Exon inclusion signatures enable accurate estimation of splicing factor activity

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

Splicing factors control exon inclusion in messenger RNA, shaping transcriptome and proteome diversity. Their catalytic activity is regulated by multiple layers, making single-omic measurements on their own fall short in identifying which splicing factors underlie a phenotype. Here, we propose splicing factor activity can be estimated by interpreting changes in exon inclusion. We benchmark methods to construct splicing factor→exon networks and calculate activity. Combining RNA-seq perturbation-based networks with VIPER (virtual inference of protein activity by enriched regulon analysis) accurately captures splicing factor activation modulated by different regulatory layers. This approach consolidates splicing factor regulation into a single score derived solely from exon inclusion signatures, allowing functional interpretation of heterogeneous conditions. As a proof of concept, we identify recurrent cancer splicing programs, revealing oncogenic- and tumor suppressor-like splicing factors missed by conventional methods. These programs correlate with patient survival and key cancer hallmarks: initiation, proliferation, and immune evasion. Altogether, we show splicing factor activity can be accurately estimated from exon inclusion changes, enabling comprehensive analyses of splicing regulation with minimal data requirements.

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

INTRODUCTION

Splicing factors carry out alternative splicing: they selectively regulate intron removal from expressed precursor mRNAs. Together with alternative transcription start and termination sites, splicing factors determine each gene's pool of mature mRNAs, impacting overall mRNA stability, translation, protein activation, and protein-protein interactions[1]. Simultaneously, splicing factor activity is 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. 1). Accurate assessment of splicing factor activity is thus challenging as it requires accounting for highly diverse regulatory mechanisms.

Splicing factor activities have been typically inferred from the presence of activating and inactivating somatic mutations or differential gene expression[2–5]. However, these represent just two of the many possible mechanisms contributing to splicing factor activity modulation. This partial picture limits identifying aberrant splicing factor activity, especially in complex genetic environments, such as cancer, where multiple molecular alterations contribute to cancer hallmarks[4,6–9]. A possible solution could be collecting more omic measurements on splicing factors. However, this would not only require large-scale efforts and resources but also methods to integrate these data and calculate splicing factor activities, which are still quite elusive.

The challenge of assessing protein activity has been effectively addressed for transcription factors (see Garcia-Alonso *et al.*[10] for a recent review). Rather than relying on the omics of a transcription factor to estimate its activity, focusing on the expression of its target genes reflects the combined influence of regulatory and signaling mechanisms that determine activity. Exploiting this strategy, however, requires addressing two challenges: first assessing the transcriptional targets of every transcription factor, and, second, estimating their activity based on the differential expression of these targets. Transcription factor→gene interactions (i.e., regulatory networks) can be inferred computationally from gene expression datasets by reverse engineering algorithms such as ARACNe[11,12], as well as empirically, by mining curated databases generated from publicly available experiments[13–15]. Then, given a high-quality regulatory network that includes the probability (likelihood) and directionality (mode of regulation) of each interaction, transcription factor activity can be accurately assessed from expression signatures through bioinformatic methods like VIPER[16] and others[17]. 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].

Methods for estimating protein activity can detect altered transcription factor activity from target gene expression signatures despite complex regulatory mechanisms. In line with this, we posit exon inclusion signatures report the activity of splicing factors. Frameworks like VIPER could be extended to estimate splicing factor activity from changes in exon inclusion across their target mRNAs, represented as networks of splicing factor→exon interactions (splicing factor networks, for simplicity) (Fig. 2a). However, it remains unclear which approaches for network construction and activity estimation best adapt protein activity methods to splicing factors, and whether this can uncover novel insights from heterogeneous data.

Here, to repurpose protein activity analysis methods for splicing factors, we benchmark different methods for splicing factor network inference and activity estimation. Combining empirically derived splicing factor networks with the VIPER framework accurately estimates splicing factor activity in our benchmark datasets. Additionally, this approach enables accurate estimation of splicing factor activity in experiments of increasing complexity compared to those used to generate splicing factor networks. To illustrate its utility in analyzing heterogeneous molecular conditions, we apply this framework to cancer-related phenotypes, revealing two biologically relevant and complementary splicing programs overlooked by conventional splicing factor expression analysis. Altogether, splicing factor activity analysis is a powerful method for identifying splicing factors that behave as mechanistic determinants of complex cellular phenotypes.

RESULTS

Benchmark datasets for splicing factor network inference and activity estimation

To evaluate how different splicing factor network inference approaches work in combination with different activity estimation methods, we developed a set of benchmark datasets. We generated a consensus list of 509 splicing factors documented in landmark literature (Supplementary Fig. 1a; Supplementary Table 1)[8,20–23]. We then compiled and uniformly processed bulk RNA-seq data from the ENCODE and ENA databases, resulting in a total of 474 experiments that perturb 248 distinct splicing factors across 39 cell lines and 54 independent studies (Supplementary Fig. 1b). For each experiment, we computed exon inclusion signatures (delta PSI) by comparing perturbed versus unperturbed exon inclusion profiles and recorded the splicing factor that was perturbed.

The splicing signatures were divided into five benchmark datasets based on their source databases: ENA (n=161), ENCODE-KD-HepG2 (n=104), ENCODE-KO-HepG2 (n=44), ENCODE-KD-K562 (n=104), and ENCODE-KO-K562 (n=61). These benchmark datasets enable computing performance metrics such as the area under the receiver operating curve (ROC-AUC) to quantify how well the predicted splicing factor activities recapitulate the splicing factor responsible for the observed splicing signatures. ROC-AUCs can be calculated along two axes for each dataset: across splicing factors and across perturbations. These axes indicate the predictive power of the estimated splicing factor activities in identifying which experiments perturbed a specific splicing factor and which splicing factor was perturbed in each experiment, respectively (Fig. 2b).

Empirical splicing factor networks combined with VIPER outperform computational ones in activity estimation

Traditionally, splicing factor networks have been reverse-engineered computationally by assessing pairwise associations between splicing factor expression and exon inclusion across multiple samples[24]. To replicate this, we adapted ARACNe[11,12] –an algorithm that calculates pairwise mutual information between splicing factor expression (TPM) and exon inclusion (PSI)– to infer splicing factor networks using RNA-seq datasets with increasing sample sizes ($n_{CardosoMoreira2020}$ = 313, $n_{PANCAN\text{-}STN}$ = 736, and $n_{PANCAN\text{-}PT}$ = 9,854) (see Methods). This resulted in three computationally-inferred splicing factor networks for the 509 annotated splicing factors. These networks consist of splicing factor→exon interactions with their associated likelihood (from 0 to 1) and mode of regulation (positive or negative).

We evaluated network quality for splicing factor activity estimation using four methods to compute protein activity: normalized enrichment scores (NES) from gene set enrichment analysis (GSEA), Spearman and Pearson correlations, and VIPER. Each method interprets splicing factor networks differently. GSEA, a widespread approach, uses interactions in an ontology (here, splicing factor networks) to compute NES on exon inclusion signatures, but it does not consider likelihoods or modes of regulation. In correlation-based methods, we incorporated likelihood and mode of regulation by multiplying them into an interaction summary value and computing a correlation coefficient (here, activity) between exon inclusion signatures and each splicing factor's target exons interaction summary values. Finally, VIPER combines aspects of both approaches by using interaction likelihoods and

modes of regulation in an enriched rank-based analysis (aREA) to estimate splicing factor activity from exon inclusion signatures.

The ARACNe-based networks resulted in splicing factor activity estimates that were close to random, with slightly better performance when more samples were used, especially when combined with VIPER (Fig. 2c). This indicates that pairwise statistical dependencies between splicing factor expression and exon inclusion are not sufficient for effective network inference.

To address this limitation, we inferred splicing factor networks using multivariate linear regression (MLR) models, which simultaneously associate expression of all splicing factors with exon inclusion, potentially accounting for combinatorial regulation (see Methods). However, this computational approach also performed poorly, yielding near-random ROC-AUC distributions in most cases, except for the largest dataset, where performance improved slightly (ROC-AUC ~ 0.6) (Fig. 2c).

Given the limitations of computational methods, we turned to empirical splicing factor networks. These networks designate exons as targets if their inclusion levels change by more than a defined threshold following splicing factor perturbation. Although such target exons may result from both direct and indirect interactions, they are functionally linked to their respective splicing factor's activity. We based empirical networks on perturbation experiments used to create benchmark datasets, cataloging exon inclusion changes associated with 248 of the 509 annotated splicing factors. Here, we set the magnitude of delta PSI as the interaction likelihood and the sign as the mode of regulation. To select an optimal delta PSI threshold, we generated and evaluated nine networks across a range of delta PSI cutoffs (from 5 to 45% > |delta PSI|) (Supplementary Fig. 1c). To avoid overfitting, we excluded interactions derived from the benchmark dataset used for evaluation.

We found using a 15% delta PSI threshold produced networks that, when combined with VIPER, achieved highly accurate activity estimations (median ROC-AUC = 0.91 across both axes) while limiting the number of target exons (n = 22,049) (Supplementary Fig. 1c). These empirical networks outperformed both computational networks and other activity estimation methods, indicating that combining empirical splicing factor networks with VIPER provides the best approach for accurate splicing factor activity estimation (Fig. 2c).

Stress testing splicing factor activity estimation with empirical networks

Despite promising results, estimating splicing factor activity using empirical networks with VIPER may still be affected by technical and biological artifacts.

From a technical standpoint, splicing changes are often highly context-specific[25], which could lead to batch effects if empirical splicing factor networks are derived from different cell lines or studies. To test robustness, we evaluated held-hout benchmark datasets independently. VIPER combined with empirical networks consistently achieved the highest ROC-AUC distributions, indicating resilience to batch effects (Supplementary Fig. 1d). As an extreme example, VIPER with ENCODE-derived networks (from 2 cell lines) achieved median ROC-AUC values of 0.90 (across splicing factors) and 0.91 (across perturbations) on the ENA benchmark dataset (37 different cell lines). Only the ENCODE-KO-HepG2 dataset

using Pearson correlations showed higher ROC-AUC than VIPER. Altogether, this suggests that combining perturbation-based networks with VIPER provides robust splicing factor activity estimation.

We further validated robustness by retaining only the top interactions in empirical networks based on likelihood. Reducing target exons by up to 73.9% did not substantially alter evaluation metrics for VIPER-assessed activities, mirroring findings from transcription factor activity benchmarks using this method (Supplementary Fig. 1e).

