Disentangling the splicing factor programs underlying complex molecular phenotypes

Miquel Anglada-Girotto[1]*, Daniel F. Moakley[2,3,4], Chaolin Zhang[2,3,4], Samuel Miravet-Verde[5], Andrea Califano[2,3,6,7,8,9]*, Luis Serrano[1,10,11]*

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain.
[2]Department of Systems Biology, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, USA 10032
[3]Department of Biochemistry & Molecular Biophysics, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, USA 10032
[4]Center for Motor Neuron Biology and Disease, Columbia University, New York, USA 10032
[5]Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zurich, Zurich, Switzerland.
[6]Department of Medicine, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, USA 10032
[7]Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, New York, USA 10032
[8]Chan Zuckerberg Biohub New York, New York, NY, USA
[9]Department of Biomedical Informatics, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, USA 10032
[10]Universitat Pompeu Fabra (UPF), Barcelona, Spain.
[11]ICREA, Pg. Lluís Companys 23, Barcelona 08010, Spain.

*corresponding author(s): miquel.anglada@crg.eu, ac2248@cumc.columbia.edu, luis.serrano@crg.eu

ABSTRACT

The regulation of exon inclusion through alternative splicing tunes the cell's behavior by increasing the functional diversity of the transcriptome and the proteome. Splicing factors work in concert to generate gene isoform pools that contribute to cell phenotypes yet their activity is controlled by multiple regulatory and signaling layers. This hinders identification of functional, phenotype-specific splicing factors using traditional single-omic measurements, such as their mutational state or expression. To address this challenge, we propose repurposing the virtual inference of protein activity by enriched regulon analysis (VIPER) to measure splicing factor activity solely from their downstream exon transcriptomic inclusion signatures. This approach is effective in assessing the effect of co-occurring splicing factor perturbations, as well as their post-translational regulation. As proof of concept, we dissect recurrent splicing factor programs underlying tumorigenesis including aberrantly activated factors acting as oncogenes and inactivated ones acting as tumor suppressors, which are undetectable by more conventional methodologies. Activation and inactivation of these cancer splicing programs effectively stratifies overall survival, as well as cancer hallmarks such as proliferation and immune evasion. Altogether, repurposing network-based inference of protein activity for splicing factor networks distills common, functionally relevant splicing programs in otherwise heterogeneous molecular contexts.

Keywords
cancer; alternative splicing; splicing factor; protein activity; VIPER; tumorigenesis.

INTRODUCTION

Splicing factors carry out alternative splicing; namely, they selectively regulate intron removal from precursor mRNAs, leading to synthesis of distinct protein isoforms. On an individual gene basis, splicing factor activity determines the pool of associated mature mRNAs, which impacts overall mRNA stability, translation, protein activation, and protein-protein interactions[1]. Splicing factor activity is thus regulated by multiple mechanisms, including transcription of their encoding genes, pre-mRNA splicing, protein synthesis and degradation, post-translational modifications, subcellular localization, and formation of functional complexes with other proteins (Fig. 1a). Accurate assessment of splicing factor activity is thus challenging as it requires accounting for a highly diverse mechanism repertoire.

Splicing factor activities have been typically inferred based on the presence of activating and inactivating somatic mutations or differential gene expression[2–5]. However, these represent just two of the many mechanisms contributing to splicing factor activity modulation. This limited scope may thus provide only partial understanding of the splicing factor repertoire that is aberrantly dysregulated in a complex genetic environment, such as cancer, where various molecular alterations and signals converge to manifest common cancer hallmarks[6,7]. Therefore, accurately assessing splicing factor activity from multi-omic data not only requires large-scale efforts and resources for data generation but also key mechanistic knowledge of how these data may help quantify splicing factor activity, which is still quite elusive.

The problem of assessing protein activity, independent of the expression of its encoding gene, has been effectively tackled for transcription factors (see Garcia-Alonso *et al.*[8] for a recent review). Rather than focusing on the expression of a transcription factor to estimate its activity, we can consider that the interplay of the complex regulatory and signaling mechanisms controlling its activity is ultimately reflected by changes in the expression of the genes it regulates. Exploiting this strategy in a proteome-wide fashion, however, requires addressing two challenges: first assessing the transcriptional targets of every transcription factor, which are generally different in different cellular contexts, and, second, deconvoluting their activity based on the differential expression of these targets. Transcription factor→target interactions (i.e., regulatory networks) can be inferred computationally from gene expression data sets by reverse engineering algorithms such as ARACNe[9,10], as well as empirically, by mining curated databases generated from publicly available experimental data[11–13]. Then, given a high-quality regulatory network that also informs on the strength (likelihood) and directionality (mode of regulation) of each interaction, the differential activity of a transcription factor can be accurately assessed from the differential expression of its transcriptional targets, through bioinformatic methods like VIPER[14] and others[15]. Although this approach only requires transcriptomic data, it has been proven highly valuable in disentangling the common functional states of transcriptional regulatory circuits in heterogeneous molecular contexts like cancer[6].

We propose that similar to transcription factors, splicing factor activity should also be measurable based on changes in the exon splicing repertoire they regulate. This suggests that methods like VIPER may be effectively extended to infer splicing factor activity from the differential exon inclusion repertoire of their target mRNAs, represented as a comprehensive network of splicing-factor→pre-mRNA exon interactions (splicing factor network, for simplicity) (Fig. 1b). Unfortunately, however, most methods for splicing factor network

deconvolution focus on RNA-binding factors[16,17] rather than on all proteins annotated as splicing factors, thus failing to provide the comprehensive networks necessary to extend the VIPER algorithm to obtain an exhaustive picture of splicing factor activity.

Using several experimental datasets, where splicing factors are individually silenced, we demonstrate that splicing factor networks can be effectively reverse-engineered and used to extend the VIPER framework to accurately estimate splicing factor activity. In particular, applying this framework to study cancer-related phenotypes revealed two biologically relevant and complementary splicing programs that have eluded more conventional approaches based on differential splicing factor expression analysis. Altogether, VIPER-based splicing factor activity analysis represents a valuable methodology to obtain a more comprehensive picture of the splicing factors that represent mechanistic determinants of complex cellular phenotypes.

RESULTS

Accurate estimation of splicing factor activities

Traditionally, splicing factor networks have been reverse-engineered by measuring the pairwise association between splicing factor expression and exon inclusion across multiple samples[18]. We thus attempted extending the ARACNe algorithm[9,10] to potentially remove indirect interactions, based on a dataset of 313 transcriptomic profiles quantifying gene expression and exon inclusion of healthy human tissues[19]. This resulted in a splicing factor network for a consensus list of 509 splicing factors documented in landmark literature (Supplementary Fig. 1a; Supplementary Table 1; see Methods)[20–24].

To evaluate the quality of the inferred network, we used the VIPER algorithm to predict differential splicing factor activity from a dataset of 1,265 exon-level transcriptomes generated following perturbation (silencing and overexpression) of individual splicing factors (Supplementary Fig. 1b; see Methods). Specifically, we computed within- and across-sample recall scores to assess the ability of the analysis to recapitulate which splicing factors had been perturbed in each specific sample, vs. which samples were targeted by a perturbation of the same splicing factor, respectively (see Methods). Surprisingly, using the ARACNe-based network produced splicing factor activity inferences that were close to random, suggesting that assessing pairwise statistical dependency between splicing factor expression and individual exon inclusion may not represent an effective approach to reverse engineer splicing factor networks (Fig. 1c). To potentially address this limitation, we inferred splicing factor networks through multivariate linear regression (MLR) models that simultaneously associate the gene expression of all splicing factors with the inclusion of each exon (see Methods). However, such an approach also performed poorly with near-random recall distributions (Fig. 1c).

Given the failure of computational approaches that had proven successful in reverse engineering transcription factor networks, we decided to leverage the perturbational datasets available for a subset of splicing factors and derive their regulatory networks empirically. Specifically, this led to effective experimental assessment of the exon inclusion repertoire associated with 247 of 509 proteins annotated as splicing factors. We nominated splicing factor→exon interactions for any exon whose inclusion changes more than a range of

thresholds (from 5 to 45%) following perturbation of each specific splicing factor (see Methods). A 15% threshold enables accurate estimation for the maximum number of splicing factors (n = 243) while limiting the total number of target exons (n = 21,934) (Supplementary Fig. 1c). These empirical splicing factor networks dramatically outperformed the computational ones, resulting in extremely high recall metrics, both within- and across-samples (median recall = 0.99 for both conditions) (Fig. 1c). To avoid overestimating recall due to overfitting, we evaluated the performance of these empirical networks using completely independent perturbational signatures that had not been used in network reconstruction. Notably, these networks identified a median number of 300 exons regulated by each splicing factor, where core splicing factors show higher connectivity than RBPs and other splicing factors (Fig. 1d). A median number of 2 splicing factors regulate each exon (Fig. 1e). Furthermore, as was also reported for the assessment of transcription factor activity using VIPER, assessment of splicing factor activity was highly robust to pruning the exon repertoire. Indeed, reducing the number of target exons by up to 77.6% did not significantly affect within-sample and across-sample recall (Supplementary Fig. 1d).

Given the quality of predictions based on the experimental splicing factor networks, we wondered what may be causing the poor performance of computational approaches. Interactions in splicing factor networks are associated with both a likelihood score in the range [0, 1] and a mode of regulation (positive or negative) (Fig. 1b). Interestingly, the computational methods assigned higher likelihoods to splicing factor-exon interactions that were experimentally observed, suggesting that this may not be the most likely source of poor performance (Supplementary Fig. 1e). On the other hand, the mode of regulation was inferred virtually at random with a prediction error close to 50% (Supplementary Fig. 1f). To confirm this observation, we tested whether combining either ARACNe or MLR-based likelihoods with the experimentally assessed mode of regulation may improve performance. Indeed, this approach increased recall to virtually match that of the experimental networks (i.e., median recall within samples: ARACNe = 0.97, MLR = 0.84, Empirical = 0.99; median recall across samples: ARACNe = 0.98, MLR = 0.86, Empirical = 0.99) (Supplementary Fig. 1g; see Methods).

ARACNe and MLR-based prediction of positive and negative mode of regulation (i.e., exon inclusion vs. exclusion), respectively, were only slightly better than random (Supplementary Fig. 1f). Moreover, integrating the predictions of these algorithms (see Methods) did not further improve performance (Supplementary Fig. 1h). This suggests that correlation-based methods are limited in assessing exon inclusion versus exclusion, leaving some margin for future improvements. Henceforth, unless stated otherwise, all splicing factor activities inferred with VIPER use empirical splicing factor networks.

VIPER-assessed splicing factor activity recapitulates complex perturbations

Similar to its application to assessing transcription factor activity, we propose that extending VIPER to leverage exon inclusion/exclusion networks should be effective in recapitulating differential splicing factor activity, as determined by established and even potentially novel mechanisms.

For example, the drug Indisulam downregulates the protein level of splicing factor RBM39 by mediating its proteasomal degradation (Fig. 2a and b)[25,26]. In contrast, the mRNA levels of its

encoding gene increase consistently upon treatment, probably as a compensatory mechanism (Supplementary Fig. 2)[25,26]. In this case, using changes in splicing factor mRNA levels to evaluate splicing factor activity would lead to the incorrect conclusion that Indisulam increases RBM39's activity. In sharp contrast, VIPER analysis of transcriptional profiles following drug perturbation identifies RBM39 as the splicing factor with the greatest decrease in splicing activity, thus effectively recapitulating Indisulam's mechanism of action (Fig. 2c; see Methods). As a result, VIPER-based analysis recapitulated protein-level effects, despite using only transcriptomic data.

Splicing factors are often modulated in concert, leading to complex, intermingled splicing signatures. To evaluate whether our approach could recapitulate the effect of such combinatorial perturbations, we used VIPER to analyze the transcriptomic data from four different studies where multiple splicing factors were simultaneously silenced[27–30] (Fig. 2d) (see Methods). Confirming the value of the proposed methodology, the silenced splicing factors were always among the most inactivated. This suggests that combinatorial splicing factor perturbations produce additive changes in exon inclusion thus allowing effective assessment of splicing factor activity by the VIPER approach (Fig. 2d).

Splicing factors interact to form a large and dynamic protein complex called the spliceosome[24]. We thus hypothesized that the VIPER-assessed activity of splicing factors interacting in the spliceosome complex via protein-protein interactions (PPIs) should be significantly correlated. At the beginning of the splicing reaction, during splice site recognition, the spliceosome forms the E, A, and B complexes comprising small ribonucleoprotein (snRNP) complexes (U1, U2, U4, U5 and U6)[22]. Within the U2 snRNP, the SF3b complex (SF3B1, SF3B2, SF3B3, SF3B4, SF3B5, SF3B6, PHF5A) plays a pivotal role in branch point recognition[31]. Different splicing drugs perturb this step of the reaction. For instance, H3B-8800 and Pladienolide B bind SF3B1[32,33]; E7107 prevents the binding of U2 snRNP to RNA[34]; and Isoginkgetin and Spliceostatin A prevent the transition from A to B spliceosome complexes in the splicing reaction[35,36]. It is thus reasonable to expect that perturbing the SF3b complex with these drugs should strongly decrease the activity of the associated splicing factors, proportionally to their distance to the SF3b complex in the PPI network (Fig. 2e and f). Consistent with this prediction, treating cell lines with the aforementioned drugs induced notable VIPER-assessed inactivation of splicing factors in the SF3b complex, followed by splicing factors in the U2 snRNP and other splicing factors with direct (shortest path length in PPI space = 1) and indirect (shortest path length in PPI space > 1) interactions with the SF3b complex (Fig. 2g). Taken together, these data show that combining VIPER with empirical splicing factor networks accurately recapitulates splicing factor activity modulation mechanisms, across a range of diverse perturbations that were not used to generate the splicing factor networks.

Elucidating aberrant regulation of splicing programs in cancer

Given VIPER's ability to accurately assess differential splicing factor activity from exon inclusion profiles, we analyzed a variety of cancer contexts to assess whether specific splicing factors may be consistently aberrantly activated and inactivated, thus suggesting the existence of a recurrently dysregulated cancer-specific splicing program.

