S. A., Liu, C. G., Lopez-Berestein, G., Berindan-Neagoe, I., Esteller, M., Kopetz, S., Lanza, G., and Calin, G. A. N-BLR, a primate-specific non-coding transcript leads to colorectal cancer invasion and migration. *Genome Biology*, 18(1), may 2017. doi: 10.1186/s13059-017-1224-0.
- Roessler, S., Jia, H.-L., Budhu, A., Forgues, M., Ye, Q.-H., Lee, J.-S., Thorgeirsson, S. S., 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 research*, 70(24): 10202–10212, 2010.
- Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., Dimitriadou, S., Liu, D. L., Kantheti, H. S., Saghafein, S., et al. Oncogenic signaling pathways in the cancer genome atlas. *Cell*, 173(2):321–337, 2018.
- Seo, J.-S., Ju, Y. S., Lee, W.-C., Shin, J.-Y., Lee, J. K., Bleazard, T., Lee, J., Jung, Y. J., Kim, J.-O., Shin, J.-Y., et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome research*, 22 (11):2109–2119, 2012.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *102*(43): 15545–15550, sep 2005. doi: 10.1073/pnas.0506580102.
- Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R. A., and Skotheim, R. I. Aberrant RNA splicing in cancer expression changes and driver mutations of splicing factor genes. *Oncogene*, 35(19):2413–2427, aug 2015. doi: 10.1038/onc.2015.318.
- Tian, J., Fan, J., Xu, J., Ren, T., Guo, H., and Zhou, L. circ-FNTA accelerates proliferation and invasion of bladder cancer. *Oncology Letters*, nov 2019. doi: 10.3892/ol.2019.11150.
- Tsirigos, A. and 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 Research*, 36(10):3484–3493, may 2008. doi: 10.1093/nar/gkn155.
- Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., Gonzalez-Vallinas, J., Lahens, N. F., Hogenesch, J. B., Lynch, K. W., and Barash, Y. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *elife*, 5:e11752, 2016.
- Wang, G., Zhang, Q., Song, Y., Wang, X., Guo, Q., Zhang, J., Li, J., Han, Y., Miao, Z., and Li, F. PAK1 regulates

- RUFY3-mediated gastric cancer cell migration and invasion. *Cell Death & Disease*, 6(3):e1682–e1682, mar 2015. doi: 10.1038/cddis.2015.50.
- Wei, I. H., Shi, Y., Jiang, H., Kumar-Sinha, C., and Chinnaiyan, A. M. RNA-seq accurately identifies cancer biomarker signatures to distinguish tissue of origin. *Neoplasia*, 16(11):918–927, nov 2014. doi: 10.1016/j.neo.2014.09.007.
- Wöss, K., Simonović, N., Strobl, B., Macho-Maschler, S., and Müller, M. TYK2: An upstream kinase of STATs in cancer. *Cancers*, 11(11):1728, nov 2019. doi: 10.3390/cancers11111728.
- 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., and Wang, J. RUFY3 interaction with FOXK1 promotes invasion and metastasis in colorectal cancer. *Scientific Reports*, 7(1), jun 2017. doi: 10.1038/s41598-017-04011-1.
- Yang, Y., Chen, L., Gu, J., Zhang, H., Yuan, J., Lian, Q., Lv, G., Wang, S., Wu, Y., Yang, Y.-C. T., et al. Recurrently deregulated lncRNAs in hepatocellular carcinoma. *Nature communications*, 8(1):1–13, 2017.
- Ye, D. Z. and Field, J. PAK signaling in cancer. *Cellular Logistics*, 2(2):105–116, apr 2012. doi: 10.4161/cl.21882.
- Yun, S. J., Kim, S.-K., Kim, J., Cha, E.-J., Kim, J.-S., Kim, S.-J., Ha, Y.-S., Kim, Y.-H., Jeong, P., Kang, H. W., et al. Transcriptomic features of primary prostate cancer and their prognostic relevance to castration-resistant prostate cancer. *Oncotarget*, 8(70):114845, 2017.
- Zhang, Y., kang Yang, W., ming Wen, G., Tang, H., an Wu, C., xia Wu, Y., liang Jing, Z., shan Tang, M., long Liu, G., zhou Li, D., hua Li, Y., and Deng, Y.-J. High expression of PRKDC promotes breast cancer cell growth via p38 MAPK signaling and is associated with poor survival. *Molecular Genetics & Genomic Medicine*, 7(11), sep 2019. doi: 10.1002/mgg3.908.