Biologically, cross-regulation among splicing factors may impact activity estimation. Overall, considering the ~140,000 annotated exons in VastDB, empirical splicing factor networks show low connectivity as each exon is regulated by a median of two splicing factors, and each splicing factor regulates a median of 300 exons (Fig. 2d). This indicates splicing changes from each splicing factor perturbation are highly specific, enhancing the accuracy of the resulting splicing factor networks.

Given the functional differences between splicing factor classes, we examined how connectivity in empirical networks correlates with their predictive power. Core factors are more connected than RNA-binding proteins (RBPs) or other splicing factors (Fig. 2e). Similarly, activity estimates for core factors reach higher ROC-AUC scores than the rest (Fig. 2f). This suggests that the exon repertoire of core splicing factors is more stable across conditions than those of RBPs or other splicing factors, which are expected to exhibit more context-dependent regulation[26].

Indirect interactions dominate empirical networks, with VIPER ensuring robust activity estimation

A possible limitation of our perturbation-based empirical splicing factor networks is that direct splicing factor→exon interactions are indistinguishable from indirect ones, potentially affecting performance. For some splicing factors, we can distinguish direct from indirect splicing factor→exon interactions. Crosslinking and immunoprecipitation (CLIP) assays identify RNA sequences physically bound by RBPs, providing direct interaction data. Since most splicing factors are RBPs, we explored whether identifying direct interactions in empirical splicing factor networks using CLIP data influences performance. Using POSTAR3[27], we generated CLIP-based networks for the 97 RBP splicing factors found. We included all splicing factor→exon interactions where CLIP peaks fell within 500 base pairs (bp) of an exon (Supplementary Fig. 2a). Only 11,626 of the 1,891,625 CLIP interactions overlapped with the 121,763 perturbation-based interactions, suggesting that most direct RBP-RNA interactions do not cause strong exon inclusion changes and that most changes in exon inclusion caused by splicing factor perturbations are likely due to indirect interactions (Supplementary Fig. 2b).

In terms of performance, CLIP-based networks alone produced near-random ROC-AUC scores. But, empirical networks containing only CLIP-verified (direct) interactions showed comparable performance to the original empirical networks using either correlation-based or VIPER for activity estimation (Supplementary Fig. 2c). Empirical networks with only non-CLIP-verified (indirect) interactions showed more variable ROC-AUC scores. Of note, only VIPER maintained median ROC-AUC scores of 0.8 or above when estimating activity

from empirical networks of indirect interactions (Supplementary Fig. 2c). These findings confirm that, when possible, identifying direct targets can enhance activity estimation, and demonstrate VIPER's robustness in accurately estimating splicing factor activity, even with empirical networks containing only indirect interactions.

Poor mode of regulation prediction limits performance of computational networks

Given the quality of predictions based on the perturbation-derived empirical splicing factor networks, we wondered what may be causing the poor performance of computational approaches. For this benchmark, we used computational splicing factor networks inferred from the largest dataset ($n_{PANCAN\text{-}PT}$ = 9,854) as they showed the best performance. Splicing factor network interactions are assigned both a likelihood score (from 0 to 1) and a mode of regulation (positive or negative) (Fig. 2a). Computational methods successfully assigned higher likelihood to splicing factor→exon interactions that were observed in perturbation experiments, suggesting this may not be the source of poor performance (Supplementary Fig. 2d). Instead, the mode of regulation, which predicts whether an exon is included or excluded, was inferred almost randomly, with a prediction error close to 50% (Supplementary Fig. 2e).

To validate this, we tested whether combining ARACNe- or MLR-based likelihoods with experimentally derived modes of regulation would improve performance. Indeed, this approach boosted ROC-AUC values, bringing them closer to those achieved by empirical networks (Supplementary Fig. 2f).

These findings suggest that computational methods struggle to accurately assess exon inclusion versus exon exclusion when splicing factor activity changes, leaving room for future improvement.

Estimated splicing factor activity recapitulates complex regulatory mechanisms

We have demonstrated that perturbation-derived splicing factor networks combined with VIPER can identify which splicing factor was perturbed under similar experimental conditions. However, we propose that this approach goes beyond that, offering the ability to functionally capture complex multi-omic regulatory mechanisms that ultimately influence splicing factor activity solely through exon inclusion signature analysis (Fig. 1). To illustrate this, we evaluate splicing factor activity analysis across several experimental validations involving perturbations of diverse regulatory mechanisms.

For example, the drug Indisulam downregulates the protein level of splicing factor RBM39 by mediating its proteasomal degradation (Fig. 3a and b)[28,29]. In contrast, the mRNA levels of its encoding gene increase consistently upon treatment, probably as a compensatory mechanism (Supplementary Fig. 3)[28,29]. 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. 3c). 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[30–33] (Fig. 3d). 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 our approach (Fig. 3d).

Splicing factors interact to form a large and dynamic protein complex called the spliceosome[23]. We thus hypothesized that the VIPER-assessed activity of splicing factors interacting in the spliceosome complex via protein-protein interactions (PPIs) should be 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)[8]. Within the U2 snRNP, the SF3b complex (SF3B1, SF3B2, SF3B3, SF3B4, SF3B5, SF3B6, PHF5A) plays a pivotal role in branch point recognition[34]. Different splicing drugs perturb this step of the reaction. For instance, H3B-8800 and Pladienolide B bind SF3B1[35,36]; E7107 prevents the binding of U2 snRNP to RNA[37]; and Isoginkgetin and Spliceostatin A prevent the transition from A to B spliceosome complexes in the splicing reaction[38,39]. 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. 3e 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. 3g).

Taken together, these data further demonstrate 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.

Recurrent aberrant splicing factor activity defines two cancer splicing programs

Cancer co-opts the splicing machinery in favor of tumor progression[4,9,40]. However, few studies have examined whether these splicing programs are unique to specific cancer cohorts or are shared, with most research focusing on single regulatory layers of splicing factors[41,42].

Our approach, based solely on exon inclusion changes, captures various regulatory mechanisms driving splicing factor activity. This suggests that estimated splicing factor activities reflect the combined output of all regulatory inputs affecting their function. To demonstrate its utility, we investigated whether splicing factor dysregulation in cancer follows a general pattern or varies by cohort. We analyzed multiple cancer cohorts to determine if certain splicing factors are consistently dysregulated, suggesting a recurrent cancer 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 assess splicing factor activity by comparing each PT sample to the median STN profiles and vice versa (Fig. 4a–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. 4b). Interestingly, some splicing factors emerged as recurrently activated or inactivated across a majority of cancer cohorts (Fig. 4c; Supplementary Fig. 4a). This suggests that oncogenic alterations may cause coordinated and recurrent dysregulation of splicing factor activity, despite their well-described heterogeneity. 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=61) putatively oncogenic splicing factors (i.e., aberrantly activated in ≥ 5 cohorts), and a set of (n=61) putatively tumor suppressor splicing factors (i.e., aberrantly inactivated in ≥ 5 cohorts) (Supplementary Table 3). Of note, PTBP1, a splicing factor with confirmed oncogenic roles in four different cancer types[43], emerged as the most frequently activated splicing factor.

If these recurrently dysregulated splicing factors play a 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. 4d; see Methods). For example, analysis of primary tumor samples across 33 TCGA cohorts revealed that aberrant activity of a recurrent oncogenic splicing factor (HNRNPD) correlates with poor prognosis in 26 of 33 cancer cohorts (Cox Proportional Hazards (PH) coefficient > 0) while the activity of a recurrent tumor suppressor splicing factor (DNMT1) correlates with favorable outcome in 27 of 33 cancer cohorts (Cox PH coefficient < 0) (Fig. 4e).

Extending the analysis to all 122 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 ($p = 1.7 \cdot 10^{-11}$, 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 ($p = 7.6 \cdot 10^{-12}$, two-sided Wilcoxon rank sum test) (Fig. 4f). These patterns are conserved if we consider different thresholds to define cancer splicing program recurrence (Supplementary Fig. 4b).

Typically, splicing factor activity is extrapolated from 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 necessarily 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 = -0.047, $p = 0.042$) (Supplementary Fig. 4c). Candidate oncogenic (n = 106) and tumor suppressor splicing factors (n = 30) identified by gene expression analysis (Supplementary Fig. 4d) were not predictive of poor and good outcome ($p > 0.05$ in both cases) (Supplementary Fig. 4e and 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.

To further characterize these cancer splicing programs, we examined their putative roles in the splicing reaction and in RBP families using SpliceosomeDB[44,45] and RBPDB[44], respectively. 32 and 54 splicing factors mapped to each database, respectively. Most were part of the C complex in the splicing reaction –involved in the catalytic steps–, and belonged to the RRMx1 and RRMx2 families of RBP domains (Fig. 4g and h).

Linking cancer splicing program dysregulation to 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. 5a). We performed an overrepresentation analysis (ORA) of the genes corresponding to the exons targeted by either oncogenic-like or tumor suppressor-like splicing factors.

To investigate the potential involvement of cancer splicing programs in cancer hallmarks, we performed an over-representation analysis (ORA) on genes corresponding to exons targeted by either oncogenic-like or tumor suppressor-like splicing factors. We selected the top 10 enriched terms (FDR < 0.05) with the most target genes for each program, highlighting both shared and unique terms, while ensuring that all significantly enriched terms from each program were included (Fig. 5b; Supplementary Table 4). Two of the enriched terms were related to antigen presentation, a critically regulated process in tumor immunology, and six others were related to cell cycle.