Of the 33 cancer cohorts included in The Cancer Genome Atlas (TCGA), 14 include ≥ 10 exon-level transcriptional profiles from primary tumor (PT) and solid tissue normal (STN) biopsies. We leveraged these samples to perform VIPER-based assessment of splicing factor activities in each PT sample versus the median exon inclusion profiles across its normal STN counterpart and in each STN sample versus the median exon inclusion profiles across its PT counterpart (Fig. 3a–b). We then focused on the splicing factors that were significantly differentially active (FDR < 0.05, two-sided Wilcoxon rank sum test) in PTs compared with their corresponding STNs across all cancer cohorts (Fig. 3b). Interestingly, some splicing factors emerged as recurrently activated or inactivated across a majority of cancer cohorts (Fig. 3c; Supplementary Fig. 3a). This suggests that oncogenic alterations may cause coordinated and recurrent dysregulation of splicing factor activity. Considering a threshold of at least 5 out of 14 cancer cohorts to be indicative of recurrent dysregulation (see Methods), the analysis identified a set of (n=63) putatively oncogenic splicing factors (i.e., aberrantly activated in >= 5 cohorts), and a set of (n=64) putatively tumor suppressor splicing factors (i.e., aberrantly inactivated in >= 5 cohorts) (Supplementary Table 3).

If these recurrently dysregulated splicing factors play a mechanistic role in tumorigenesis and progression, then, activation of oncogenic and tumor suppressor factors should be predictive of poor and good prognosis, respectively. Indeed, outcome analyses confirmed these expectations (Fig. 3d; see Methods). For example, analysis of primary tumor samples across 33 TCGA cohorts revealed that aberrant activity of a recurrent oncogenic splicing factor (HNRNPD) significantly correlates with poor prognosis in 27 of 33 cancer cohorts (81.8%; Cox Proportional Hazards (PH) coefficient > 0) while the activity of a recurrent tumor suppressor splicing factor (SNRNP200) correlates with good outcome in 28 of 33 cancer cohorts (84.8%; Cox PH coefficient < 0) (Fig. 3e).

Extending the analysis to all 127 recurrently dysregulated splicing factors showed the same trends. Indeed, aberrant activation of oncogenic splicing factors correlates with poor outcome across significantly more cancer cohorts than tumor suppressor ones ($p < 0.0001$, two-sided Wilcoxon rank sum test). The opposite happens for tumor suppressor-like splicing factors; their activity correlates with good outcome across significantly more cancer cohorts than the oncogenic ones ($p < 0.0001$, two-sided Wilcoxon rank sum test) (Fig. 3f). These patterns are conserved if we consider different thresholds to define cancer splicing program recurrence (Supplementary Fig. 3b).

In the literature, splicing factor activity is typically extrapolated from splicing factor expression. We argue splicing factor expression is only one of the molecular mechanisms regulating splicing factor activity. Hence, we expect that our approach should capture activity differences that are not reflected by differences in mRNA levels. To evaluate this, we repeated the same analysis with splicing factor gene expression. Confirming our expectations, gene expression fold-changes in tumor vs matched normal samples did not correlate with their differential activity (Spearman correlation = –0.055, $p = 0.011$) (Supplementary Fig. 3c). Candidate oncogenic (n = 90) and tumor suppressor splicing factors (n = 23) identified by gene expression analysis (Supplementary Fig. 3d) were not predictive of poor and good outcome ($p > 0.05$ in both cases) (Supplementary Fig. 3e-f). Thus, by focusing on changes in splicing factor targets rather than in splicing factor expression, our approach systematically elucidated the potential contribution of their dysregulation to tumorigenesis and outcome.

Cancer splicing programs regulate cancer hallmarks

Given the functional relevance of the cancer splicing programs identified through our approach, we wondered which biological processes they may potentially regulate (Fig. 4a). For this purpose, we performed an overrepresentation analysis (ORA) of the genes corresponding to the exons targeted by either oncogenic-like or tumor suppressor-like splicing factors and only considered the top 10 most enriched terms in each program (FDR < 0.05, ORA) (see Methods). While two of the enriched terms were related to antigen presentation, a critically regulated process in tumor immunology, most enriched terms were related to cell cycle (Fig. 4b). Further confirming the biological relevance of the cancer splicing programs identified by these analyses, the target exons of splicing factors dysregulated in cancer reside in genes enriched in biological processes involved in two hallmarks of cancer: immune evasion and aberrant proliferation.

Immune evasion by cancer cells hinders the efficacy of immune checkpoint blockade (ICB) treatments in cancer patients making response prediction challenging[37]. Interestingly, a total of 181 genes belonging to the terms “Antigen Processing Ubiquitination Proteasome Degradation” and “Class I MHC-mediated Antigen Processing Presentation” related to immune evasion contain at least one exon regulated by the tumor suppressor-like splicing program (Fig. 4b). Conversely, exons regulated by the oncogenic-like splicing program are not enriched in immune evasion-related terms. To pinpoint exons of the tumor suppressor-like splicing program with a potentially important role in ICB efficacy, we performed differential splicing analyses between ICB-responding and non-responding patients classified following RECIST[38] guidelines—based on changes in tumor size—from Riaz *et al.*[39]. Of the 608 exons belonging to the genes in enriched terms and to the lists of targets of tumor suppressor-like splicing factors, 16 exons were differentially spliced (p < 0.05, two-sided Wilcoxon rank sum tests). To further prioritize these exons by their potential impact on ICB efficacy, we obtained data from an in vivo genome-wide CRISPR knockout screen in cancer cell lines inoculated in immunocompetent mice ICB-treated versus an immuno-deficient NOD SCID Il2rg−/− control mice[40]. With these data, we could ask which of the 16 differentially spliced exons belong to genes whose mouse orthologs inhibit (negative score) or promote (positive score) immune evasion (Fig. 4c). In particular, the differentially spliced exon HsaEX1036341 belongs to a gene encoding for one of the strongest inhibitors of immune evasion screened, SEC22B[40,41]. In other words, these results suggest that disrupting SEC22B hampers immune evasion mechanisms in cancer cells, making them more susceptible to ICB treatment. Considering that exon HsaEX1036341_SEC22B is annotated to disrupt its gene’s open reading frame upon exclusion, we hypothesize patients with lower inclusion of the exon in their tumor should respond better to ICB treatment due to a decrease in the amount of functional SEC22B protein. In fact, HsaEX1036341_SEC22B is slightly less included in responder (median exon inclusion (PSI) = 95.9, n = 10) than non-responder (median exon inclusion (PSI) = 99.4, n = 22) patients (p = 0.024, two-sided Wilcoxon rank sum tests) (Fig. 4c). Further, performing overall survival analysis of the ICB-treated patients stratified by HsaEX1036341_SEC22B inclusion, we find that patients with a functionally disrupted SEC22B, namely low HsaEX1036341_SEC22B inclusion, exhibit better prognosis than patients with high HsaEX1036341_SEC22B inclusion (p = 0.0079, Log-rank test; best cutpoint = 98.91 PSI) (Fig. 4d; See Methods). We conclude that analyzing the interactions within the identified cancer splicing programs could unveil exons with potentially relevant roles in ICB efficacy.

To validate the involvement of cancer splicing programs in cell proliferation, we estimated splicing factor activities for 1,015 cell lines in the Cancer Cell Line Encyclopedia (CCLE) and correlated the median activities among either oncogenic-like or tumor suppressor-like splicing factors against MKI67 expression, a marker of cell proliferation[42]. The activity of the oncogenic-like splicing program correlates positively with the proliferation marker (Pearson correlation = 0.21, $p < 2.8 \cdot 10^{-11}$) and the activity of the tumor suppressor-like splicing program correlates negatively (Pearson correlation = -0.33, $p < 2.2 \cdot 10^{-16}$) (Fig. 4e). Further, the sign of these relationships is reproducible when computing the correlation by grouping CCLE cell lines by cancer type of origin (Supplementary Fig. 4). This not only confirms the link between cell proliferative state and the activation of cancer splicing programs but also corroborates the cancer driver-like behavior of the splicing programs identified in an independent dataset as high proliferation marker expression is simultaneously associated with activation of the oncogenic program and inactivation of the tumor suppressor program.

Coordinated regulation of cancer splicing programs during carcinogenesis

Carcinogenesis is the process through which a normal cell becomes a cancer cell with the potential to originate a tumor. This process is typically initiated by the acquisition of cancer-driver mutations that permanently activate oncogenes and inactivate tumor suppressor genes[43]. To achieve full transformation into cancer cells, carcinogenic mutations rewire normal cells at all regulatory levels, including splicing. Then, if cancer splicing programs exhibit a cancer driver-like behavior, they should be modulated to mediate the establishment of the cancerous state during carcinogenesis (Fig. 4a).

To test this hypothesis, we analyzed how splicing factor activities change during carcinogenic transformation. Danielsson *et al.*[44] measured the transcriptomes of a model of carcinogenesis consisting of primary BJ fibroblasts immortalized with telomerase reverse transcriptase (TERT), transformed into cancer cells by adding the SV40 large-T antigen, and made metastatic by introducing the oncogenic H-Ras mutation RASG12V (Fig. 4f)[45]. We then estimated the splicing factor activities for the immortalized, transformed, and metastatic fibroblasts using the primary fibroblasts as a reference (see Methods). While oncogenic-like splicing factors are less active than tumor suppressor-like splicing factors in immortalized fibroblasts, oncogenic- and tumor suppressor-like splicing programs are respectively switched on and off in both transformed and metastatic fibroblasts (Fig. 4g). These results indicate the cancer splicing programs identified are coordinately activated during carcinogenesis, further validating their cancer driver-like behavior.

DISCUSSION

Repurposing VIPER with empirically derived splicing factor networks accurately estimated differential splicing factor activity solely based on the differential inclusion of target exons in transcriptomic data. Although empirically derived splicing factor networks were obtained from independent single-splicing factor knockout, knockdown, or overexpression experiments, our approach recapitulated multi-omic regulation of splicing factors including protein-level degradation, combinatorial silencing, and SF3b complex protein-protein interactions. This benchmarked our approach as a more universal reporter assay compared to transcription factors[14], without having to silence each splicing factor in every possible cellular context to

obtain highly predictive VIPER-compatible networks. This predictive power under diverse types of direct and indirect perturbations suggested we can infer splicing factor activities in conditions as complex as tumor samples. Performing splicing factor activity analysis on primary tumors compared with solid tissue normal samples revealed two new cancer splicing programs with splicing factors that behave like oncogenes and tumor suppressor genes. Our analyses suggest these cancer splicing programs are likely involved in carcinogenesis and disease progression.

Traditional methods to infer aberrant splicing factor activity focus on molecular events affecting their activity, such as the presence of genetic alterations or their differential expression. These not only require multi-omic profiling but also correct assessment of whether each specific event —e.g., a non-synonymous somatic mutation— may affect splicing factor function. In contrast, VIPER analysis shifts the focus from the splicing factors to the exons they regulate. This allowed precise assessment of splicing factor activity due to direct and indirect molecular alterations using only transcriptomic data. For example, inducing RBM39 degradation with Indisulam increased RBM39 mRNA levels but decreased its protein levels. While traditional methods would produce conflicting results depending on the omic information available, such as an increase in RBM39 activity based on its increase in mRNA levels or an activity decrease when inspecting its protein levels, the splicing signature of the perturbation assessed through VIPER with splicing factor networks showed a clear decrease in RBM39 activity. Further, only the cancer splicing programs identified through splicing factor activity inference showed disease-relevant behaviors. This places VIPER-assessed splicing factor activity analysis as a general approach capable of extracting new and functionally relevant insights from widely available transcriptomic data.

Cancer is a genomically complex disease leading to aberrant cell growth[46]. In cancer cells, the splicing machinery is hijacked, resulting in gene isoforms that initiate and sustain the disease[4,22]. Thousands of changes in splicing occur in cancer cells compared with healthy counterparts, suggesting there is a change in the active programs of splicing factors. While certain splicing factors have shown the ability to initiate cancer (e.g. SF3B1[47], RBFOX2[48], PTBP1[49]), the existence of general cancer-driver splicing programs remains underexplored[50]. We identified two recurrent splicing factor programs with oncogenic- and tumor suppressor-apparent behaviors that not only hold prognostic power but also are informative of multiple cancer hallmarks: (i) they are selectively activated during carcinogenesis; (ii) their activation correlates with cell proliferative state; and (iii) studying the splicing factor interactions in our networks highlighted relevant exons potentially influencing ICB treatment efficacy. This not only corroborates the power of splicing factor activity analysis with VIPER to distill biologically relevant splicing programs in complex molecular contexts but also sets the stage for future research on how the activation of these cancer driver-like splicing programs is mechanistically regulated.

A current limitation of our approach for splicing factor activity analysis with VIPER is the requirement for splicing factor networks obtained from experiments perturbing single splicing factors to obtain accurate activity estimations. Differently from transcription factors, we showed that the splicing factor networks obtained with ARACNe –the algorithm extensively used to infer transcription factor-gene networks[9,10]– generate informative splicing factor-exon likelihood scores but fail at inferring their mode of regulation sign using bulk RNA sequencing observational data. Interestingly, in a study developed concurrently using the

same ARACNe-based pipeline for splicing factor network reconstruction with a single-cell RNA sequencing dataset of mouse cortex, mode of regulation inference is consistently accurate[51]. This suggests cell type abundance in bulk RNA sequencing samples could confound the splicing factor-exon associations inferred computationally. Nevertheless, we show splicing factor activities can be accurately inferred for 243 (47.7%) out of the 509 annotated splicing factors using empirical splicing factor networks. These experimentally derived networks can be easily extended as more perturbation experiments become available.

Given the high recall of our approach to recapitulate altered splicing factor activities in highly diverse molecular contexts while only requiring transcriptomic data, we envision splicing factor activity analysis with VIPER to be instrumental in extracting the splicing programs underlying a variety of phenotypes.

METHODS

Compilation of the consensus list of splicing factors

As illustrated in Supplementary Fig. 1a, we combined the splicing factors from 5 different sources –five publications[20–24]– to create a consensus list of 509 splicing factors. Hegele *et al.*[24] performed a protein interaction screen among the proteins that copurify with spliceosomes. Papasaikas *et al.*[20] perturbed many of the components of the splicing machinery. Seiler *et al.*[21] studied the splicing factors with putative driver mutations across TCGA cancers. Rogalska *et al.*[22] reviewed the regulation of alternative splicing. Head *et al.*[23] studied the expression of splicing factors in cancer. This list can be found in Supplementary Table 1.