Immune evasion hampers the effectiveness of immune checkpoint blockade (ICB) therapies in cancer patients, complicating treatment response predictions[46]. A total of 181 genes belonging to the immune evasion terms "Antigen Processing Ubiquitination Proteasome Degradation" and "Class I MHC-mediated Antigen Processing Presentation" contain exons regulated by the tumor suppressor-like splicing program (Fig. 5b). To prioritize tumor suppressor-associated exons potentially involved in ICB efficacy, we analyzed exon inclusion differences between ICB-responding and non-responding patients based on RECIST criteria[47], using data from Riaz *et al.*[48]. Out of 608 exons from genes in immune-related terms, 16 were differentially spliced ($p < 0.05$, two-sided Wilcoxon rank sum tests). For additional prioritization, we used data from an in vivo genome-wide CRISPR screen comparing ICB response tumors in immunocompetent versus immuno-deficient NOD SCID Il2rg−/− control mice[48], identifying exons within genes affecting ICB response (Fig. 5c). Specifically, exon HsaEX1036341 in SEC22B –one of the strongest immune evasion inhibitors identified (impairment score = -1.2)– emerged as a promising candidate. Exclusion of this exon disrupts SEC22B's open reading frame, potentially reducing functional protein and improving ICB response, as evidenced by its lower inclusion in responder (PSI = 95.9, n = 10) than in non-responder (PSI = 99.4, n = 22) patients ($p = 0.024$, two-sided Wilcoxon rank sum tests). Survival analysis further supported this finding, with low exon inclusion in ICB-treated patients correlating with better prognosis ($p = 0.0079$, Log-rank test; best cutpoint = 98.91 PSI) (Fig. 5d). Examining empirical networks, we identified three splicing factors —PRPF8, SRSF1, and SF3B3– as regulators of HsaEX1036341. An ORA of their targets revealed enrichment (FDR < 0.05) in cell cycle control pathways and "Class I MHC-mediated Antigen Processing Presentation" (Supplementary Fig. 5a), suggesting

inactivating these splicing factors should impair immune responses as well. In vivo CRISPR data revealed silencing mouse homolog of SF3B3 closely mirrors SEC22B silencing (impairment score = -0.89) compared to silencing SRSF1 and PRPF8 (impairment scores = -0.61 and 0.68, respectively). This suggests SF3B3 as the most likely regulator of HsaEX1036341 in SEC22B. 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[49]. The activity of the oncogenic-like splicing program correlates positively with the proliferation marker (Pearson correlation = 0.2, $p < 6.9 \cdot 10^{-11}$) and the activity of the tumor suppressor-like splicing program correlates negatively (Pearson correlation = -0.32, $p < 2.2 \cdot 10^{-16}$) (Fig. 5e). Further, the sign of these relationships is reproducible when computing the correlation by grouping CCLE cell lines by cancer type of origin (Supplementary Fig. 5). 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.

Taken together, we confirmed the biological relevance of the cancer splicing programs identified by splicing factor analyses exploring their associations with two hallmarks of cancer: immune evasion and aberrant proliferation.

Cancer splicing programs exhibit coordinated regulation 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[50]. 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. 5a).

To test this, we analyzed how splicing factor activities change during carcinogenic transformation. Danielsson *et al.*[51] 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. 5f)[52]. We then estimated the splicing factor activities for the immortalized, tumorigenic, and metastatic cells using primary fibroblasts as the baseline reference (see Methods). While oncogenic-like splicing factors are less active than tumor suppressor-like splicing factors in immortalized cells, oncogenic programs activate while tumor suppressor programs inactivate in transformed and metastatic cells (Fig. 5g).

To assess whether gene expression drives these activity changes, we repeated the analysis using gene expression. We find that splicing factor expression generally increases with

tumorigenesis, independent of the oncogenic or suppressor splicing program (Supplementary Fig. 5c). This indicates that cancer splicing programs are activated coordinately during carcinogenesis through more complex regulation than solely changes in their gene expression. This result also underscores how our approach interprets exon inclusion signatures as a comprehensive measure of splicing factor activity beyond what gene expression analysis can capture.

Given the regulation of cancer splicing programs during carcinogenesis, we evaluated their potential role as carcinogenic drivers by examining DEMETER2 gene dependency scores from DepMap[53]. Dependency scores summarize gene silencing fitness effects and indicate the cancer driver potential of non-essential genes. Silencing essential genes hampers cell viability broadly, while perturbing non-essential tumor suppressors is expected to increase viability (dependency > 0), and silencing non-essential oncogenes should decrease it (dependency < 0)[54–56]. As expected, essential splicing factors have negative dependencies across all cell types and splicing programs (cancerous (n = 626), engineered (HEK, OELE, PRECLH), or fibroblasts (HS729)) (Supplementary Fig. 5d, top). Non-essential, tumor suppressor-like splicing factors exhibit higher dependencies than oncogenic-like ones in fibroblasts (p = 0.028, two-sided Wilcoxon rank sum tests) (Supplementary Fig. 5d, bottom). Cancerous and engineered cells show similar trends, but dependencies did not significantly differ between cancer programs (p > 0.05, two-sided Wilcoxon rank sum tests). Regulating splicing factors in cancer programs can influence cell proliferation –a cancer hallmark– suggesting that their coordinated regulation could contribute to cancer initiation and progression.

DISCUSSION

Repurposing VIPER with empirically derived splicing factor networks accurately estimated splicing factor activity solely from inclusion signatures of target exons. 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[16], 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 nonsynonymous somatic mutation— may affect splicing factor function. In contrast, splicing factor activity 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 empirical 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 splicing factor activity analysis from exon inclusion signatures 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[57]. In cancer cells, the splicing machinery is hijacked, resulting in gene isoforms that initiate and sustain the disease[4,8]. 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[58], RBFOX2[59], PTBP1[60]), the existence of general cancer-driver splicing programs remains underexplored[41]. As a case study, our systematic analysis of cancer samples identified two recurrent splicing factor programs with oncogenic- and tumor suppressor-like 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 and regulators potentially influencing ICB treatment efficacy. This not only corroborates the power of splicing factor activity analysis 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 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[11,12]– 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[61]. This suggests cell type abundance in bulk RNA sequencing samples could confound the splicing factor-exon associations inferred computationally, and sets the stage for future improvements. Nevertheless, we show splicing factor activities can be accurately inferred for 243 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 capability of our approach to capture altered splicing factor activity across diverse contexts using only transcriptomic data, we anticipate that splicing factor activity analysis will be instrumental in identifying and studying functionally relevant splicing programs underlying complex 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[8,20–23]– to create a consensus list of 509 splicing factors. Hegele *et al.*[23] 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.*[8] reviewed the regulation of alternative splicing. Head *et al.*[22] 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*[62] 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 -C*". 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 dataset, this workflow results in two gene expression tables corresponding to raw transcript counts and 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[63] 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 datasets for computational inference of splicing factor networks

In this study, we use different algorithms to estimate splicing factor activity from transcriptomic data (see below). These algorithms require a splicing factor network that includes both the likelihood (a score between 0 and 1) and the mode of regulation (positive or negative) of each splicing factor→exon interaction. While splicing factor networks can be derived empirically, they can also be inferred computationally from observational datasets. Here, we generated computationally inferred networks by linking splicing factor gene expression levels with exon inclusion profiles across a large sample set, which provides sufficient variation in splicing factor expression and exon inclusion.

To obtain diverse observational data, we used RNA sequencing datasets from two studies:

- Cardoso-Moreira *et al.*[64]: this dataset includes 313 human samples across 6 tissues (brain, heart, ovary, kidney, testis, and liver) at various stages of development and age.
- The Cancer Genome Atlas (TCGA): this dataset includes RNA sequencing data from 9,854 primary tumor samples and 736 adjacent solid tissue normal samples, offering a wide range of tissue types and cancer cohorts.

We processed these datasets using the *vast-tools* pipeline[62] to quantify gene expression TPM and exon inclusion PSI tables as described in previous sections. For each exon, *vast-tools* assigns missing values (NA) when it detects fewer than 10 reads, resulting in some missing data in exon inclusion profiles. Because computational network inference methods cannot operate with missing values, we performed missing data imputation using k-nearest neighbors (KNN) algorithm[65].

We applied the KNN imputation using the *sklearn.impute.KNNImputer* function with k=5 nearest neighbors, selected to balance data availability with imputation robustness. The imputation estimates the missing PSIs by calculating the Euclidean distance based on non-missing values across samples, filling in each missing PSI using the k closest samples as reference points. Using a moderate k=5 provides robust estimates without over-smoothing the inclusion levels (Supplementary Fig. 6). These imputed PSI values were used exclusively for the computational inference of splicing factor networks.

ARACNe

To infer splicing factor networks computationally, we adapted the ARACNe-AP tool[12]. Originally, ARACNe calculates pairwise mutual information (MI) scores between genes in a

single matrix. For our purposes, we modified ARACNe-AP to compute MI between two matrices with different row names –one for splicing factor log-normalized gene expression TPMs and another for imputed exon inclusion PSIs. Minimal updates were made to the ARACNe-AP code, and these changes are documented in the repository's version history (https://github.com/chaolinzhanglab/ARACNe-AP).

We ran ARACNe adapted to two different matrices on each of the observational datasets using default mutual information threshold of $1 \cdot 10^{-8}$ and 100 bootstrap iterations. For each bootstrap, we treated the splicing factor expression matrix as the regulator and the exon inclusion matrix as the target, applying data processing inequality specifically to splicing factors to limit their indirect associations. Each bootstrap run generated a network of splicing factor→exon interactions, where MI scores indicated the strength of each interaction.

To refine bootstrapped networks, we kept the top 1,000 exons per splicing factor, ranked by MI score. This selection showed minimal impact on performance across different numbers of top target exons retained (Supplementary Fig. 6). We then consolidated these bootstrap networks using ARACNe's module with the default p-value threshold of 0.05. This consolidation produced a consensus network containing significant splicing factor→exon interactions with MI values as interaction likelihoods.

To complete the splicing factor networks, we assigned a mode of regulation (values between -1 and 1) to each interaction, indicating the regulatory direction. Following VIPER's guidelines (https://www.bioconductor.org/packages/release/bioc/vignettes/viper/inst/doc/viper.pdf), we used the *viper::aracne2regulon* function to calculate these values based on the consolidated network's MI scores. This step created networks with both interaction likelihoods and mode of regulation (inhibition/activation), making them fully compatible with the methods tested for splicing factor activity estimation.

Multivariate linear regression (MLR)

To identify pairwise interactions between splicing factors and exons while controlling for other potential splicing factor influences, we applied MLR to each observational dataset. For each exon, we modeled its PSI as the dependent variable, with the standardized, log-normalized TPMs of 509 splicing factors as independent variables. This setup allowed us to determine the contribution of each splicing factor to exon inclusion, producing a regression coefficient and associated p-value for each splicing factor→exon pair.

In these models, the regression coefficient represents the strength of the putative interaction, while the p-value assesses the significance of the interaction. To ensure compatibility with methods used for splicing factor activity estimation, we included only interactions with a nominal p-value below 0.01. For each significant interaction, we used the absolute value of the coefficient as the interaction likelihood and the sign of the coefficient as the mode of regulation (-1 or 1).

This approach produced splicing factor networks with both likelihood and mode of regulation for each splicing factor→exon interaction.

Empirical

We constructed empirical splicing factor networks using experimental perturbation data from publicly available sources collected to generate our benchmark datasets. For each perturbation experiment (e.g., splicing factor knockdown, knockout, or over expression), we calculated a delta PSI by subtracting the exon inclusion in the control condition from the perturbed condition. To maintain consistency across perturbation types, we multiplied the delta PSI values by -1 for overexpression experiments, assuming that exons more included upon splicing factor upregulation would be excluded upon downregulation in the same cell system. This approach only affects the sign of the delta PSI values, not the original PSI values, which remain within the 0-100% range.

Next, we selected exons showing substantial inclusion changes (|delta PSI| ≥ threshold) upon splicing factor perturbation. The absolute delta PSI served as the interaction likelihood, while the delta PSI sign indicated the interaction's mode of regulation. Across thresholds from 5% to 45% (in 5% increments), we assessed the resulting networks' accuracy by calculating ROC-AUC evaluation metrics. A 15% threshold achieved optimal performance and practical network sizes, encompassing 243 of the 248 splicing factors with available experimental data, with high ROC-AUC scores for VIPER-based activity estimation (Supplementary Fig. 1c).

For splicing factors with multiple perturbation experiments available, we aggregated individual splicing factor networks at the dataset level, preserving independent biological replicates. In the ENA dataset, with multiple independent exon inclusion signatures for some splicing factors, we grouped networks by sets of studies to reflect different experiments.

To ensure comparability with the ENCODE-derived networks –which is separated by cell line (HepG2 or K562) and perturbation type (CRISPR knockdown or shRNA knockdown)– we also categorized ENA-derived networks into four distinct groups of experiments perturbing different splicing factors. This resulted in a total of eight networks:

I. ENCORE knockdowns in HepG2 (41,477 splicing factor-exon interactions, 104 splicing factors);
II. ENCORE knockdowns in K562 (45,396 splicing factor-exon interactions, 104 splicing factors);
III. ENCORE knockouts in HepG2 (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 group 1 (14,158 splicing factor-exon interactions, 108 splicing factors, 48 studies);
VI. ENA perturbations in different cell systems group 2 (3,042 splicing factor-exon interactions, 23 splicing factors, 9 studies);
VII. ENA perturbations in different cell systems group 3 (522 splicing factor-exon interactions, 12 splicing factors, 3 studies); and
VIII. ENA perturbations in different cell systems group 4 (963 splicing factor-exon interactions, 2 splicing factors, 2 studies)

Cross-linking and immunoprecipitation (CLIP)

To construct CLIP-based splicing factor networks, we obtained CLIP peaks from the POSTAR3 database[27]. We associated each splicing factor with exons by mapping CLIP peaks to the nearest exon in VastDB within a 500 bp window. In this network, we assigned both the likelihood and mode of regulation to 1, as CLIP data only confirms the presence of an interaction but does not indicate its interaction probability or regulation direction.

ARACNe likelihood and Empirical mode of regulation

To assess whether the likelihood values generated by ARACNe contribute to the low performance in splicing factor activity estimation, we created splicing factor networks by replacing empirical likelihoods with those calculated by ARACNe. We included only splicing factor→exon interactions present in both the empirical and ARACNe-derived networks, ensuring a shared interaction basis for comparison.

MLR likelihood and Empirical mode of regulation

Using a similar approach to that applied to ARACNe-derived networks, we substituted the empirical likelihoods with likelihoods derived from MLR. As with ARACNe, we retained only splicing factor→exon interactions common in both empirical and MLR networks to allow a direct comparison.

Estimation of splicing factor activities using splicing factor networks

We estimated splicing factor activities using four complementary methods: normalized enrichment scores (NES) from gene set enrichment analysis (GSEA), Spearman correlation, Pearson correlation, and VIPER. Each method uses different principles for estimating activity, allowing us to benchmark different approaches for splicing factor activity estimation from a splicing factor network and an exon inclusion signature.

NES from GSEA

Using GSEA, we estimated splicing factor activities as NES values. Here, each splicing factor network was used as an ontology where splicing factors represent exon sets and exon signatures are ranked by their inclusion levels. Missing PSIs were omitted. We applied the *clusterProfiler::GSEA*[66] function with the default *fgsea* algorithm[67]. Likelihood and mode of regulation data could not be included in this algorithm. The GSEA function was set with *pvalueCutoff=1.1* to ensure NES scores were produced for all splicing factors, using default settings for other parameters.

Correlation-based activity estimation (Spearman and Pearson correlation)

To capture direct relationships between splicing factors and exon inclusion, we estimated splicing factor activities as correlation coefficients. For Spearman and Pearson correlations, we used *stats::cor*, specifying “spearman” and “pearson” methods, respectively. To incorporate the interaction likelihoods and modes of regulation, each splicing factor network was preprocessed by multiplying these values, producing a refined interaction metric. Correlation of this metric with exon inclusion levels provided activity estimates for each splicing factor.

VIPER

VIPER predicts splicing factor activity by assessing the enrichment of target exons in splicing signatures. It requires two inputs: a molecular signature, such as differences in exon inclusion between perturbed and unperturbed conditions, and a splicing factor network that specifies each factor's target exons, along with a likelihood score and a mode of regulation.

To calculate protein activity, VIPER applies the analytic Rank-based Enrichment Analysis (aREA)[16], testing the enrichment of a regulator's target exons in the input molecular signature. If any exon inclusion values are missing, VIPER omits these values during the enrichment analysis.

VIPER evaluates the significance of activity predictions by comparing enrichment scores to a null model generated through samples permutation. When sample size is limited, VIPER uses permutation or an analytic approximation to assess enrichment. It provides activity scores for all network regulators as output.

Evaluation metrics of splicing factor activity estimation

To evaluate the accuracy of different methods for estimating splicing factor activity, we constructed benchmark datasets from publicly available single-splicing factor perturbation experiments. These datasets included all relevant experiments from ENCODE and ENA that perturb a single splicing factor, each producing an exon inclusion signature tied to a specific splicing factor's alteration. We divided these experiments into five benchmark datasets based on source ENCODE-KD-HepG2, ENCODE-KD-K562, ENCODE-KO-HepG2, ENCODE-KO-K562, and ENA.

The goal is for a reliable network and estimation method to accurately identify which splicing factor was perturbed in each experiment. For example, given an exon inclusion signature from a knockdown experiment, the best-performing approach should correctly assign the lowest activity score to the knocked-down splicing factor.

To evaluate performance, we calculated area under the receiver operator curve (ROC-AUC) scores using the *pROC::roc* function[68], comparing performance across two perspectives: by splicing factor and by perturbation (Fig. 2b). First, we estimated splicing factor activities for each perturbation signature, resulting in a table where rows represent splicing factors and columns represent perturbation experiments.

To calculate ROC-AUC scores across splicing factors, we used each row in this table as a predictor (the activity scores) and a binary indicator marking whether the splicing factor was perturbed as the response. This tests how well the method identifies a perturbation for each splicing factor across all experiments. Similarly, we calculated ROC-AUC scores across perturbations by using each column as a predictor and a binary vector marking the perturbed factor for that experiment as the response. This dual approach helps evaluate how well the method detects perturbations consistently across both splicing factors and individual perturbation experiments.

Unless stated otherwise, all performance metrics were computed using held-out benchmark datasets.

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 degradation of RBM39 splicing factor protein

We downloaded the proteomic and transcriptomic data produced by Nijhuis *et al.*[28] upon perturbation of IMR-32 neuroblastoma 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.*[28] dataset had a single biological replicate for transcriptomic data condition (see Supplementary Fig. 2a), we obtained RNA sequencing data from Lu *et al.*[29], 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 present measured transcriptomic changes upon combinatorial knockdown of splicing factors[30–33], 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[21,69–73]. 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 as 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 identify cancer splicing programs based on gene expression, we used DESeq2[74] to perform differential expression analysis. Starting with raw gene count data, we compared primary tumor samples to normal solid tissue samples for each of the 14 cancer cohorts included in our study.

For each splicing factor, we assessed whether its expression was significantly different (FDR < 0.05). We then counted the number of cancer cohorts where each splicing factor as consistently upregulated or downregulated. To define "oncogenic-like" and "tumor suppressor-like" programs, we required a splicing factor to be consistently upregulated in at least 5 cancer types to be classified as "oncogenic-like". Similarly, splicing factors downregulated in at least 5 cancer cohorts were classified as tumor suppressor-like.

Association sign 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)

To assess gene set enrichment, we used the *clusterProfier::enricher* function, which performs statistical tests to identify significant overlaps between our gene lists and predefined pathways. We used ReactomeDB as our reference database.

For background (or "universe") genes, we kept the default setting, which uses all genes annotated in ReactomeDB.

Identification of exons in splicing programs associated with immune evasion

Analysis of melanoma patients treated with immune checkpoint blockade drugs

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.*[48] study. We then classified each patient into a responder or non-responder considering their RECIST[47] 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.*[75] 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 that 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.*[51] 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

Intermediate files generated from data analyses

Intermediate files generated throughout this study are available in this Figshare repository: https://doi.org/10.6084/m9.figshare.27835518.v1

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.[23]
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.[8]
Supplementary Table 1 in the publication.

List of splicing factors from Head et al.[22]
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.

POSTAR3 CLIP data to construct CLIP-based splicing factor networks
http://postar.ncrnalab.org/

Observational RNA sequencing data from human samples from diverse molecular contexts
We downloaded raw FASTQ files for Cardoso-Moreira *et al.*[64] through the ENA website.

Perturbing RBM39 with Indisulam
For the Nijhuis *et al.*[28] 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.*[29] 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[33], PRJNA498529[32], PRJNA587741[31], and PRJNA321560[30].

Protein-protein interaction network
We downloaded STRING DB's human protein interaction network from their webpage[76].
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[35], PRJNA685790[69], PRJNA662572[70], PRJNA380104[71], PRJNA354957[72], and PRJNA292827[73]

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.

SpliceosomeDB complexes
We obtained annotations of splicing factors to splicing reaction complexes from the SpliceosomeDB website.
http://spliceosomedb.ucsc.edu/

RBP domain families
We obtained annotations of splicing factors to RBP domain families from the RBPDB website.
http://rbpdb.ccbr.utoronto.ca/

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.[75]
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/MediaObje

cts/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.[48]
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.[51]
We downloaded raw FASTQ files from the ENA website (PRJNA193487).