Quantification of gene expression and alternative splicing with vast-tools

All RNA sequencing samples considered in this study were processed with *vast-tools*[52] to quantify gene expression and exon inclusion. For each sample, we aligned sequencing reads to the hg38 genome assembly (Hs2 in VastDB) with the command "*vast-tools align --sp Hs2 --EEJ_counts --expr*". We combined them using "*vast-tools combine --sp Hs2 --keep_raw_reads --keep_raw_incl --TPM*". And, we set to NA the PSI of exons whose inclusion or exclusion was detected with less than 10 reads by running "*vast-tools tidy -min_N 1 -min_SD 0 --min_ALT_use 25 --noVLOW*". For each datast, this workflow results in a gene expression table corresponding to transcripts per million (TPM) –these values can range from 0 to potentially infinite– and an exon inclusion table quantifying the percentage of transcripts that include a certain exon or percentage spliced in (PSI) –these values can range from 0 to 100–. The details of each command can be found at https://github.com/vastgroup/vast-tools.

Assembly of splicing factor perturbation dataset from publicly-available RNA sequencing data

ENCORE

The ENCODE subproject called ENCORE[53] contains baseline and perturbed RNA sequencing data of K562 and HepG2 cell lines targeting RNA binding proteins either through shRNA knockdown or CRISPR knockout. Out of the different RNA binding proteins targeted in ENCORE, 164 overlapped with our consensus list of splicing factors. We then used *vast-tools* to analyze the transcriptomes of baseline and perturbed conditions for the splicing factors.

European Nucleotide Archive (ENA)

ENA is a public database of sequencing experiments. To find more splicing factors, we performed an advanced search querying for experimental data with the following filters:

- human samples: *tax_eq(9606)*
- RNA sequencing files: *library_strategy="RNA-Seq"* and *library_source="TRANSCRIPTOMIC"*
- title (*sample_title* and *experiment_title*), run (*run_alias*), or sample alias (*sample_alias*) containing any of the splicing factors in the consensus list

This search resulted in a table with 20,659 entries that we hand-curated to only contain experiments that specifically perturbed at least one splicing factor and the corresponding baseline condition for each perturbation. This resulted in 1,597 different experiments measuring the perturbation effects in the transcriptome for 114 splicing factors through knockout, knockdown, and overexpression. See Supplementary Table 2 for more details on each experiment and corresponding ENA identifiers.

Inference of splicing factor networks

Observational data for the computational inference of splicing factor networks (CardosoMoreira2019 dataset)

In this study, we use the VIPER algorithm to estimate splicing factor activity. To make these predictions, VIPER requires a splicing factor network with the likelihood (from 0 to 1) and mode of regulation (positive or negative) of each splicing factor-exon interaction. In principle, these networks cannot only be inferred empirically but also computationally. We used computational methods to infer splicing factor networks compatible with VIPER requirements by associating gene expression changes in splicing factors with exon inclusion of the exons in all genes. We reasoned this task requires samples with diverse expression of splicing factors and exon inclusions. Cardoso-Moreira *et al.*[19] performed RNA sequencing of human samples at different stages of differentiation and age, across 6 different tissues (brain, heart, ovary, kidney, testis, and liver). We then downloaded and processed these data with *vast-tools* as described above to obtain gene expression and exon inclusion for a total of 313 samples. Missing values in each exon inclusion table were imputed using the function *sklearn.impute.KNNImputer* with k=5 nearest neighbors and default parameters. This imputed exon inclusion table was required to infer SF-exon networks with ARACNe and MLR algorithms (see below).

ARACNe

We inferred splicing factor networks using ARACNE-AP[10]. Specifically, we used the MRAS pipeline (https://github.com/chaolinzhanglab/mras)[51] which uses a forked version of the original ARACNE-AP that enables the computation of mutual information between two different matrices, in our case, gene expression and exon inclusion matrices (https://github.com/chaolinzhanglab/ARACNe-AP). We ran the standard pipeline on the CardosoMoreira2019 with a mutual information threshold of $1 \cdot 10^{-8}$ and 100 bootstraps using the table of gene expression of splicing factors as the regulator matrix and the table of imputed exon inclusion of all genes as the target matrix. Data processing inequality was applied with respect to splicing factors only. In addition, we pruned each bootstraped network keeping only the top 1,000 targets (exons) for each regulator (splicing factor) based on the pairwise mutual information. We consolidated the bootstrapped networks into a single network with a p-value threshold of 0.05. Mutual information scores in consolidated networks correspond to the likelihood of splicing factor networks. Finally, we estimated the mode of regulation for each splicing factor-exon interaction using the *viper::aracne2regulon* function.

Multivariate linear regression (MLR)

To obtain pairwise interactions between a splicing factor and an exon while controlling for the other possible interactions between that exon and other splicing factors, we performed multivariate linear regression. We identified putative splicing factor-exon interactions by fitting linear models on the CardosoMoreira2019 gene expression and imputed exon inclusion datasets. Specifically, we set the standardized log-normalized gene expression of each of the 509 splicing factors as the independent variables and the inclusion (PSI) of each exon as the dependent variable. This resulted in a regression coefficient and a p-value for that coefficient value for each splicing factor and exon. In this setting, regression coefficients correspond to the strength of the putative splicing factor-exon interactions and p-values to the statistical significance of that interaction. To build splicing factor networks compatible with VIPER, we kept those regression coefficients with a corresponding nominal p-value lower than 0.01. We considered the absolute value of regression coefficients as the interaction likelihood, and the sign of these coefficients as the mode of regulation.

Empirical

We created empirical splicing factor networks using experimental data from the publicly available perturbation datasets described above. For each experiment, including splicing factor knockdown, knockout, or overexpression, we computed a delta PSI by subtracting the exon inclusion in the perturbed condition from the control condition. To keep consistency between all types of perturbations, we multiplied the delta PSI in overexpression experiments by -1 assuming that exons more included upon upregulation of a splicing factor would be excluded upon downregulation in the same cell system. Then, for each experiment, we kept those exons exhibiting strong inclusion changes upon splicing factor perturbation (|delta PSI| ≥ threshold). For a range of thresholds, from 5 to 45% in steps of 5, we evaluated within- and across-sample recall of the resulting networks. A threshold of 15% leads to highly accurate and reasonably-sized networks comprising 243 of the 247 splicing factors with experimental data available. The VIPER-compatible splicing factor network for a single splicing factor in a given experiment consists of the absolute delta PSI as the likelihood of the interaction and the sign of the delta PSI as the mode of regulation of the interaction. We then combined splicing factor networks for single splicing factors considering their dataset of origin. Datasets obtained from ENA are heterogeneous and contain more than one perturbation experiment, which results in a splicing factor network for some splicing factors. Instead of ignoring these valuable independent biological replicates to consider only one experiment per splicing factor, we grouped splicing factor networks obtained independently into new dataset-level networks. Altogether, this resulted in a total of eight networks: (i) ENCORE knockdowns in HepG2 cells (41,477 splicing factor-exon interactions, 104 splicing factors); (ii) ENCORE knockdowns in K562 cells (45,396 splicing factor-exon interactions, 104 splicing factors); (iii) ENCORE knockouts in HepG2 cells (6,340 splicing factor-exon interactions, 44 splicing factors); (iv) ENCORE knockouts in K562 (17,119 splicing factor-exon interactions, 61 splicing factors); (v) ENA perturbations in different cell systems 1 (14,158 splicing factor-exon interactions, 108 splicing factors, 48 studies); (vi) ENA perturbations in different cell systems 2 (3,042 splicing factor-exon interactions, 23 splicing factors, 9 studies); (vii) ENA perturbations in different cell systems 3 (522 splicing factor-exon interactions, 12 splicing factors, 3 studies); and (viii) ENA perturbations in different cell systems 4 (963 splicing factor-exon interactions, 2 splicing factors, 2 studies).

ARACNe likelihood and Empirical mode of regulation

To test whether the likelihoods of splicing factor networks obtained with ARACNe could cause the low recall of VIPER splicing factor activities, we created splicing factor networks substituting likelihoods in empirical networks with likelihoods obtained with ARACNe. Note that we could only consider those splicing factor-exon interactions found in both networks.

MLR likelihood and Empirical mode of regulation

We performed the same procedure described above for ARACNe-derived networks with the difference that we substituted empirical likelihoods with MLR-derived likelihoods and, again, could only keep those splicing factor-exon interactions found in both empirical and MLR networks.

ARACNe negative modes of regulation and MLR positive modes of regulation

We created a version of splicing factor networks that combined all those splicing factor-exon interactions in ARACNe networks that had a negative mode of regulation with all those splicing factor-exon interactions in MLR networks that had a positive mode of regulation.

Estimation of splicing factor activities using splicing factor networks with VIPER

To predict protein activity, in this case splicing factor activity, with the *viper::viper* function we require two inputs: (i) a molecular signature, namely the differences in exon inclusion occurring in a sample compared to a reference condition (e.g. perturbed versus unperturbed); and (ii) a network or list of networks associating each regulator of interest (splicing factors) to their targets (exons) with a certain likelihood (a positive value ranging from 0 to 1) and mode of regulation (positive or negative). Precisely, while the likelihood score quantifies the probability of the interaction, the mode of regulation value quantifies the direction of that interaction, i.e. whether a given exon is more included (positive) or excluded (negative) with the increase in the activity of its splicing factor regulator. This input network can be either a single consensus network typically obtained by combining many sources of information or a list of networks obtained from different sources. While the former runs the classical version of the VIPER algorithm for estimation of protein activity, the latter runs metaVIPER[54], a newer version of the algorithm capable of integrating the protein activities obtained from multiple networks of the same regulators increasing the robustness of predicted activities. In this study, we used the classical VIPER algorithm only for the evaluation and comparison of splicing factor activity prediction with computationally- and empirically-derived splicing factor networks. After observing the high recall of empirical splicing factor networks, we always ran VIPER in metaVIPER mode to seamlessly integrate the multiple splicing factor networks available for some splicing factors that we obtained from different publicly available sources. Note that VIPER predicts the activity for each of the regulators in the network used as input. While we can potentially predict interactions for any splicing factor using computationally inferred networks, empirical networks only contain splicing factor-exon interactions for a total of 243 of the 509 splicing factors in our consensus list.

Evaluation metrics of splicing factor networks

We evaluated the ability of the different splicing factor networks to estimate accurate splicing factor activities with VIPER by creating a ground truth dataset from the publicly available perturbation data analyzed above. In this dataset, comprising all ENCORE and ENA experiments perturbing a single splicing factor, we know which experimental perturbation caused every splicing signature. Hence, VIPER-predicted protein activities with a good splicing factor network should quantitatively pinpoint which splicing factor was perturbed in each experiment. In other words, the best networks for VIPER inference of splicing factor activity should return an extreme value in those experiments where the corresponding splicing factor had been perturbed. Then, splicing factor activities can be predicted accurately in two manners considering whether the predicted activity of the perturbed splicing factor is extreme compared to the rest of the splicing factors "within" the perturbed sample or compared to itself "across" the rest of the samples that did not perturb this splicing factor.

<u>Across-sample recall</u>

This score quantifies how calibrated predicted splicing factor activities are for a given splicing factor across a set of samples in which the majority do not perturb the splicing factor of interest. To compute the across-sample recall for a given splicing factor, we sort samples based on splicing factor activity predicted with VIPER, we store the median rank of the sample(s) in which the splicing factor was perturbed experimentally, and we normalize that rank by the total number of samples considered to have a score ranging from 0 to 1, being 1 the best possible performance.

<u>Within-sample recall</u>

On the other hand, within-sample recall quantifies how calibrated predicted splicing factor activities are for a given splicing factor within each sample in which that splicing factor was perturbed. To calculate the within-sample recall, we sort splicing factors according to their splicing factor activity predicted with VIPER, we store the rank of the perturbed splicing factor, and we normalize by the total number of splicing factors considered which results in a score ranging from 0 to 1, being 1 the best possible performance.

Of note, because empirical splicing factor networks derive from the same data used for network evaluation, we excluded those recall scores obtained when using as ground truth the experiment from which the network was created.

Robustness of empirical splicing factor networks

To evaluate the robustness of empirical splicing factor networks based on the likelihood scores of splicing factor→exon interactions, we pruned the targets of each splicing factor keeping the 40, 50, 60, 70, 80, 90, and 100 interactions with the highest likelihood. We then evaluated pruned versions of splicing factor networks by computing within- and across-sample recall as described above.

Experimental validations of splicing factor activity estimation with diverse perturbations

Indisulam-induced degradataion of RBM39 splicing factor protein

We downloaded the proteomic and transcriptomic data produced by Nijhuis *et al.*[25] upon perturbation of IMR-32 neuroblast cells with either Indisulam (5 μM, 6 hours) or DMS vehicle control. Proteomic experiments had 3 biological replicates per condition while transcriptomic experiments had 1 biological replicate per condition. On the proteomics side, we obtained processed proteomic data in the form of LFQ values for each protein and sample (see Data Availability). Then, considering only measurements for splicing factors, we performed a statistical differential analysis of protein expression comparing Indisulam-treated to control samples through a T-test for each splicing factor using the function *ggpubr::compare_means(method="t.test")* and computed the average log2 fold changes between conditions. On the transcriptomics side, we processed raw RNA sequencing FASTQ files with *vast-tools* as described above to quantify exon inclusion (PSI) and gene expression (TPM). To estimate splicing factor activities with VIPER, we computed the splicing signatures of the Indisulam-treated cells by subtracting exon inclusion measured in control samples and ran VIPER with empirical splicing factor networks as explained above.

Finally, because the Nijhuis *et al.*[25] dataset had a single biological replicate for transcriptomic data condition (see Supplementary Fig. 2a), we obtained RNA sequencing data from Lu *et al.*[26], who measured the transcriptomic changes upon treatment with Indisulam (1 μM, 96 hours) compared with DMSO vehicle control with three biological replicates per condition for three cell lines (A375, MEL501, SKMEL239). After processing raw files with *vast-tools* as described above, we confirmed the changes in mRNA levels of RBM39 upon Indisulam treatment (see Supplementary Fig. 2b).