DEMETER2 gene dependencies from DepMap
https://ndownloader.figshare.com/articles/6025238/versions/6

CCLE cell line metadata
https://ndownloader.figshare.com/articles/13681534/versions/1

CODE AVAILABILITY

To improve the reproducibility of our results, the full data analysis pipeline of this manuscript will be made 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 at https://github.com/MiqG/viper_splicing.

FIGURE LEGENDS

Figure 1. Splicing factor activity is determined by multiple levels of regulation.

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.

Figure 2. Reverse engineering of splicing factor networks to estimate splicing factor activities with VIPER.

(**a**) Estimation of splicing factor activity from exon inclusion signatures. 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. We

estimated splicing factor activity using normalized enrichment scores (NES) from gene set enrichment analysis (GSEA), Pearson and Spearman correlation coefficients, and VIPER.

(**b**) Benchmark outline for splicing factor activity estimation algorithms. We generated 5 benchmark datasets including single-splicing factor perturbation experiments from different publicly available sources. To compute performance metrics (ROC-AUC), we consider predicted splicing factor activities and which splicing factor was actually perturbed across perturbations (horizontal) and across splicing factors (vertical).

(**c**) Evaluation of splicing factor activity estimation with different methods for activity estimation from an exon inclusion signature and a splicing factor→exon network. Computational networks (ARACNe, MLR) were reverse-engineered using three different datasets of increasing size.

(**d**) Distribution of number splicing factors per target exon (left), and number of target exons per splicing factor (right) in empirical splicing factor networks.

(**e**) Distributions of the number of targets (exons) per regulator (splicing factor) considering either all splicing factors categorized by "Core" splicing factors appear in Spliceosome Database[45], "RBP" splicing factors that appear in Gene Ontology[77] category "RNA binding protein", or "Other" splicing factors.

(**f**) Evaluation of splicing factor activity estimation for each splicing factor in benchmark datasets using empirical splicing factor networks.

In box and whisker plots in panels (**c**), (**d**), (**e**), and (**f**), the median is marked by a horizontal line, with the first and third quartiles as box edges. Whiskers extend up to 1.5 times the interquartile range, and individual outliers are plotted beyond.

Figure 3. 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 and p-values obtained from a two-sided T-test in IMR-32 cells. Red, RBM39.
(**c**) Ranking of splicing factors according to their estimated protein activities upon treatment with Indisulam (5 μM, 6 hours) in IMR-32 cells. Red, RBM39.

(**d**) Distributions of estimated protein activities upon combinatorial perturbation of splicing factors. Red, corresponding knocked down genes in each case. Top, experimental outline. Studies analyzed include PRJNA223244[33], PRJNA498529[32], PRJNA587741[31], and PRJNA321560[30].

(**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.*[8]

(**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 upon

perturbation with different splicing-targeting drugs considering the shortest path lengths of each splicing factor to the SF3b complex. Color, whether genes belong to the U2snRNP. Studies analyzed include PRJNA371421[35], PRJNA685790[69], PRJNA662572[70], PRJNA380104[71], PRJNA354957[72], and PRJNA292827[73]. In box and whiskers plots in the panel, the median is marked by a horizontal line, with the first and third quartiles as box edges. Whiskers extend up to 1.5 times the interquartile range, and individual outliers are plotted beyond.

Figure 4. Splicing factors recurrently active in tumors exhibit cancer driver-like behavior.

(**a**) TCGA cancer cohorts overview. Among the 33 cohorts with tumor samples, 14 also have corresponding healthy (solid tissue normal) samples. Cancer type abbreviations are available 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**) Frequency of splicing factors showing significant activation (top) or inactivation (bottom) in tumor samples versus solid tissue normal (FDR < 0.05, two-sided Wilcoxon rank sum tests). Dashed lines mark the recurrence threshold for a splicing factor's inclusion in either the oncogenic-like or tumor suppressor-like cancer splicing program, with selected splicing factors above each line. Labels, top 5 most recurrently activated or inactivated splicing factors.

(**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 in which either HNRNPD (left) or DNMT1 (right) are associated with either high or low risk of death. Color, cancer splicing program. Top, number of cancer cohorts for each type of association.

(**f**) Distributions of the number of cancer cohorts that each splicing factor is associated with a high or low risk of death. Color, cancer splicing program. Top, p-value based on two-sided Wilcoxon rank sum tests. Bottom, number of splicing factors in the corresponding cancer splicing program.

(**g** and **h**) Splicing factors from cancer splicing programs identified found in (**g**) the spliceosome reaction complexes according to SpliceosomeDB[45], and in (**h**) RNA-binding protein (RBP) domain families.

In box and whiskers plots in panel (**f**), the median is marked by a horizontal line, with the first and third quartiles as box edges. Whiskers extend up to 1.5 times the interquartile range, and individual outliers are plotted beyond.

Figure 5. Contribution of cancer splicing programs to disease progression.

(**a**) Cancer splicing program activation during carcinogenesis. Carcinogenic transformation begins with genomic lesions that activate oncogenes and inactivate tumor suppressors, reprogramming the cell into a cancerous state. We hypothesize that cancer splicing programs are regulated alongside these transformations.

(**b**) ORA of the top 10 significantly (FDR < 0.05) enriched ReactomeDB terms for genes corresponding to target exons of splicing factors in the oncogenic-like and tumor suppressor-like cancer splicing programs. Enriched terms only in the top 10 for one of the splicing programs are also shown.

(**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), we considered the orthologous gene-level immune evasion score 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.*[48] 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 with cancer cell proliferative state. Top, Pearson correlations (R) and corresponding p-values (p) between the variables in the axes. Error shadows, 95% confidence intervals.

(**f**) Danielsson *et al.*[51] experimental design. Primary fibroblasts were induced into different states transfecting either hTERT (immortalized), hTERT and SV40 large-T (tumorigenic), hTERT, SV40 large-T, and RAS G12V (metastasizing).

(**g**) Splicing factor activities by cancer splicing program in different stages of carcinogenesis. Top, p-value 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 is marked by a horizontal line, with the first and third quartiles as box edges. Whiskers extend up to 1.5 times the interquartile range, and individual outliers are plotted beyond.

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**) Evaluation metrics of empirical splicing factor networks defined with different thresholds for absolute delta PSI.

(**d**) Evaluation metrics of empirical splicing factor networks defined with four of the five benchmark datasets and evaluated on the fifth one.

(**e**) Robustness analysis of the selected empirical splicing factor network (threshold delta PSI = |delta PSI|) consisting of keeping only top K interactions for each splicing factor sorted by splicing factor→exon interaction likelihood, or delta PSI magnitude.

In box and whisker plots in panels (**c**), (**d**), and (**e**), the median is marked by a horizontal line, with the first and third quartiles as box edges. Whiskers extend up to 1.5 times the interquartile range, and individual outliers are plotted beyond.

Supplementary Figure 2. Evaluation of methods to reverse engineer splicing factor networks.

(**a**) Distributions of distances between mapped CLIP peaks and target exons to generate CLIP-based splicing factor networks.

(**b**) Venn diagrams of the overlaps between the splicing factors of the empirical splicing factor networks and their target interactions.

(**c**) Evaluation of CLIP-based splicing factor networks compared to empirical networks. “Empirical in CLIP” are networks made of the overlapping interactions between perturbation-based networks and CLIP-based networks; direct empirical interactions. “Empirical not in CLIP” are empirical networks without CLIP-based interaction evidence; indirect empirical interactions. Note that correlation-based activities cannot be computed for CLIP-based networks as all likelihood and mode of regulation values are constant.

(**d**) Likelihood scores from ARACNe and MLR match empirical interactions. Distributions of splicing factor→exon likelihood scores in ARACNe- or MLR-derived networks for experimentally observed vs. unobserved interactions. Top, two-sided Wilcoxon rank sum p-value. Bottom, interaction counts per category.

(**e**) MoR inferred by ARACNe and MLR approximate random (0.5). Proportion of correctly inferred positive and negative modes of regulation (MoR) interactions by ARANCe and MLR algorithms compared to experimental data.

(**f**) Combining ARACNe or MLR likelihoods with empirical MoR restores predictive power. Across perturbations and across splicing factors, ROC-AUCs for network versions using computational likelihoods with empirical MoR signs.

In box and whisker plots in panels (**c**), (**d**), and (**f**), the median is marked by a horizontal line, with the first and third quartiles as box edges. Whiskers extend up to 1.5 times the interquartile range, and individual outliers are plotted beyond.

Supplementary Figure 3. Changes in transcriptomic gene expression upon treatment with Indisulam.

(**a**) mRNA levels of RBM39 treated with either Indisulam (5 μM, 6 hours) or DMSO vehicle control in the IMR-32 cell line from study Nijhuis *et al.*[28]. Top, corresponding gene expression values. Note this corresponds to a single biological replicate.

(**b**) mRNA levels of RBM39 of A375, MEL501, and SKMEL239 cell lines treated with either Indisulam (1 μM, 96 hours) or DMSO vehicle control from study Lu *et al.*[29]. Top, p-value based on two-sided Wilcoxon rank sum tests comparing the mRNA levels of cell lines in each condition.

In box and whiskers plots in panel (**b**), the median is marked by a horizontal line, with the first and third quartiles as box edges. Whiskers extend up to 1.5 times the interquartile range, and individual outliers are plotted beyond.

Supplementary Figure 4. Splicing programs defined by gene expression lack cancer driver-like behavior.

(**a**) Differential activation status of splicing factors (FDR < 0.05, two-sided Wilcoxon rank sum test) across cancer cohorts.
(**b**) Distributions of cancer cohorts in which each splicing factor is associated with high or low death risk, based on thresholds for defining recurrent activation in cancer splicing programs. Top, p-value based on two-sided Wilcoxon rank sum tests. Bottom, splicing factor count in resulting each program.

(**c**) Relationship between differential splicing factor protein activities and splicing factor gene expression fold-changes 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 (top) and downregulated (bottom) splicing factors. 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 that each splicing factor expression is associated with a high or low risk of death in gene expression-based cancer splicing programs defined with recurrently differentially expressed splicing factors. Color, cancer splicing program. Top, p-value 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 that each splicing factor is associated with a high or low risk of death considering different thresholds of differential expression recurrence to define cancer splicing programs. Color, cancer splicing program. Top, p-value based on two-sided Wilcoxon rank sum tests. Bottom, number of splicing factors in the corresponding cancer splicing program.

In box and whisker plots in panels (**b**), (**e**), and (**f**), the median is marked by a horizontal line, with the first and third quartiles as box edges. Whiskers extend up to 1.5 times the interquartile range, and individual outliers are plotted beyond.

Supplementary Figure 5. Functional insights and dependencies of cancer splicing programs in carcinogenesis and immune response.

(**a**) The 10 most significantly enriched ReactomeDB terms (FDR < 0.05, ORA), ranked by gene set size, for the target genes of splicing factors PRPF8, SRSF1, and SF3B3. These factors are putative regulators of exon HsaEX1036341 in SEC22B.

(**b**) Distributions of Pearson correlations between proliferative state marker MKI67 and median splicing factor activities of each cancer splicing program. We considered the 22 cancer types with more than 10 cell lines in the database. Top, *p*-value based on two-sided Wilcoxon rank sum test.

(**c**) Distributions of log2 fold-change in gene expression for splicing factors in cancer splicing programs, comparing untreated BJ fibroblasts and fibroblasts at different carcinogenesis stages. Top, p-value based on Wilcoxon rank sum tests.

(**d**) DEMETER2[53] gene dependency scores for splicing factors in each cancer splicing program, stratified by cell type and essentiality status of the splicing factor. Top, p-value based on Wilcoxon rank sum tests.

In box and whisker plots in panels (**b**), (**c**), and (**d**), the median is marked by a horizontal line, with the first and third quartiles as box edges. Whiskers extend up to 1.5 times the interquartile range, and individual outliers are plotted beyond.

Supplementary Figure 6. Parameter sensitivity analysis for PSI imputation and network inference with ARACNe

(**a**) Evaluation of KNNImputer for PSI imputation with different k values. Distributions of real and imputed PSI values for 5 randomly selected samples from the CardosoMoreira2020 exon inclusion PSI table. “Synthetically Missing” PSI values were artificially set to NA for imputation evaluation. “Real” PSI values, left unaltered, are shown for comparison.

(**b**) Impact of “max_targets” on ARACNe inference. ROC AUC distributions for computational splicing factor networks inferred with ARACNe using different “max_targets” parameters. Networks were inferred using the CardosoMoreira2020 dataset. In box and whisker plots, the median is marked by a horizontal line, with the first and third quartiles as box edges. Whiskers extend up to 1.5 times the interquartile range, and individual outliers are plotted beyond.

FIGURES

Figure 1

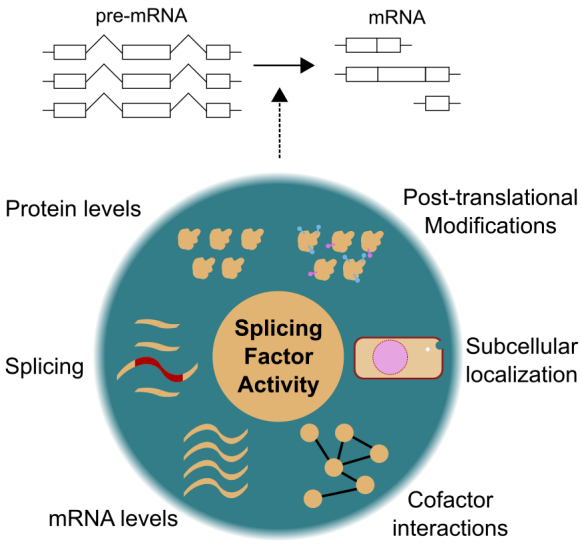


Figure 2

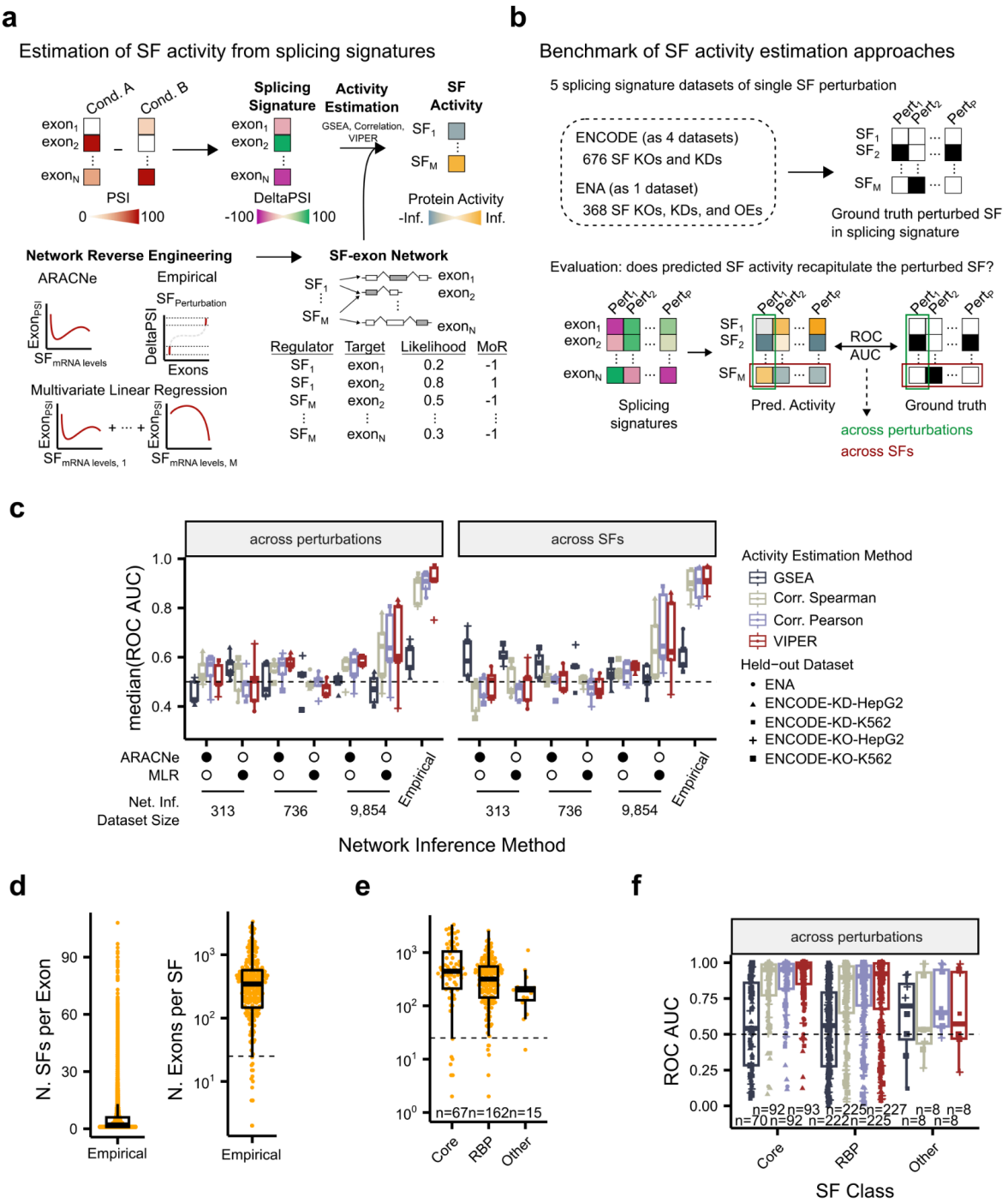

a
Estimation of SF activity from splicing signatures
Cond. A
Cond. B
exon1
exon2
exonN
PSI
0
100
Splicing Signature
DeltaPSI
-100
100
Activity Estimation
GSEA, Correlation, VIPER
SF Activity
SF1
SFM
Protein Activity
-Inf.
Inf.
Network Reverse Engineering
ARACNe
Empirical
SFPerturbation
ExonPSI
SFmRNA levels
DeltaPSI
Exons
Multivariate Linear Regression
SFmRNA levels, 1
SFmRNA levels, M
SF-exon Network
Regulator Target Likelihood MoR
SF1 exon1 0.2 -1
SF1 exon2 0.8 1
SFM exon2 0.5 -1
SFM exonN 0.3 -1
b
Benchmark of SF activity estimation approaches
5 splicing signature datasets of single SF perturbation
ENCODE (as 4 datasets)
676 SF KOs and KDs
ENA (as 1 dataset)
368 SF KOs, KDs, and OEs
Pert1
Pert2
PertP
SF1
SF2
SFM
Ground truth perturbed SF in splicing signature
Evaluation: does predicted SF activity recapitulate the perturbed SF?
Splicing signatures
Pred. Activity
ROC AUC
Ground truth
across perturbations
across SFs
c
across perturbations
across SFs
median(ROC AUC)
1.0
0.8
0.6
0.4
ARACNe
MLR
Net. Inf.
Dataset Size
313
736
9,854
Empirical
Network Inference Method
Activity Estimation Method
GSEA
Corr. Spearman
Corr. Pearson
VIPER
Held−out Dataset
ENA
ENCODE-KD-HepG2
ENCODE-KD-K562
ENCODE-KO-HepG2
ENCODE-KO-K562
d
N. SFs per Exon
90
60
30
0
Empirical
N. Exons per SF
10^3
10^2
10^1
Empirical
e
10^3
10^2
10^1
10^0
n=67n=162n=15
Core
RBP
Other
f
across perturbations
ROC AUC
1.00
0.75
0.50
0.25
0.00
n=92 n=93 n=225n=227 n=8 n=8
n=70 n=92 n=222n=225 n=8 n=8
Core
RBP
Other
SF Class