Combinatorial knockdown of splicing factors

For the four studies considered that measured transcriptomic changes upon combinatorial knockdown of splicing factors[27–30], we processed their raw RNA sequencing data with *vast-tools* as described above to obtain exon inclusion profiles. For each study, cell line, and condition, we obtained the splicing signatures required for VIPER inference of splicing factor activity by subtracting the average exon inclusion in unperturbed control samples from the corresponding perturbed samples. We estimated splicing factor activities with empirical splicing factor networks as described above. And, when applicable, we median-summarized splicing factor activities across biological replicates. This resulted in a table of predicted activities per condition and splicing factor in the empirical splicing factor networks.

Recapitulation of SF3b complex protein-protein interactions

We measured the shortest paths from each protein to each splicing factor in the SF3b complex within the STRINGDB protein-protein interaction network using the *networkx.shortest_path_length* function. Then, we summarized shortest path lengths to the SF3b complex level considering the minimal shortest path lengths among the splicing factors in the protein complex. We placed all interactions with a minimal shortest path length equal to or greater than 4 in the same bin.

We estimated splicing factor activities for six studies that measured how the transcriptome changes when perturbing cell lines with spliceosome drugs[21,55–59]. We processed raw RNA sequencing data with *vast-tools* as described above to obtain exon inclusion profiles. For each study, cell line, drug concentration, and incubation time, we obtained splicing signatures by subtracting the average exon inclusion in unperturbed vehicle controls from the corresponding perturbed condition. We then computed splicing factor activities with VIPER using empirical splicing factor networks as described above. Finally, when applicable, we summarized the estimated splicing factor activities across biological replicates in each condition by computing the median.

Identification of recurrent splicing programs in cancer

We processed raw RNA sequencing samples from TCGA with *vast-tools* as described above to obtain exon inclusion and gene expression tables for all 33 types of cancer available. To identify recurrent splicing programs associated with cancer, for each type of cancer, we compared primary tumor samples to their healthy solid tissue normal counterparts provided each condition had at least 10 samples. Hence, we could only perform this comparison for 14 types of cancer.

Cancer splicing programs based on splicing factor activity

To perform differential activity analysis, we require predicting splicing factor activity in both tumor and healthy samples. We reasoned we could obtain splicing signatures for both conditions using the median of the other as a reference. Specifically, for each cancer cohort with sufficient data, we obtained splicing signatures for each tumoral sample by subtracting the median exon inclusion across healthy samples and we obtained splicing signatures for each healthy sample by subtracting the median exon inclusion across tumor samples. We used these signatures to compute VIPER splicing factor activities with empirical splicing factor networks as explained above. This resulted in a table of splicing factor activities for each tumor- and healthy-derived sample in each type of cancer. We performed differential activity analysis for each splicing factor by computing the median activity difference between tumor and healthy conditions and by statistically testing the significance of these differences through two-sided Wilcoxon rank sum tests using the function *scipy.stats.mannwhitneyu*. We considered significant all those differences in splicing factor activity with an FDR (adjusted *p*-value) lower than 0.05. We quantified the recurrence of splicing factor activation and inactivation by counting in how many of the 14 types of cancers considered each splicing factor was differentially activated (oncogenic-like) or inactivated (tumor suppressor-like) in tumor samples. We assigned splicing factors to the oncogenic cancer splicing program if they were more frequently activated than inactivated in tumors with a threshold difference of 5. And, correspondingly, we assigned splicing factors to the tumor suppressor cancer splicing program if they were more frequently inactivated than activated in tumors with a threshold difference of 5.

Cancer splicing programs based on splicing factor gene expression

To perform differential gene expression analysis, we followed the same approach described for splicing factor activity to obtain comparable results. Briefly, for each splicing factor in the same 14 types of cancers considered for differential activity analysis, we used gene expression matrices obtained with *vast-tools* to compute the median gene expression

difference between tumor and healthy samples and we statistically tested these differences through two-sided Wilcoxon rank sum tests using the function *scipy.stats.mannwhitneyu*. We then considered significant differences in gene expression with an FDR lower than 0.05, and counted in how many types of cancer considered each splicing factor is up- or down-regulated. Finally, following the same procedure above, we defined recurrent cancer splicing programs based on differential gene expression considering those splicing factors preferentially upregulated (oncogenic-like) or downregulated (tumor suppressor-like) with a minimum difference of at least 5 different cancer cohorts.

Association of molecular measurements with patient outcome

For the 33 types of cancer in the TCGA, we studied the prognostic power of the splicing factors in recurrent cancer splicing programs identified either through splicing factor activities predicted with VIPER or through splicing factor mRNA levels by fitting Cox Proportional Hazards (Cox PH) models using the *survival::coxph* function that associate splicing factor activities or gene expression in tumor samples to patient clinical data (overall survival and survival status). Fitted model coefficients inform how the increase of the variable of interest is associated with an increase (positive coefficient) or decrease (negative coefficient) in the risk of death. Note that we could perform this analysis across all available cancer cohorts because we only required tumor samples.

For this analysis, we had to re-estimate splicing factor activities with VIPER as we aimed to study how splicing factor activities varied across tumors regardless of the molecular state of healthy counterparts. We obtained splicing signatures for each type of cancer by subtracting median exon inclusion across the tumor samples from each cancer cohort. Then, we computed splicing factor activities with VIPER using empirical splicing factor networks as described above.

Overrepresentation analysis (ORA)

All statistical tests for gene set overlap analysis were performed using the function *clusterProfiler::enricher* with default parameters. We used the ReactomeDB pathways as the gene set-term database.

Identification of exons in splicing programs associated with immune evasion

<u>Analysis of melanoma patients treated with immune checkpoint blockade drugs</u>

As mentioned in the results section, genes carrying exons in the tumor suppressor-like cancer splicing program were enriched in immune evasion-related terms. To study whether exons in genes belonging to these gene sets could play a role in response to ICB treatment, we processed raw RNA sequencing data with *vast-tools* as described above for the 109 tumor samples of cancer patients treated with ICB from the *Riaz et al.*[39] study. We then classified each patient into a responder or non-responder considering their RECIST[38] classification; CR (n=3) and PR (n=7) patients were considered as responders (n=10), and PD patients as non-responders (n=23). For each of the 608 exons in the 181 genes found in enriched immune invasion terms, we performed a differential splicing analysis comparing responders to non-responders. We computed exon inclusion differences between the groups by subtracting median exon inclusion in non-responders from median exon inclusion in

responders and we tested these differences statistically through two-sided Wilcoxon rank sum tests using the function *ggpubr::compare_means(method="wilcox.test")*. We only reported as significant those exon inclusion differences with a nominal *p*-value < 0.05.

In addition, for the case of exon HsaEX1036341_SEC22B, we performed a Kaplan-Meier curve analysis to relate overall patient survival probability over time while stratifying by the inclusion of this exon. To find the best possible split in exon HsaEX1036341_SEC22B inclusion, we used the function *survminer::surv_cutpoint* with default parameters.

In vivo genome-wide CRISPR knockout screening data

We used data from the Dubrot *et al.*[40] study to obtain the score quantifying the contribution of each gene to immune evasion in immunocompetent mice model treated with ICB compared with immuno-deficient mice. After downloading those data, we translated mouse gene names to human gene names using a table generated through BioMart (https://www.ensembl.org/info/data/biomart/index.html). Then, we multiplied the "Average Score" column by the "Sign" column to obtain signed scores that are informative of whether knocking out a given gene favors (positive) or hampers (negative) immune evasion and we kept only those signed scores measured for the condition "ICB vs NSG". Finally, we only considered those genes whose exon was differentially spliced in the analysis above.

Association of splicing factor activity with proliferative state marker MKI67

We processed raw RNA sequencing data from the CCLE with *vast-tools* as described above to quantify both exon inclusion and gene expression for the 1,015 cell lines with sufficient sequencing coverage. To estimate splicing factor activities with VIPER, we computed splicing signatures for each cell line by subtracting the median exon inclusion from the exon inclusion matrix. We predicted splicing factor activities with VIPER using empirical splicing factor networks as described above.

Activities of cancer splicing programs during carcinogenesis

We processed raw RNA sequencing data from the Danielsson *et al.*[44] study with *vast-tools* as described above. To estimate how splicing factor activities change at each stage of BJ fibroblast transformation, we computed the splicing signatures by subtracting average exon inclusion in unperturbed cells from transformed cells. We then predicted splicing factor activity with VIPER using empirical splicing factor networks as explained above.

DATA AVAILABILITY

Previously published datasets used

List of splicing factors from Papasaikas et al.[20]
Supplementary Table 2 in the publication.

List of splicing factors from Hegele et al.[24]
Supplementary Table 1 in the publication.

List of splicing factors from Seiler et al.[21]

Supplementary Table 1 in the publication.

List of splicing factors from Rogalska et al.[22]
Supplementary Table 1 in the publication.

List of splicing factors from Head et al.[23]
Supplementary Table 1 in the publication.

RNA sequencing data of splicing factor perturbations from ENCORE project
We downloaded raw FASTQ files from the ENCORE website for all shRNA and CRISPR experiments targeting RNA binding proteins: https://www.encodeproject.org/encore-matrix/?type=Experiment&status=released&internal_tags=ENCORE.

RNA sequencing data of splicing factor perturbations from ENA
We downloaded raw FASTQ files from the ENA website listed in Supplementary Table 2 in this manuscript.

Observational RNA sequencing data from human samples from diverse molecular contexts
We downloaded raw FASTQ files for Cardoso-Moreira *et al.*[19] through the ENA website.
Perturbing RBM39 with Indisulam
For the Nijhuis *et al.*[25] study, we downloaded raw FASTQ files from the ENA website (PRJNA673205) and processed LFQ proteomics data from the PRIDE website (PXD022164). Metadata for proteomics experiments were kindly provided by the authors.
For the Lu *et al.*[26] study, we downloaded raw FASTQ files from the ENA website (PRJNA683080).

Combinatorial knockdowns of splicing factors
We downloaded raw FASTQ files for four different studies from the ENA website: PRJNA223244[30], PRJNA498529[29], PRJNA587741[28], and PRJNA321560[27].

Protein-protein interaction network
We downloaded STRING DB's human protein interaction network from their webpage[60].
https://stringdb-static.org/download/protein.links.full.v11.5/9606.protein.links.full.v11.5.txt.gz

List of genes in SF3b complex
https://signor.uniroma2.it/relation_result.php?id=SIGNOR-C442

List of genes in U2snRNP complex
https://signor.uniroma2.it/relation_result.php?id=SIGNOR-C479

Perturbing SF3b complex reaction step with splicing drugs
We downloaded raw FASTQ files for six different studies from the ENA website: PRJNA371421[32], PRJNA685790[55], PRJNA662572[56], PRJNA380104[57], PRJNA354957[58], and PRJNA292827[59]

Splicing event information

We downloaded information and their predicted protein impact of splicing events from VastDB.
https://vastdb.crg.eu/wiki/Downloads#Homo_sapiens_.28hg38.29

Molecular and clinical data from The Cancer Genome Atlas (TCGA)
We used the GDC portal to download the clinical, survival, and raw RNA-seq data.

Reactome pathways processed by MSigDB
We obtained ReactomeDB pathways from the MSigDB website.
https://www.gsea-msigdb.org/gsea/msigdb/human/genesets.jsp?collection=CP:REACTOME

Immune evasion CRISPR screen from Dubrot et al.[40]
We obtained the immune evasion score for each gene from Supplementary Table 13 in the publication.
https://static-content.springer.com/esm/art%3A10.1038%2Fs41590-022-01315-x/MediaObjects/41590_2022_1315_MOESM2_ESM.xlsx

Human to mouse gene name translations
We obtained tables of human-to-mouse gene name translations from BioMart.

Data from patients treated with ICB from Riaz et al.[39]
We obtained raw FASTQ files from the ENA website (PRJNA356761).
We obtained patient clinical metadata from Supplementary Table 2 in the publication.

Molecular data from the Cancer Cell Line Encyclopedia (CCLE)
We downloaded raw RNA-seq .fastq files for the cancer cell lines in the CCLE from the ENA website (PRJNA523380).
We downloaded cell line metadata ("sample_info.csv") and processed mutation mapping across CCLE cell lines ("CCLE_mutations.csv") from DepMap's figshare repository (download link: https://ndownloader.figshare.com/articles/13681534/versions/1).

RNA sequencing data of carcinogenesis from Danielsson et al.[44]
We downloaded raw FASTQ files from the ENA website (PRJNA193487).

CODE AVAILABILITY

To improve the reproducibility of our results, the full data analysis pipeline of this manuscript is publicly available at https://github.com/MiqG/publication_viper_splicing and the corresponding scripts to run VIPER with the SF-exon networks identified in this study are available at https://github.com/MiqG/viper_splicing.

FIGURE LEGENDS

Figure 1. Reverse engineering of splicing factor networks to estimate splicing factor activities with VIPER.

(**a**) Splicing factors regulate the removal of introns from pre-mRNA. The activity of splicing factors is regulated by multiple molecular mechanisms such as their mRNA levels, alternative splicing, protein

levels, post-translational modifications, subcellular localization, and cofactor interactions. (**b**) Inference of differential splicing factor activity with VIPER. The algorithm requires two inputs: (i) a splicing signature, the difference between exon inclusion in a condition of interest (Cond. A) and a reference condition (Cond. B); (ii) a splicing factor (SF) network, the splicing factor-exon interactions and their corresponding likelihood and mode of regulation (MoR). Splicing factor networks can be obtained using either empirical –from splicing factor experimental perturbations– or computational –here we consider ARACNe and Multivariate Linear Regression (MLR)– methods. (**c**) Evaluation of splicing factor activity estimation (y-axis) with VIPER using different methods to reverse engineer splicing factor networks (x-axis). "across"-sample and "within"-sample recall evaluate how well estimated splicing factor activities predict either in which experiment was a splicing factor perturbed and which splicing factor was perturbed in a given experiment, respectively (see Methods). Color, whether the corresponding network was randomly shuffled or not. (**d** and **e**) Distributions of the number of (**d**) targets (exons) per regulator (splicing factor) considering either all splicing factors (left) or categorized (right), and (**e**) regulators (splicing factors) per target (exon) in empirical splicing factor networks. "Core" splicing factors appear in Splicesome Database[61], "RBP" splicing factors appear in Gene Ontology[62] category "RNA binding protein". In box and whiskers plots in panels (**c**), (**d**), and (**e**), the median value is indicated as a horizontal line, and the lower and upper bounds of the box correspond to the first- and third quartiles, respectively. The upper and lower whiskers range from the corresponding box hinges to the largest value no further than 1.5 times the interquartile range from the hinge. All outlying data points beyond the whiskers are plotted individually.