Figure 3

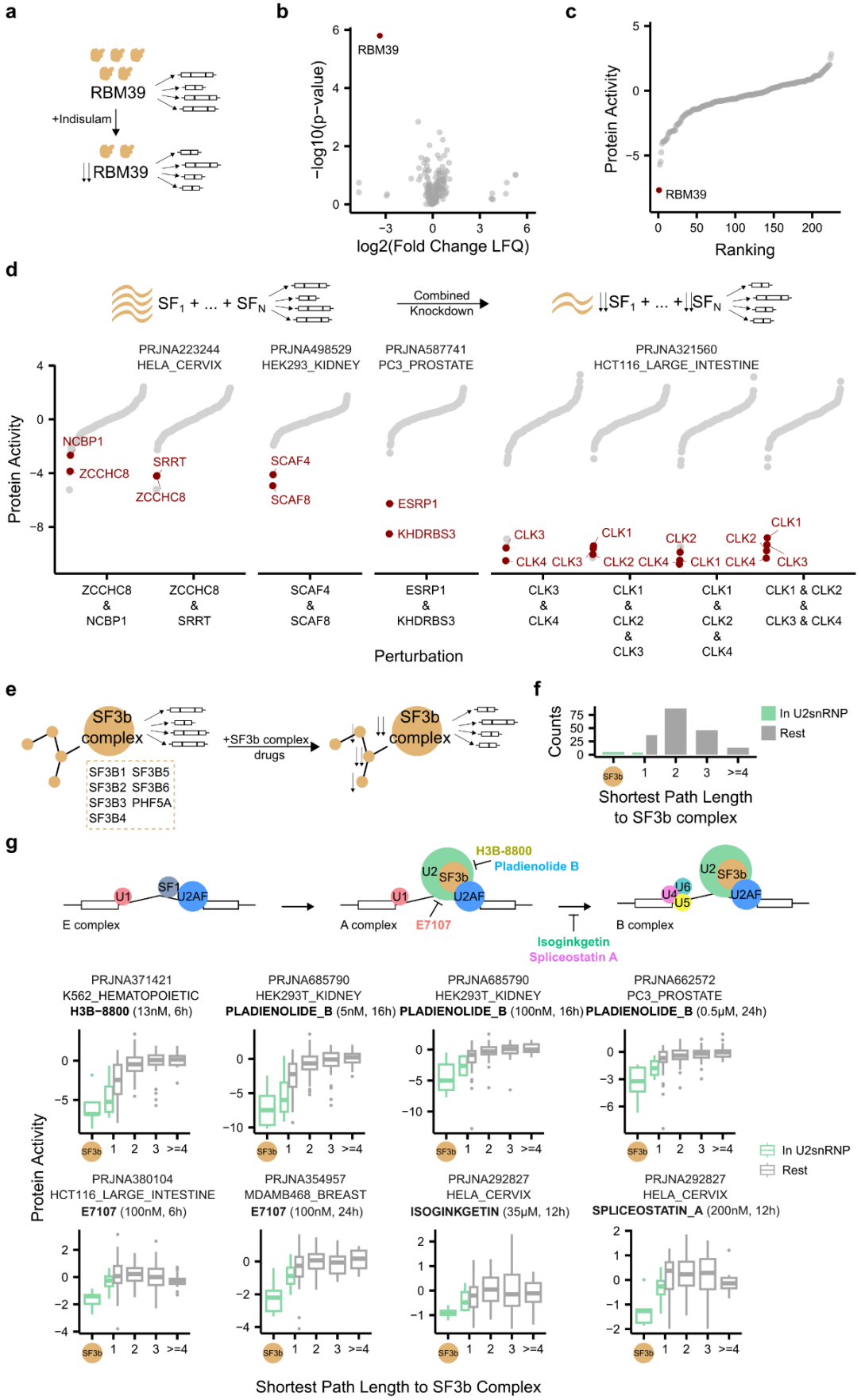

a
RBM39
+Indisulam
RBM39
b
RBM39
−log10(p−value)
log2(Fold Change LFQ)
c
Protein Activity
RBM39
Ranking
d
SF1 + ... + SFN
Combined Knockdown
PRJNA223244 HELA_CERVIX
PRJNA498529 HEK293_KIDNEY
PRJNA587741 PC3_PROSTATE
PRJNA321560 HCT116_LARGE_INTESTINE
Protein Activity
NCBP1
ZCCHC8
SRRT
ZCCHC8
SCAF4
SCAF8
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 4

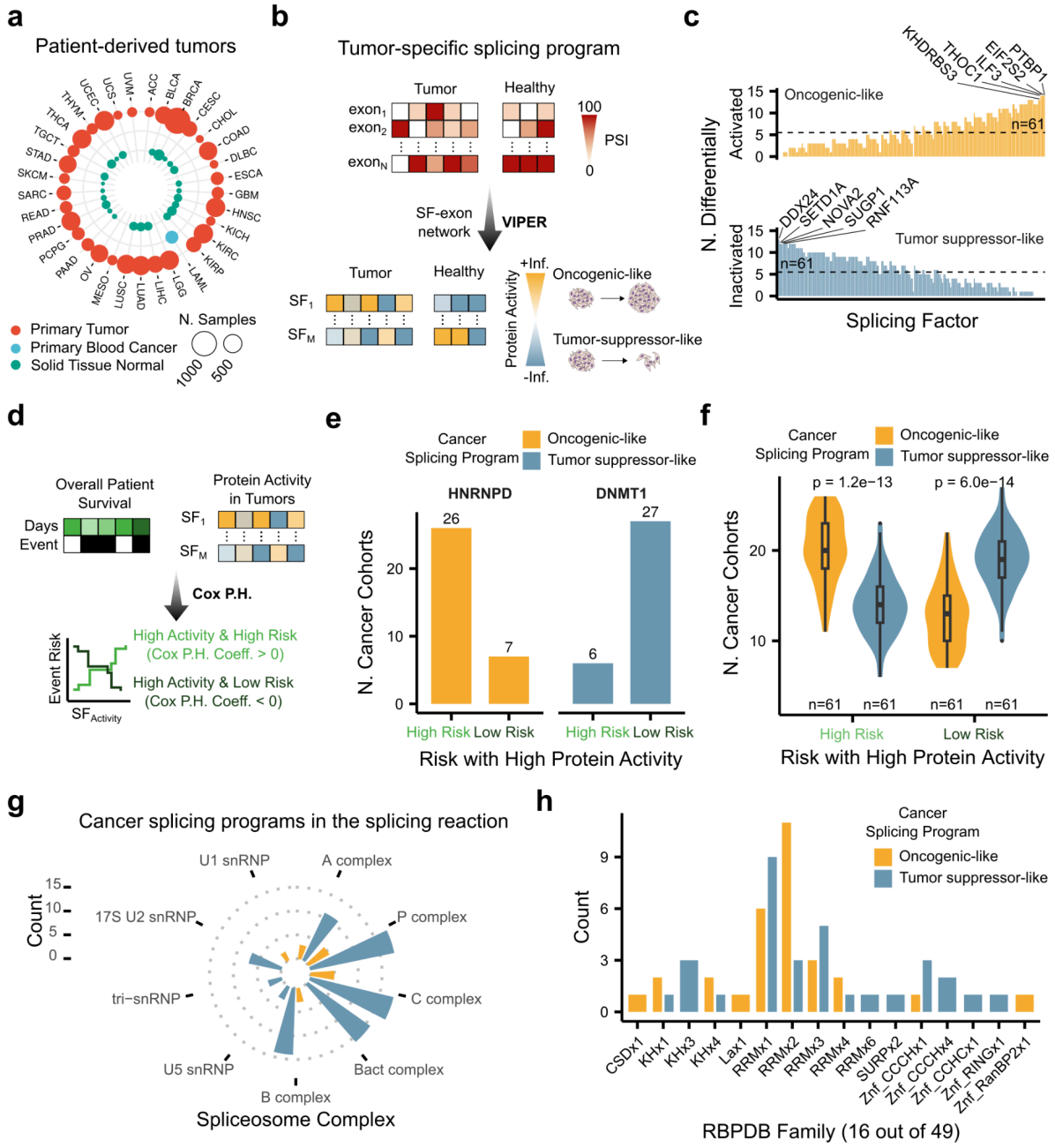
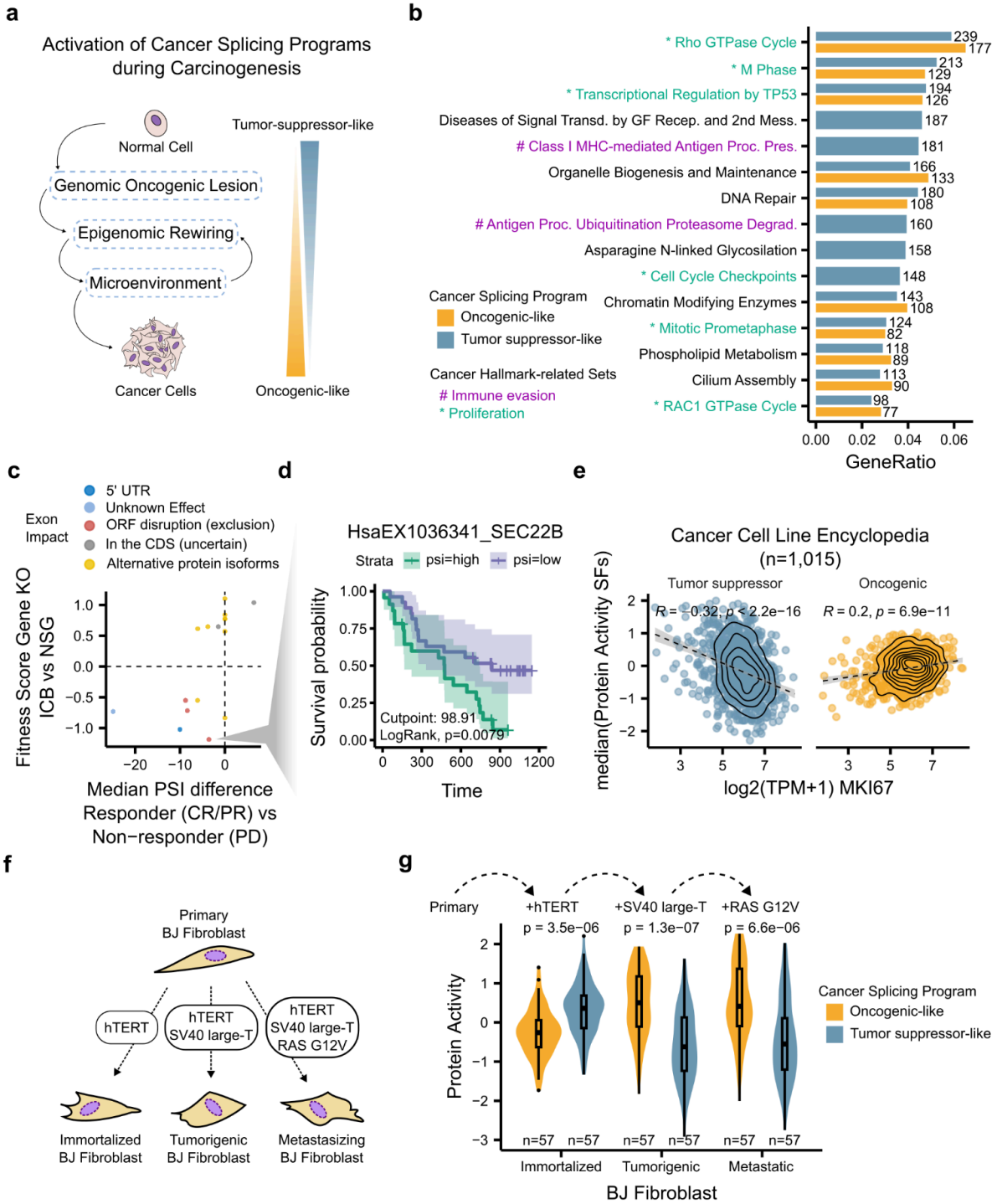
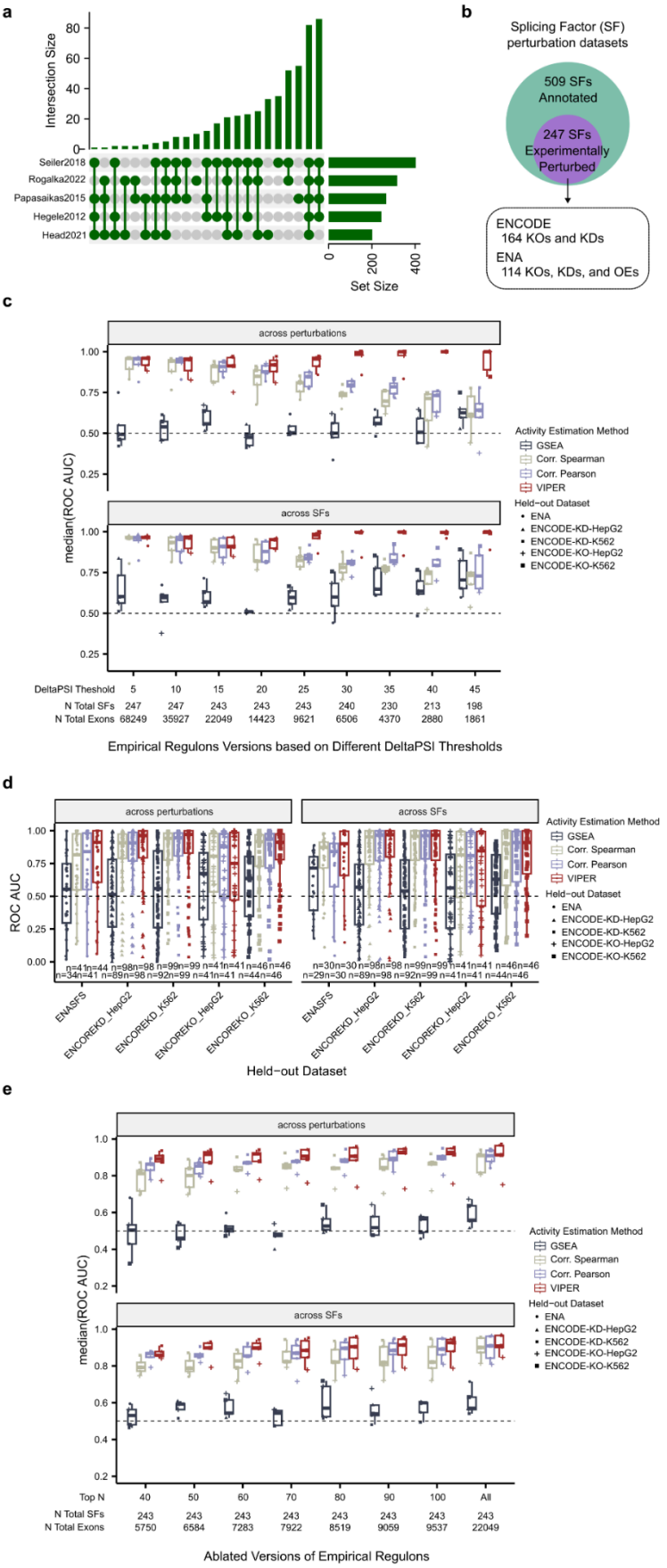