Figure 2. Assessing splicing factor activity in complex molecular contexts.

(**a**) Indisulam promotes protein-level degradation of splicing factor RBM39, inducing significant changes in the exon inclusion repertoire of the cell. (**b**) Differential proteomic expression of splicing factors. Relationship between fold change LFQ values comparing Indisulam-treated (5 μM, 6 hours) and untreated samples (x-axis) and p-values obtained from a two-sided T-test (y-axis) in IMR-32 cells. Red, RBM39. (**c**) Ranking of splicing factors (x-axis) according to their estimated protein activities (y-axis) upon treatment with Indisulam (5 μM, 6 hours) in IMR-32 cells. Red, RBM39. (**d**) Distributions of estimated protein activities (y-axis) upon combinatorial perturbation of splicing factors (x-axis). Red, corresponding knocked down genes in each case. Top, experimental outline. Studies analyzed include PRJNA223244[30], PRJNA498529[29], PRJNA587741[28], and PRJNA321560[27]. (**e**) The SF3b complex is composed of 8 proteins. When perturbing spliceosome assembly with drugs, we hypothesize proteins in the SF3b complex and proteins interacting with the complex should decrease their activity. Adapted from Rogalska *et al.*[22] (**f**) Distribution of shortest path length to reach any protein of the SF3b complex in the STRINGDB protein-protein interaction network. Color, whether genes belong to the U2snRNP. (**g**) Top, mechanisms of action of splicing-targeting drugs: H3B-8800 and Pladienolide B bind the SF3b complex; E7107 prevents the binding of U2 complex to RNA; Isoginkgetin and Spliceostatin A impede the transition from A to B complex. Bottom, distributions of splicing factor activities (y-axis) upon perturbation with different splicing-targeting drugs considering the shortest path lengths of each splicing factor to the SF3b complex (x-axis). Color, whether genes belong to the U2snRNP. Studies analyzed include PRJNA371421[32], PRJNA685790[55], PRJNA662572[56], PRJNA380104[57], PRJNA354957[58], and PRJNA292827[59]. In box and whiskers plots in the panel, the median value is indicated as a horizontal line, and the lower and upper bounds of the box correspond to the first- and third quartiles, respectively. The upper and lower whiskers range from the corresponding box hinges to the largest value no further than 1.5 times the interquartile range from the hinge. All outlying data points beyond the whiskers are plotted individually.

Figure 3. Splicing factors recurrently active in tumors exhibit cancer driver-like behavior.

(**a**) Overview of cancer cohorts considered from TCGA. Out of the 33 cancer cohorts with tumor-derived samples available, there are 14 cancer cohorts for which solid tissue normal (healthy)

samples are found. Cancer-type abbreviation meanings can be found at https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations. (**b**) Workflow to identify tumor-specific splicing programs. From healthy and tumor samples, we quantified their exon inclusion and computed the splicing signatures to obtain splicing activities for each splicing factor and sample. (**c**) Number of differentially (FDR < 0.05, two-sided Wilcoxon rank sum test) activated (top) and inactivated (bottom) splicing factors (x-axis) in tumor samples compared with solid tissue normal samples. Labels highlight the top 5 most frequently activated and inactivated splicing factors. Horizontal dashed lines correspond to the selected recurrency threshold to be considered a splicing factor in either cancer splicing program. On top of the dashed line, the number of splicing factors selected as part of each cancer splicing program. (**d**) Workflow to associate protein activities in primary tumor samples with patient overall survival using a Cox Proportional Hazards (Cox P.H.) survival model. Positive Cox P.H. coefficients indicate that high splicing factor activity is associated with a high risk of death. Conversely, negative Cox P.H. coefficients indicate that high splicing factor activity is associated with a low risk of death. (**e**) Number of cancer cohorts (y-axis) in which either HNRNPD (left) or SNRNP200 (right) are associated with either high or low risk of death (x-axis). Color, cancer splicing program. Top, number of cancer cohorts for each type of association. (**f**) Distributions of the number of cancer cohorts (y-axis) that each splicing factor is associated with a high or low risk of death (x-axis). Color, cancer splicing program. Top, ****$p \leq 0.0001$, ***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$ based on two-sided Wilcoxon rank sum tests. Bottom, number of splicing factors in the corresponding cancer splicing program. In box and whiskers plots in panel (**f**), the median value is indicated as a horizontal line, and the lower and upper bounds of the box correspond to the first- and third quartiles, respectively. The upper and lower whiskers range from the corresponding box hinges to the largest value no further than 1.5 times the interquartile range from the hinge. All outlying data points beyond the whiskers are plotted individually.

Figure 4. Contribution of cancer splicing programs to disease progression.

(**a**) The activation of cancer splicing programs during carcinogenesis. To transition from normal to cancer, cells undergo a series of changes starting from a series of genomic lesions that typically activate oncogenes and inactivate tumor suppressor genes. These alterations rewire the epigenomic state of the cell into cancerous. Provided the microenvironment does not recognize transformed cells, they will start proliferating generating a tumor. In parallel to this process, we hypothesize cancer splicing programs are regulated. (**b**) ORA results for the top 10 enriched terms from ReactomeDB considering the genes corresponding to the target exons of splicing factors in the cancer splicing programs. (**c**) Prioritization of exons with the potential to contribute to ICB treatment response. For each of the differentially spliced exons between responder and non-responder patients (p-value < 0.05, two-sided Wilcoxon rank sum test) (x-axis), we considered the orthologous gene-level immune evasion score (y-axis) obtained from CRISPR screens comparing the effects in cancer cell viability of knocking out genes in either immunocompetent mice treated with ICB versus immunodeficient mice. Color, exon impact obtained from VastDB annotations. (**d**) Kaplan Meier survival curves of ICB patients in Riaz *et al.*[39] stratified according to the HsaEX1036341_SEC22B inclusion levels that maximized log-rank test differences. Color, patients with high and low exon inclusion. Bottom, exon inclusion PSI cut point between the two groups of patients and log-rank test p-value. (**e**) Relationship between median activities of cancer splicing programs (y-axis) with cancer cell proliferative state (x-axis). Top, Pearson correlations (R) and corresponding p-values (p) between the variables in the axes. Error shadows, 95% confidence intervals. (**f**) Danielsson *et al.*[44] experimental design. Primary fibroblasts were induced into different states transfecting either hTERT (immortalized), hTERT and SV40 large-T (transformed), hTERT, SV40 large-T, and RAS G12V (metastasizing). (**g**) Splicing factor activities by cancer splicing program (y-axis) in different stages of carcinogenesis (x-axis). ****$p \leq 0.0001$ based on two-sided Wilcoxon rank sum tests comparing the activity of the splicing factors in each cancer program. In box and whiskers plots in panel (**g**), the median value is indicated as a horizontal line, and the lower and upper bounds of the box correspond to the first- and third quartiles, respectively. The upper and lower whiskers range from the corresponding box hinges to the largest

value no further than 1.5 times the interquartile range from the hinge. All outlying data points beyond the whiskers are plotted individually.

Supplementary Figure 1. Evaluation of methods to reverse engineer splicing factor networks.

(**a**) UpSet plot illustrating the overlaps between the lists of splicing factors collected from the literature to obtain our consensus list of splicing factors. (**b**) Overview of the data processed from publicly available repositories to evaluate the ability of inferred splicing factor networks to estimate splicing factor activities with VIPER and used to generate the empirical splicing factor network. (**c**) Ablation analysis of the empirical splicing factor network. We evaluated across- and within-recall (y-axis) of empirical splicing factor networks depending on the delta PSI threshold selected (x-xaxis). Bottom, number of splicing factors and exons considered in each network version. (**d**) Robustness analysis of the empirical splicing factor network. We evaluated how between and within accuracies (see Methods) changed in this network if we only considered the top N strongest splicing factor-exon interactions for each splicing factor. Of note, to avoid information leakage between data used to generate and evaluate empirical networks, we did not evaluate empirical splicing factor networks with the same perturbation experiment used to obtain that network. (**e**) Splicing factor network likelihoods inferred by ARACNe and MLR recapitulate ground truth experimental interactions. Distributions of splicing factor-exon likelihood scores (y-axis) in either ARACNe- or MLR-derived networks considering whether a splicing factor-exon interaction is observed experimentally (x-axis). Top, the *p*-value of two-sided Wilcoxon rank sum test. Bottom, the number of interactions in each category. (**f**) splicing factor network modes of regulation inferred by ARACNe and MLR are close to random (0.5). Proportion (y-axis) of negative and positive (x-axis) splicing factor-exon modes of regulation (MoR) interactions inferred by ARACNe and MLR algorithms that match experimental observations (color). (**g**) splicing factor networks resulting from combining ARACNe or MLR likelihoods with empirical modes of regulation recover prediction recall. Distributions of between and within accuracies (y-axis) for different splicing factor network versions (x-axis) that use the corresponding likelihood values but the experimentally derived mode of regulation signs. Color, whether the corresponding network was randomly shuffled or not. (**h**) splicing factor networks resulting from combining ARACNe and MLR edges with negative and positive modes of regulation, respectively, leads to near-random accuracies. Distributions of between and within accuracies (y-axis) of networks obtained by combining interactions with negative MoR in ARACNe networks and interactions with positive MoR in MLR networks. Color, whether the corresponding network was randomly shuffled or not. In box and whiskers plots in panels (**c**), (**d**), (**f**), and (**g**), the median value is indicated as a horizontal line, and the lower and upper bounds of the box correspond to the first- and third quartiles, respectively. The upper and lower whiskers range from the corresponding box hinges to the largest value no further than 1.5 times the interquartile range from the hinge. All outlying data points beyond the whiskers are plotted individually.

Supplementary Figure 2. Changes in transcriptomic gene expression upon treatment with Indisulam.

(**a**) mRNA levels of RBM39 (y-axis) treated with either Indisulam (5 μM, 6 hours) or DMSO vehicle control (x-axis) in the IMR-32 cell line from study Nijhuis *et al.*[25]. Top, corresponding gene expression values. Note this corresponds to a single biological replicate. (**b**) mRNA levels of RBM39 (y-axis) of A375, MEL501, and SKMEL239 cell lines (color) treated with either Indisulam (1 μM, 96 hours) or DMSO vehicle control (x-axis) from study Lu *et al.*[26]. ****$p \leq 0.0001$ based on two-sided Wilcoxon rank sum tests comparing the mRNA levels of cell lines in each condition. (**b**). In box and whiskers plots in panel (**b**), the median value is indicated as a horizontal line, and the lower and upper bounds of the box correspond to the first- and third quartiles, respectively. The upper and lower whiskers range from the corresponding box hinges to the largest value no further than 1.5 times the interquartile range from the hinge. All outlying data points beyond the whiskers are plotted individually.

Supplementary Figure 3. Considering gene expression instead of splicing activities results in splicing cancer programs without cancer driver-like behavior.

(**a**). Differentially activated (FDR < 0.05, two-sided Wilcoxon rank sum test) splicing factors (x-axis) across cancer cohorts (y-axis). Color, splicing factor is activated (orange; oncogenic) or inactivated (blue; tumor suppressor) in tumors, or not differentially activated (gray). (**b**) Distributions of the number of cancer cohorts (y-axis) that each splicing factor is associated with a high or low risk of death (x-axis) considering different thresholds of differential activation recurrence to define cancer splicing programs. Color, cancer splicing program. Top, ****$p \leq 0.0001$, **$p \leq 0.01$ based on two-sided Wilcoxon rank sum tests. Bottom, number of splicing factors in the corresponding cancer splicing program. (**c**) Relationship between differential splicing factor protein activities (x-axis) and splicing factor gene expression fold-changes (y-axis) computed across 14 different types of cancer from TCGA comparing tumor samples to solid tissue normal (healthy) samples. Top, the Pearson correlation coefficient (R) and p-value (p). (**d**) Frequency of differentially (FDR < 0.05) upregulated (y-axis, top) and downregulated (y-axis, bottom) splicing factors (x-axis). We highlighted the top 5 splicing factors that were recurrently activated and inactivated in Fig. 3c. Horizontal dashed lines correspond to the selected recurrency threshold to be considered a splicing factor in either cancer splicing program. On top of the dashed line, the number of splicing factors selected as part of each gene expression-based cancer splicing program. (**e**) Distributions of the number of cancer cohorts (y-axis) that each splicing factor expression is associated with a high or low risk of death (x-axis) in gene expression-based cancer splicing programs defined with recurrently differentially expressed splicing factors. Color, cancer splicing program. Top, ns $p > 0.05$ based on two-sided Wilcoxon rank sum tests. Bottom, number of splicing factors in the corresponding gene expression-based cancer splicing program. (**f**) Distributions of the number of cancer cohorts (y-axis) that each splicing factor is associated with a high or low risk of death (x-axis) considering different thresholds of differential expression recurrence to define cancer splicing programs. Color, cancer splicing program. Top, *$p \leq 0.05$, ns $p > 0.05$ based on two-sided Wilcoxon rank sum tests. Bottom, number of splicing factors in the corresponding cancer splicing program. In box and whiskers plots in panels (**b**), (**e**), and (**f**), the median value is indicated as a horizontal line, and the lower and upper bounds of the box correspond to the first- and third quartiles, respectively. The upper and lower whiskers range from the corresponding box hinges to the largest value no further than 1.5 times the interquartile range from the hinge. All outlying data points beyond the whiskers are plotted individually.

Supplementary Figure 4. Activities of cancer splicing programs correlate with proliferative state marker MKI67 across CCLE cancer types.

Distributions of Pearson correlations between proliferative state marker MKI67 and median splicing factor activities (y-axis) of each cancer splicing program (x-axis). We considered the 22 cancer types with more than 10 cell lines in the database. Top, two-sided Wilcoxon rank sum test p-value. In box and whiskers plots, the median value is indicated as a horizontal line, and the lower and upper bounds of the box correspond to the first- and third quartiles, respectively. The upper and lower whiskers range from the corresponding box hinges to the largest value no further than 1.5 times the interquartile range from the hinge. All outlying data points beyond the whiskers are plotted individually.

FIGURES

Figure 1

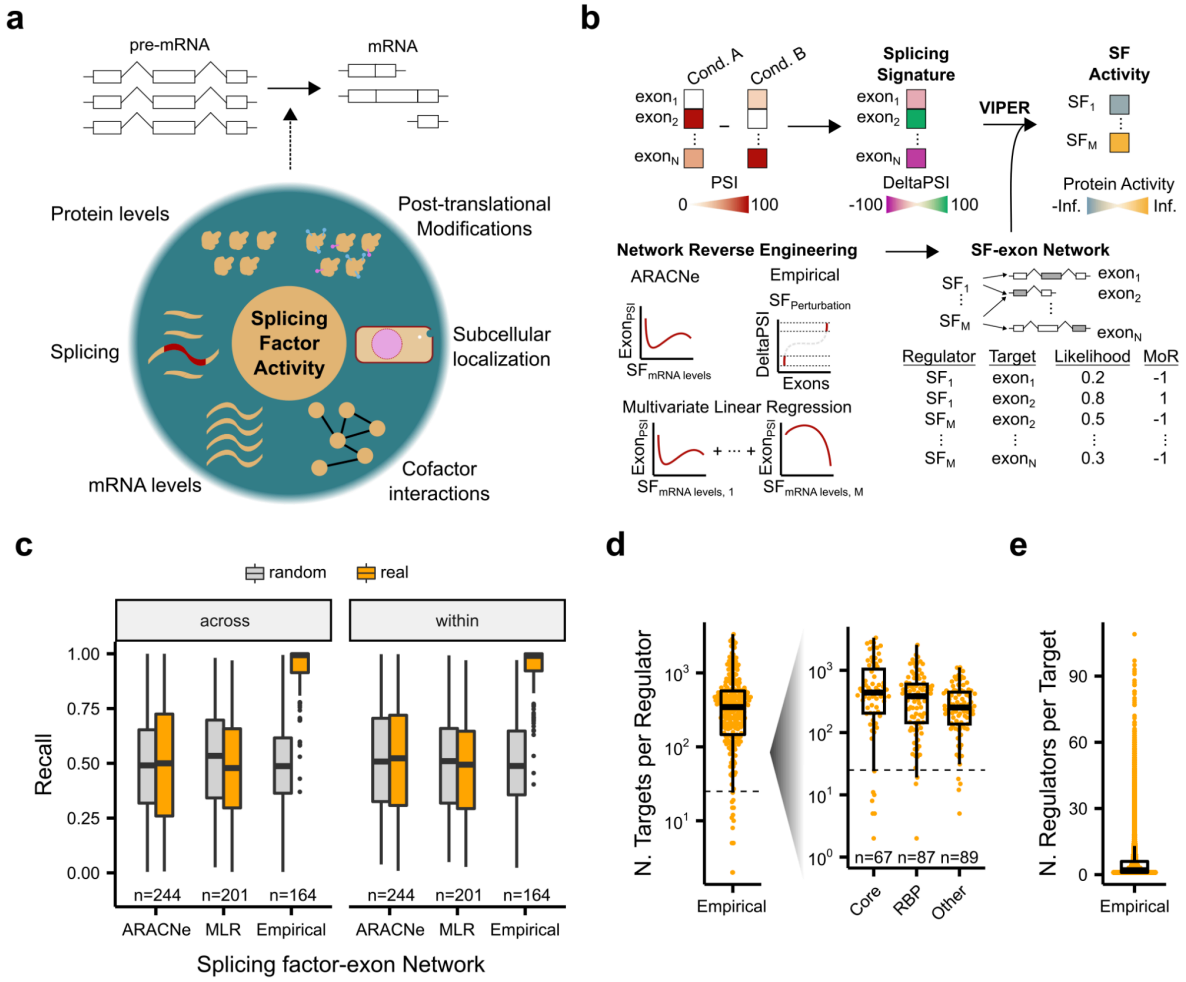


Figure 2

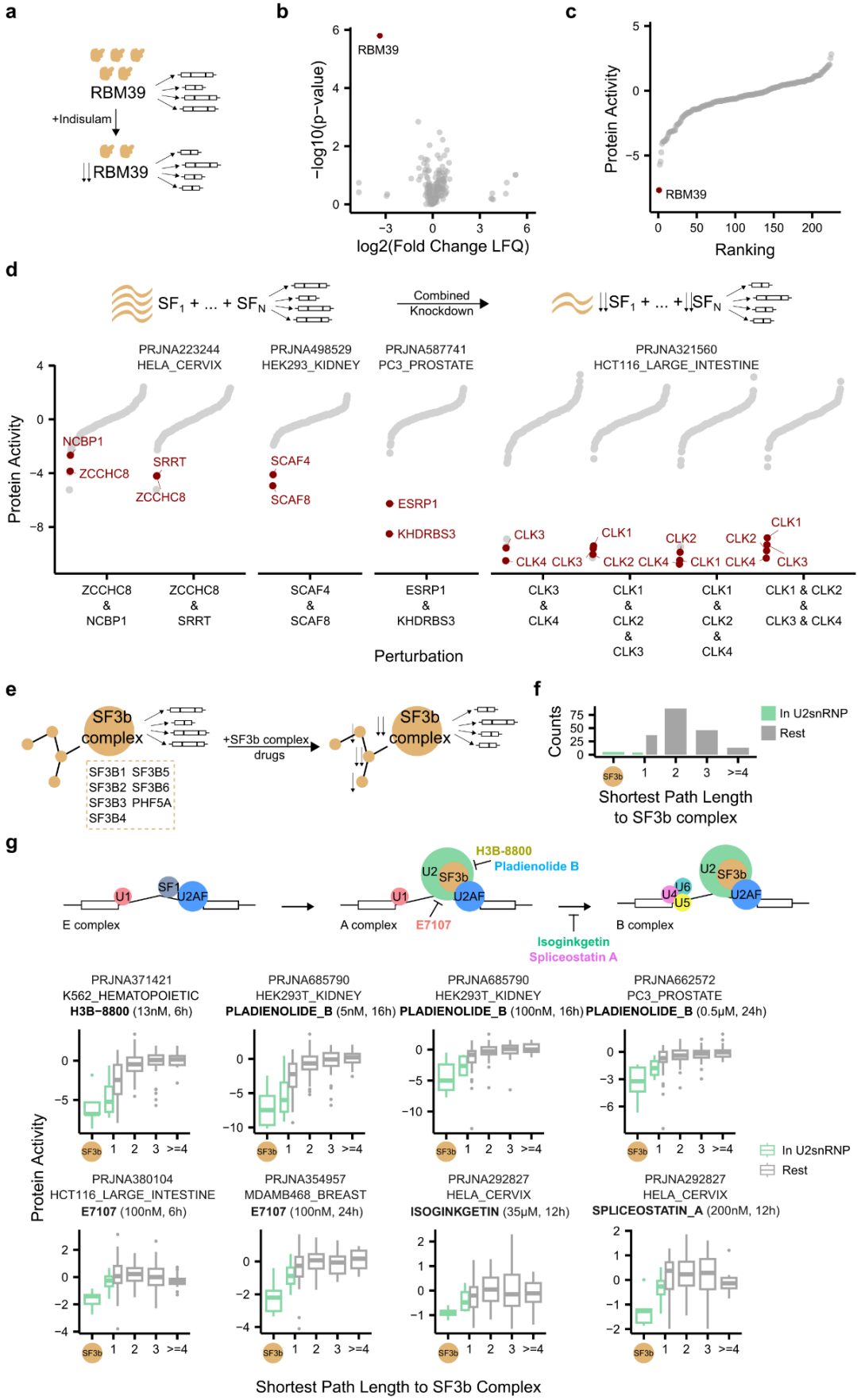

a
RBM39
+Indisulam
RBM39
b
RBM39
−log10(p−value)
log2(Fold Change LFQ)
c
Protein Activity
RBM39
Ranking
d
SF1 + ... + SFN
Combined Knockdown
PRJNA223244 HELA_CERVIX
PRJNA498529 HEK293_KIDNEY
PRJNA587741 PC3_PROSTATE
PRJNA321560 HCT116_LARGE_INTESTINE
Protein Activity
NCBP1
ZCCHC8
SRRT
ZCCHC8
SCAF4
SCAF8
ESRP1
KHDRBS3
CLK3
CLK4 CLK3
CLK1
CLK2 CLK4
CLK2
CLK1 CLK4
CLK2
CLK1
CLK3
ZCCHC8 & NCBP1
ZCCHC8 & SRRT
SCAF4 & SCAF8
ESRP1 & KHDRBS3
CLK3 & CLK4
CLK1 & CLK2 & CLK3
CLK1 & CLK2 & CLK4
CLK1 & CLK2 & CLK3 & CLK4
Perturbation
e
SF3b complex
SF3B1 SF3B5
SF3B2 SF3B6
SF3B3 PHF5A
SF3B4
+SF3b complex drugs
SF3b complex
f
Counts
In U2snRNP
Rest
SF3b 1 2 3 >=4
Shortest Path Length to SF3b complex
g
U1
SF1
U2AF
E complex
U1
U2
SF3b
U2AF
A complex
E7107
H3B-8800
Pladienolide B
Isoginkgetin
Spliceostatin A
U2
SF3b
U4
U6
U5
U2AF
B complex
PRJNA371421 K562_HEMATOPOIETIC H3B−8800 (13nM, 6h)
PRJNA685790 HEK293T_KIDNEY PLADIENOLIDE_B (5nM, 16h)
PRJNA685790 HEK293T_KIDNEY PLADIENOLIDE_B (100nM, 16h)
PRJNA662572 PC3_PROSTATE PLADIENOLIDE_B (0.5μM, 24h)
Protein Activity
In U2snRNP
Rest
PRJNA380104 HCT116_LARGE_INTESTINE E7107 (100nM, 6h)
PRJNA354957 MDAMB468_BREAST E7107 (100nM, 24h)
PRJNA292827 HELA_CERVIX ISOGINKGETIN (35μM, 12h)
PRJNA292827 HELA_CERVIX SPLICEOSTATIN_A (200nM, 12h)
Shortest Path Length to SF3b Complex

Figure 3

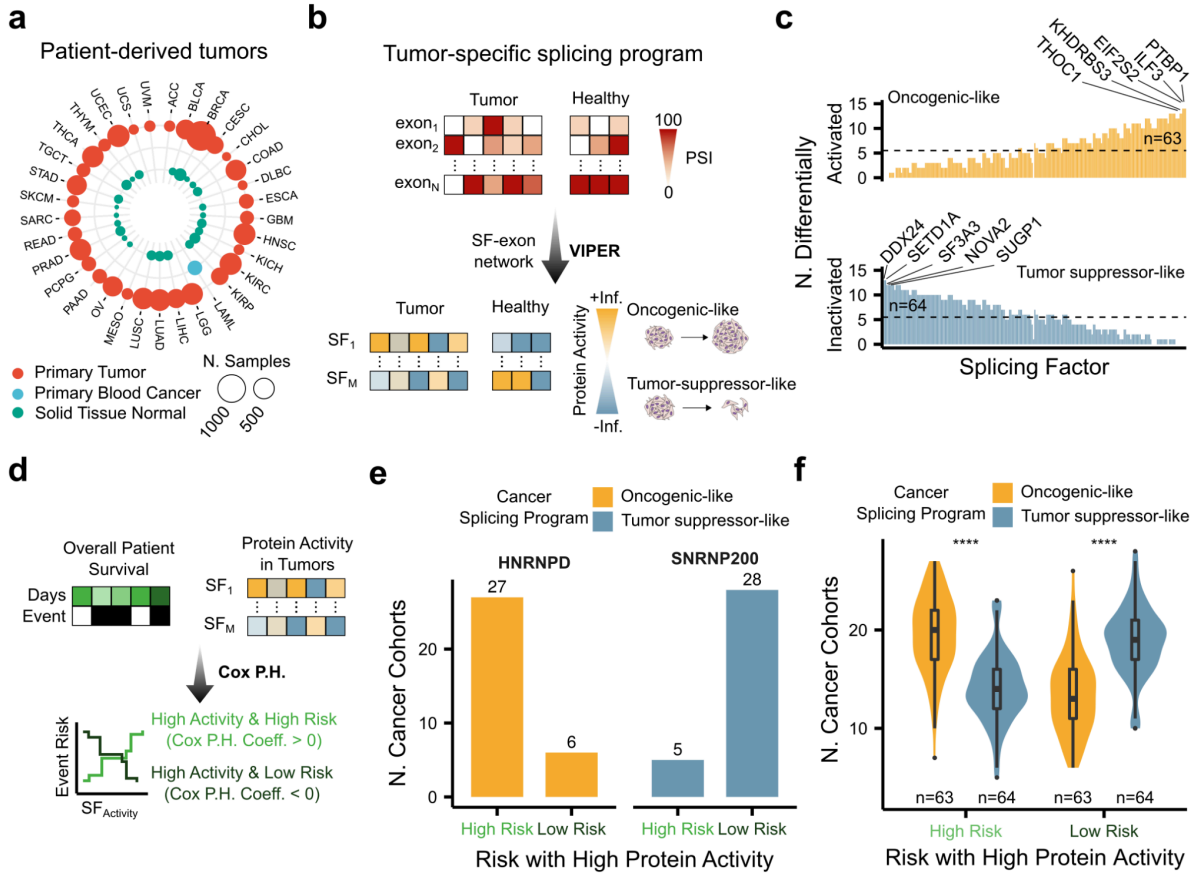
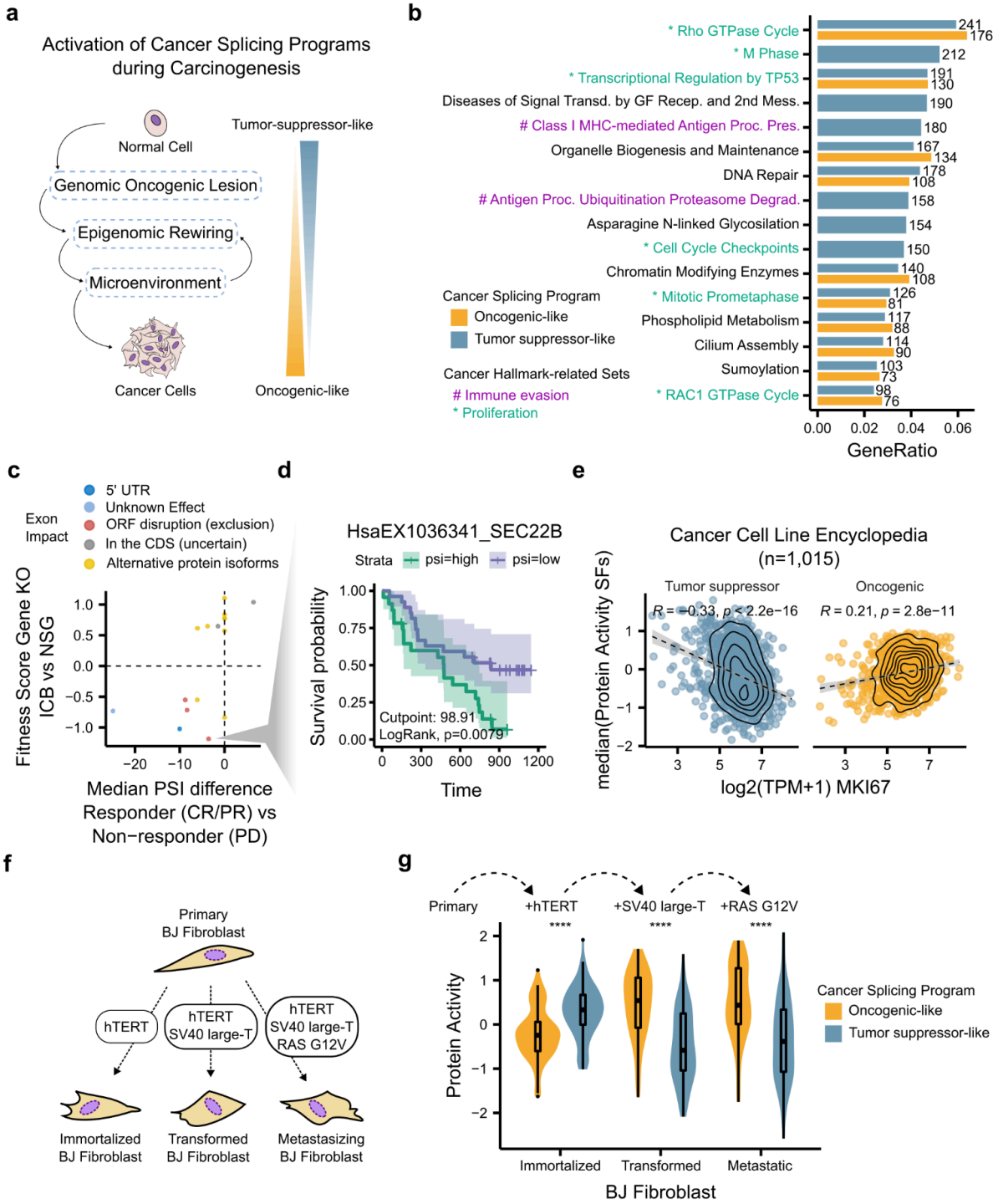
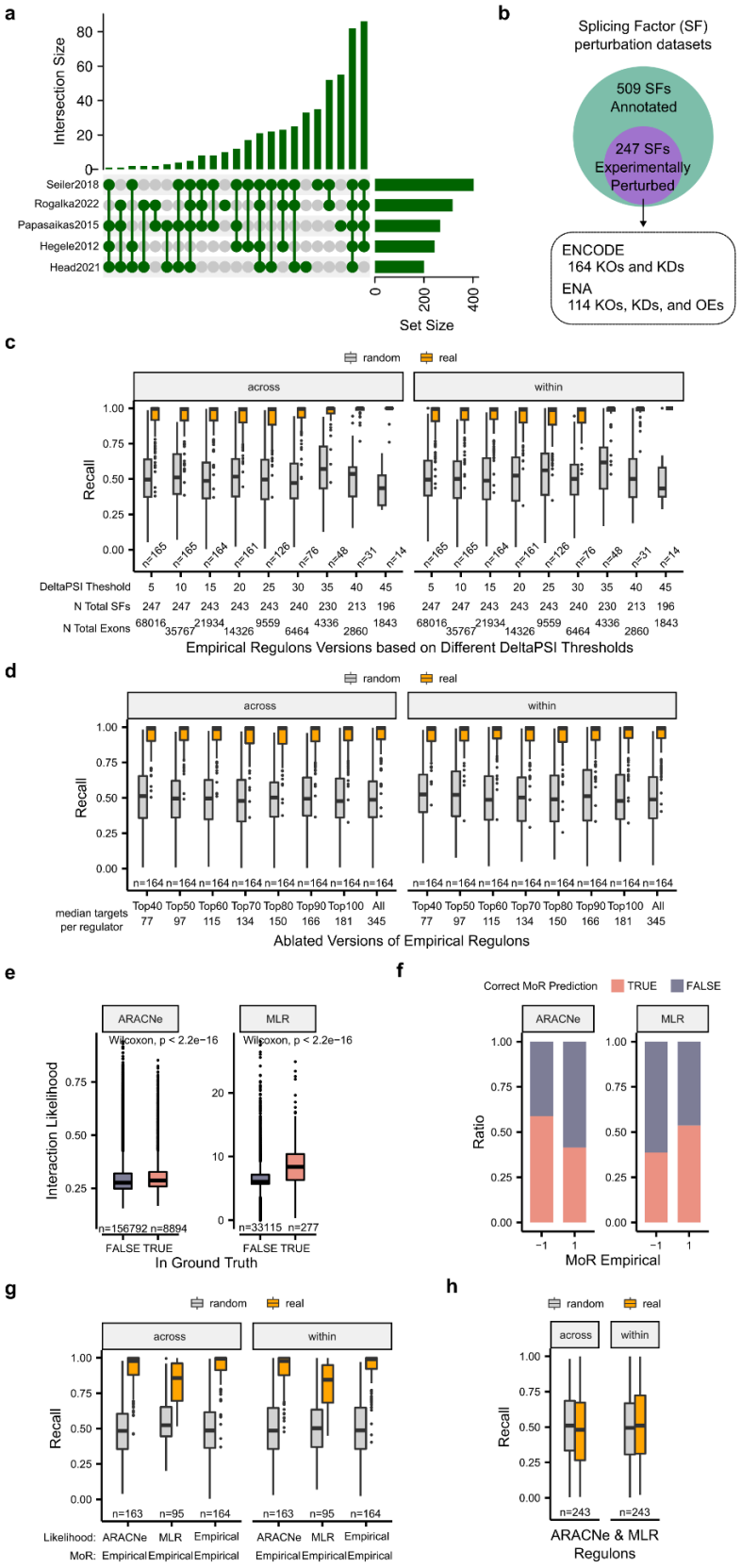

a
Patient-derived tumors
ACC
BLCA
BRCA
CESC
CHOL
COAD
DLBC
ESCA
GBM
HNSC
KICH
KIRC
KIRP
LAML
LGG
LIHC
LUAD
LUSC
MESO
OV
PAAD
PCPG
PRAD
READ
SARC
SKCM
STAD
TGCT
THCA
THYM
UCEC
UCS
UVM
Primary Tumor
Primary Blood Cancer
Solid Tissue Normal
N. Samples
1000
500
b
Tumor-specific splicing program
Tumor
Healthy
exon1
exon2
exonN
100
PSI
0
SF-exon network
VIPER
SF1
SFM
+Inf.
Protein Activity
-Inf.
Oncogenic-like
Tumor-suppressor-like
c
KHDRBS3
THOC1
EIF2S2
ILF3
PTBP1
Oncogenic-like
n=63
N. Differentially
Activated
Inactivated
DDX24
SETD1A
SF3A3
NOVA2
SUGP1
Tumor suppressor-like
n=64
Splicing Factor
d
Overall Patient Survival
Days
Event
Protein Activity in Tumors
SF1
SFM
Cox P.H.
Event Risk
SFActivity
High Activity & High Risk (Cox P.H. Coeff. > 0)
High Activity & Low Risk (Cox P.H. Coeff. < 0)
e
Cancer Splicing Program
Oncogenic-like
Tumor suppressor-like
HNRNPD
SNRNP200
27
6
5
28
N. Cancer Cohorts
High Risk
Low Risk
Risk with High Protein Activity
f
Cancer Splicing Program
Oncogenic-like
Tumor suppressor-like
****
****
N. Cancer Cohorts
n=63
n=64
n=63
n=64
High Risk
Low Risk
Risk with High Protein Activity

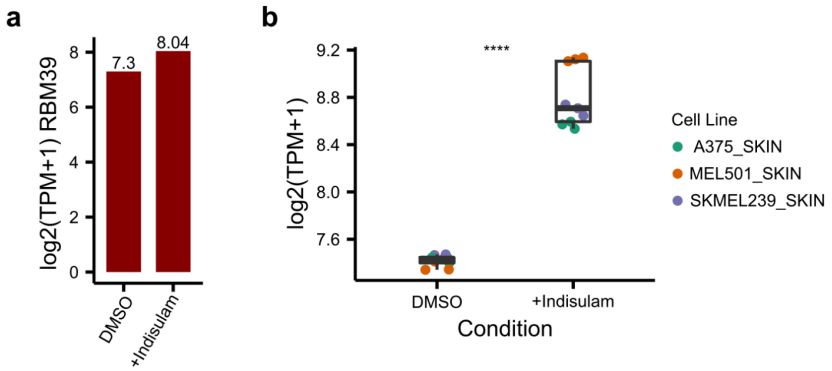
Figure 4


a
Activation of Cancer Splicing Programs during Carcinogenesis
Normal Cell
Genomic Oncogenic Lesion
Epigenomic Rewiring
Microenvironment
Cancer Cells
Tumor-suppressor-like
Oncogenic-like
b
* Rho GTPase Cycle
* M Phase
* Transcriptional Regulation by TP53
Diseases of Signal Transd. by GF Recep. and 2nd Mess.
Class I MHC-mediated Antigen Proc. Pres.
Organelle Biogenesis and Maintenance
DNA Repair
Antigen Proc. Ubiquitination Proteasome Degrad.
Asparagine N-linked Glycosilation
* Cell Cycle Checkpoints
Chromatin Modifying Enzymes
* Mitotic Prometaphase
Phospholipid Metabolism
Cilium Assembly
Sumoylation
* RAC1 GTPase Cycle
Cancer Splicing Program
Oncogenic-like
Tumor suppressor-like
Cancer Hallmark-related Sets
Immune evasion
* Proliferation
GeneRatio
c
Exon Impact
5' UTR
Unknown Effect
ORF disruption (exclusion)
In the CDS (uncertain)
Alternative protein isoforms
Fitness Score Gene KO ICB vs NSG
Median PSI difference Responder (CR/PR) vs Non−responder (PD)
d
HsaEX1036341_SEC22B
Strata psi=high psi=low
Survival probability
Cutpoint: 98.91
LogRank, p=0.0079
Time
e
Cancer Cell Line Encyclopedia (n=1,015)
Tumor suppressor
Oncogenic
R = −0.33, p < 2.2e−16
R = 0.21, p = 2.8e−11
median(Protein Activity SFs)
log2(TPM+1) MKI67
f
Primary BJ Fibroblast
hTERT
hTERT SV40 large-T
hTERT SV40 large-T RAS G12V
Immortalized BJ Fibroblast
Transformed BJ Fibroblast
Metastasizing BJ Fibroblast
g
Primary
+hTERT
+SV40 large-T
+RAS G12V
Protein Activity
Cancer Splicing Program
Oncogenic-like
Tumor suppressor-like
Immortalized
Transformed
Metastatic
BJ Fibroblast

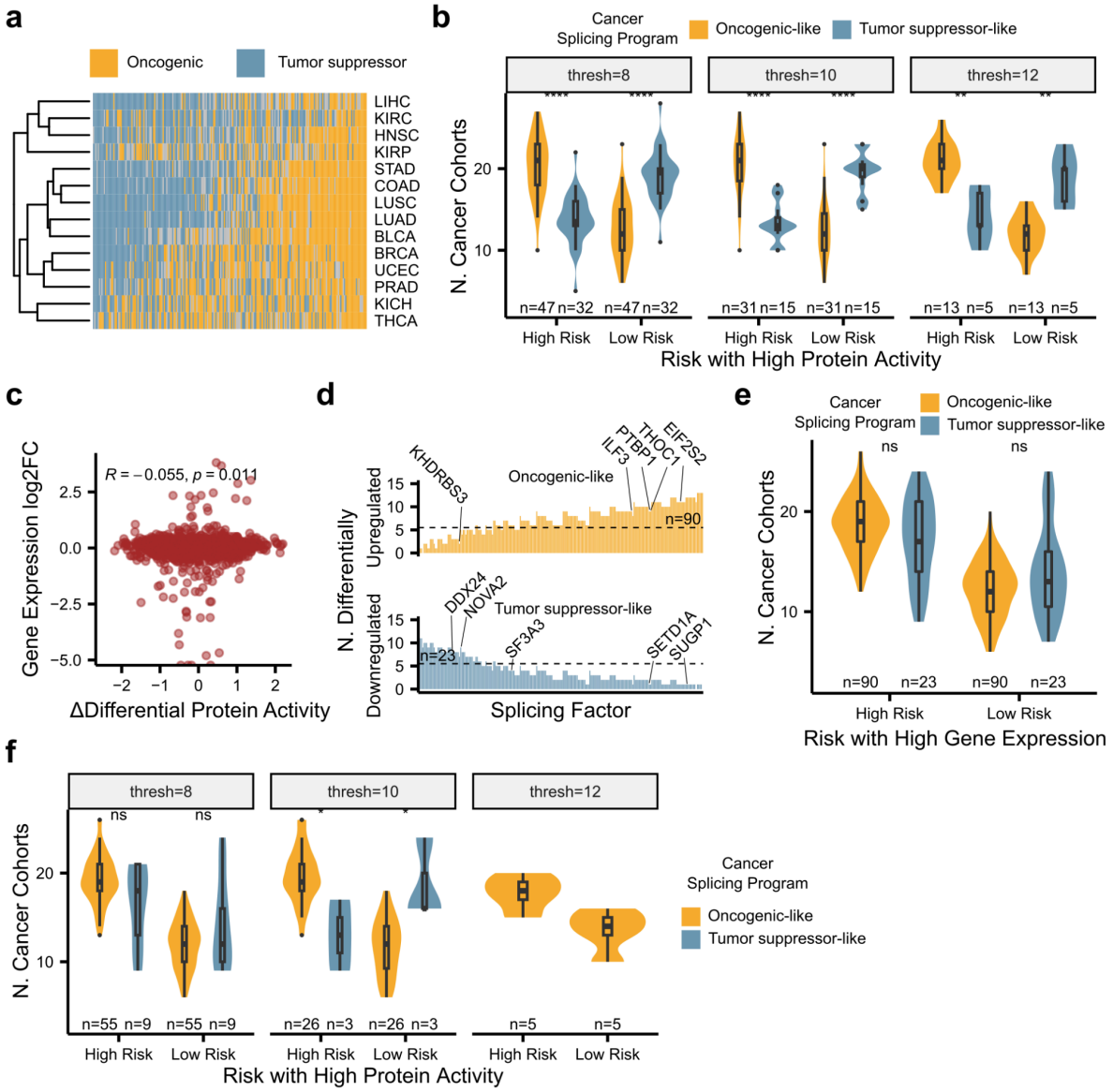
Supplementary Figure 1



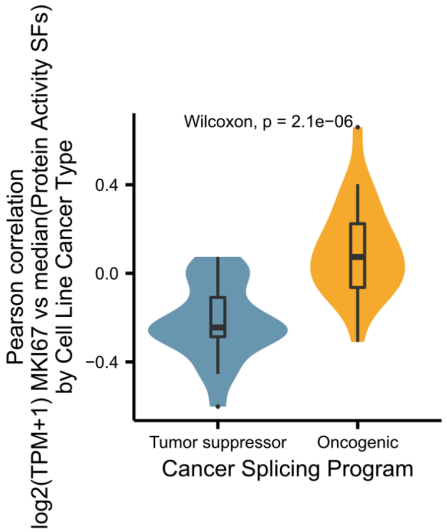
Supplementary Figure 2



Supplementary Figure 3



Supplementary Figure 4



SUPPLEMENTARY TABLES

Supplementary Table 1. Consensus list of 509 splicing factors.

Supplementary Table 2. Project identifiers of hand-curated datasets perturbing splicing factors available in ENA.

Supplementary Table 3. Cancer splicing programs.

REFERENCES

1. Bludau, I. & Aebersold, R. Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Nat. Rev. Mol. Cell Biol.* **21**, 327–340 (2020).
2. Torres-Méndez, A. *et al.* Parallel evolution of a splicing program controlling neuronal excitability in flies and mammals. *Science Advances* (2022) doi:10.1126/sciadv.abk0445.
3. Olivieri, J. E. *et al.* RNA splicing programs define tissue compartments and cell types at single-cell resolution. (2021) doi:10.7554/eLife.70692.
4. Bradley, R. K. & Anczuków, O. RNA splicing dysregulation and the hallmarks of cancer. *Nat. Rev. Cancer* **23**, 135–155 (2023).
5. Ciampi, L., Serrano, L. & Irimia, M. Unique transcriptomes of sensory and non-sensory neurons: insights from Splicing Regulatory States. *Mol. Syst. Biol.* (2024) doi:10.1038/s44320-024-00020-1.
6. Califano, A. & Alvarez, M. J. The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nat. Rev. Cancer* **17**, 116–130 (2016).
7. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).
8. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**, 1363–1375 (2019).
9. Margolin, A. A. *et al.* ARACNE: An Algorithm for the Reconstruction of Gene Regulatory

Networks in a Mammalian Cellular Context. *BMC Bioinformatics* **7**, 1–15 (2006).
10. Lachmann, A., Giorgi, F. M., Lopez, G. & Califano, A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32**, 2233–2235 (2016).
11. Garcia-Alonso, L. *et al.* Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer. *Cancer Res.* **78**, 769–780 (2018).
12. Müller-Dott, S. *et al.* Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Res.* **51**, 10934–10949 (2023).
13. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
14. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 (2016).
15. Badia-I-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinform Adv* **2**, vbac016 (2022).
16. Zhang, C. *et al.* Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science* **329**, 439–443 (2010).
17. Aghamirzaie, D., Collakova, E., Li, S. & Grene, R. CoSpliceNet: a framework for co-splicing network inference from transcriptomics data. *BMC Genomics* **17**, 845 (2016).
18. Chen, L. & Zheng, S. Studying alternative splicing regulatory networks through partial correlation analysis. *Genome Biol.* **10**, R3 (2009).
19. Cardoso-Moreira, M. *et al.* Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).
20. Papasaikas, P., Tejedor, J. R., Vigevani, L. & Valcárcel, J. Functional splicing network reveals extensive regulatory potential of the core spliceosomal machinery. *Mol. Cell* **57**, 7–22 (2015).
21. Seiler, M. *et al.* Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. *Cell Rep.* **23**, 282–296.e4 (2018).
22. Rogalska, M. E., Vivori, C. & Valcárcel, J. Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. *Nat. Rev. Genet.* **24**, 251–269 (2023).
23. Head, S. A. *et al.* Silencing of SRRM4 suppresses microexon inclusion and promotes tumor growth across cancers. *PLoS Biol.* **19**, e3001138 (2021).
24. Hegele, A. *et al.* Dynamic protein-protein interaction wiring of the human spliceosome. *Mol. Cell* **45**, 567–580 (2012).
25. Nijhuis, A. *et al.* Indisulam targets RNA splicing and metabolism to serve as a therapeutic strategy for high-risk neuroblastoma. *Nat. Commun.* **13**, 1380 (2022).
26. Lu, S. X. *et al.* Pharmacologic modulation of RNA splicing enhances anti-tumor immunity. *Cell* **184**, 4032–4047.e31 (2021).
27. Funnell, T. *et al.* CLK-dependent exon recognition and conjoined gene formation revealed with a novel small molecule inhibitor. *Nat. Commun.* **8**, 7 (2017).
28. Zhang, D. *et al.* Intron retention is a hallmark and spliceosome represents a therapeutic vulnerability in aggressive prostate cancer. *Nat. Commun.* **11**, 2089 (2020).
29. Gregersen, L. H. *et al.* SCAF4 and SCAF8, mRNA Anti-Terminator Proteins. *Cell* **177**, 1797–1813.e18 (2019).
30. Andersen, P. R. *et al.* The human cap-binding complex is functionally connected to the nuclear RNA exosome. *Nat. Struct. Mol. Biol.* **20**, 1367–1376 (2013).
31. Sun, C. The SF3b complex: splicing and beyond. *Cell. Mol. Life Sci.* **77**, 3583–3595 (2020).
32. Seiler, M. *et al.* H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. *Nat. Med.* **24**, 497–504 (2018).
33. Cretu, C. *et al.* Structural Basis of Splicing Modulation by Antitumor Macrolide Compounds. *Mol. Cell* **70**, 265–273.e8 (2018).
34. Folco, E. G., Coil, K. E. & Reed, R. The anti-tumor drug E7107 reveals an essential role for SF3b in remodeling U2 snRNP to expose the branch point-binding region. *Genes*

Dev. **25**, 440–444 (2011).
35. Roybal, G. A. & Jurica, M. S. Spliceostatin A inhibits spliceosome assembly subsequent to prespliceosome formation. *Nucleic Acids Res.* **38**, 6664–6672 (2010).
36. O'Brien, K., Matlin, A. J., Lowell, A. M. & Moore, M. J. The biflavonoid isoginkgetin is a general inhibitor of Pre-mRNA splicing. *J. Biol. Chem.* **283**, 33147–33154 (2008).
37. Chowell, D. *et al.* Improved prediction of immune checkpoint blockade efficacy across multiple cancer types. *Nat. Biotechnol.* **40**, 499–506 (2022).
38. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).
39. Riaz, N. *et al.* Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* **171**, 934–949.e16 (2017).
40. Dubrot, J. *et al.* In vivo CRISPR screens reveal the landscape of immune evasion pathways across cancer. *Nat. Immunol.* **23**, 1495–1506 (2022).
41. Alloatti, A. *et al.* Critical role for Sec22b-dependent antigen cross-presentation in antitumor immunity. *J. Exp. Med.* **214**, 2231–2241 (2017).
42. Sun, X. & Kaufman, P. D. Ki-67: more than a proliferation marker. *Chromosoma* **127**, 175–186 (2018).
43. Pitot, H. C. The molecular biology of carcinogenesis. *Cancer* **72**, 962–970 (1993).
44. Danielsson, F. *et al.* Majority of differentially expressed genes are down-regulated during malignant transformation in a four-stage model. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 6853–6858 (2013).
45. Hahn, W. C. *et al.* Creation of human tumour cells with defined genetic elements. *Nature* **400**, 464–468 (1999).
46. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
47. Cilloni, D., Itri, F., Bonuomo, V. & Petiti, J. Mutations in Hematological Malignancies. *Cancers* **14**, (2022).
48. Maurin, M. *et al.* RBFOX2 deregulation promotes pancreatic cancer progression and metastasis through alternative splicing. *Nat. Commun.* **14**, 8444 (2023).
49. Bielli, P., Bordi, M., Di Biasio, V. & Sette, C. Regulation of BCL-X splicing reveals a role for the polypyrimidine tract binding protein (PTBP1/hnRNP I) in alternative 5' splice site selection. *Nucleic Acids Res.* **42**, 12070–12081 (2014).
50. Koedoot, E. *et al.* Co-regulated gene expression of splicing factors as drivers of cancer progression. *Sci. Rep.* **9**, 5484 (2019).
51. Moakley, D. F. *et al.* Reverse engineering neuron type-specific and type-orthogonal splicing-regulatory networks using single-cell transcriptomes. *bioRxiv* 2024.06.13.597128 (2024) doi:10.1101/2024.06.13.597128.
52. Tapial, J. *et al.* An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res.* **27**, 1759–1768 (2017).
53. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
54. Ding, H. *et al.* Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nat. Commun.* **9**, 1471 (2018).
55. Kim Guisbert, K. S., Mossiah, I. & Guisbert, E. Titration of SF3B1 Activity Reveals Distinct Effects on the Transcriptome and Cell Physiology. *Int. J. Mol. Sci.* **21**, (2020).
56. Matsumoto, K. *et al.* G-quadruplex-forming nucleic acids interact with splicing factor 3B subunit 2 and suppress innate immune gene expression. *Genes Cells* **26**, 65–82 (2021).
57. Teng, T. *et al.* Splicing modulators act at the branch point adenosine binding pocket defined by the PHF5A-SF3b complex. *Nat. Commun.* **8**, 15522 (2017).
58. Chan, S. *et al.* Basal-A Triple-Negative Breast Cancer Cells Selectively Rely on RNA Splicing for Survival. *Mol. Cancer Ther.* **16**, 2849–2861 (2017).
59. Tseng, C.-K. *et al.* Human Telomerase RNA Processing and Quality Control. *Cell Rep.* **13**, 2232–2243 (2015).
60. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased

coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
61. Cvitkovic, I. & Jurica, M. S. Spliceosome database: a tool for tracking components of the spliceosome. *Nucleic Acids Res.* **41**, D132–41 (2013).
62. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

ACKNOWLEDGEMENTS

We would like to thank Luca Zanella, Lucas ZhongMing Hu, Alessandro Vasciaveo, Alec Wang for their support in the reimplementation of VIPER. We would like to thank Sophie Bonnal, Malgorzata Ewa Rogalska, Juan Valcárcel, Manuel Irimia, Yamile Ana, Javier Gonzalez-de-Miguel and Xavier Hernandez-Alias for the insightful discussions regarding the elucidation of splicing factor programs. Also, we would like to thank Hector C Keun, Clare Eckold and Anke Nijhuis for kindly sharing proteomics sample metadata and raw RNA sequencing data from cells treated with indisulam. We acknowledge support of the Spanish Ministry of Science and Innovation through the Centro de Excelencia Severo Ochoa (CEX2020-001049-S, MCIN/AEI /10.13039/501100011033), and the Generalitat de Catalunya through the CERCA programme and to the EMBL partnership. We are grateful to the CRG Core Technologies Programme for their support and assistance in this work. The results shown here are in part based upon data generated by the TCGA Research Network (https://www.cancer.gov/tcga).

FUNDING

This project was funded by grants from the Plan Estatal de Investigación Científica y Técnica y de Innovación to LS (PID2021-122341NB-I00 project funded MICIU / AEI / 10.13039 / 501100011033 / FEDER, UE). Further, was also supported by the NCI U54A274506 (Center for Cancer Systems Therapeutics, CaST), the NCI Outstanding Investigator Award R35 CA197745, and the NIH Shared Instrumentation Grants S10 0D012351 and S10 OD021764 and S10 0D032433 all to AC. In addition, CZ was supported by grants R01NS125018 and R35GM145279. And, MA-G was supported in part by Boehringer Ingelheim Fonds travel grant.

AUTHOR INFORMATION

Conceptualization: MA-G, SM-V, AC, LS; Methodology: MA-G, DFM, SM-V, AC, LS; Software: MA-G, DFM, CZ; Validation: MA-G, DFM, SM-V, AC, LS; Formal analysis: MA-G, DFM, SM-V, AC, LS; Investigation: MA-G; Resources: AC, LS; Data Curation: MA-G; Writing - original draft: MA-G; Writing - review & editing: MA-G, DFM, SM-V, AC, LS; Visualization: MA-G; Supervision: SM-V, AC, LS; Project administration: MA-G; Funding acquisition: MA-G, CZ, AC, LS. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

Dr. Califano is founder, equity holder, and consultant of DarwinHealth Inc., a company that has licensed some of the algorithms used in this manuscript from Columbia University. Columbia University is also an equity holder in DarwinHealth Inc.

The rest of authors declare no conflicts of interest.