a
Patient-derived tumors
Primary Tumor
Primary Blood Cancer
Solid Tissue Normal
N. Samples
1000
500
b
Tumor-specific splicing program
Tumor
Healthy
PSI
SF-exon network
VIPER
Protein Activity
Oncogenic-like
Tumor-suppressor-like
c
N. Differentially
Activated
Inactivated
Oncogenic-like
Tumor suppressor-like
n=61
Splicing Factor
d
Overall Patient Survival
Protein Activity in Tumors
Days
Event
Cox P.H.
Event Risk
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
DNMT1
N. Cancer Cohorts
High Risk
Low Risk
Risk with High Protein Activity
f
p = 1.2e−13
p = 6.0e−14
n=61
Risk with High Protein Activity
g
Cancer splicing programs in the splicing reaction
Count
U1 snRNP
A complex
17S U2 snRNP
P complex
tri−snRNP
C complex
U5 snRNP
Bact complex
B complex
Spliceosome Complex
h
Count
RBPDB Family (16 out of 49)

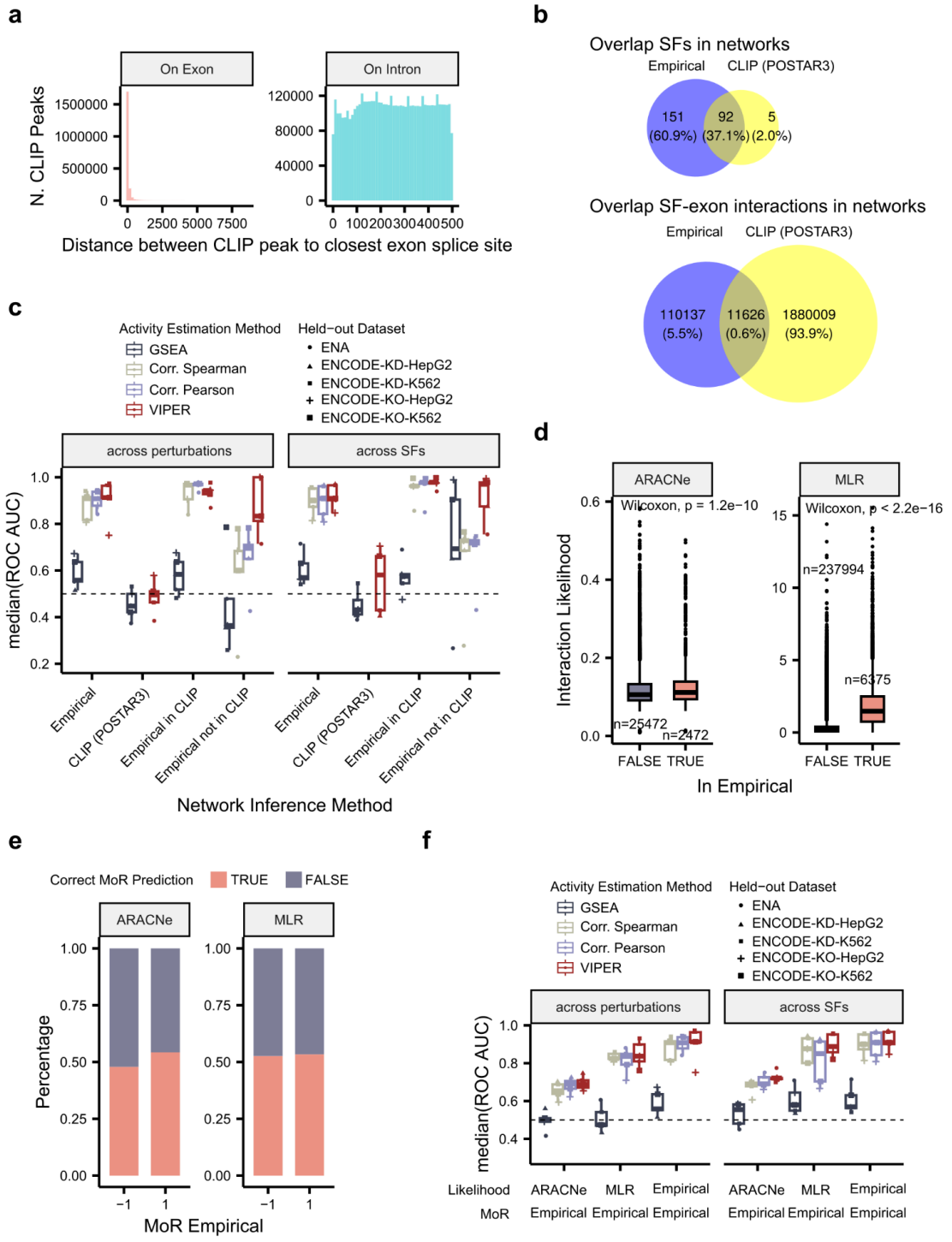
Figure 5


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
* RAC1 GTPase Cycle
239 177 213 129 194 126 187 181 166 133 180 108 160 158 148 143 108 124 82 118 89 113 90 98 77
Cancer Splicing Program
Oncogenic-like
Tumor suppressor-like
Cancer Hallmark-related Sets
Immune evasion
* Proliferation
0.00 0.02 0.04 0.06
GeneRatio
c
5' UTR
Unknown Effect
ORF disruption (exclusion)
In the CDS (uncertain)
Alternative protein isoforms
Exon Impact
Fitness Score Gene KO ICB vs NSG
1.0 0.5 0.0 −0.5 −1.0
−20 −10 0
Median PSI difference Responder (CR/PR) vs Non−responder (PD)
d
HsaEX1036341_SEC22B
Strata psi=high psi=low
Survival probability
1.00 0.75 0.50 0.25 0.00
Cutpoint: 98.91
LogRank, p=0.0079
0 300 600 900 1200
Time
e
Cancer Cell Line Encyclopedia (n=1,015)
Tumor suppressor
Oncogenic
R = −0.32, p < 2.2e−16
R = 0.2, p = 6.9e−11
median(Protein Activity SFs)
2 1 0 −1 −2
3 5 7
log2(TPM+1) MKI67
f
Primary BJ Fibroblast
hTERT
hTERT SV40 large-T
hTERT SV40 large-T RAS G12V
Immortalized BJ Fibroblast
Tumorigenic BJ Fibroblast
Metastasizing BJ Fibroblast
g
Primary
+hTERT p = 3.5e−06
+SV40 large-T p = 1.3e−07
+RAS G12V p = 6.6e−06
Protein Activity
2 1 0 −1 −2 −3
n=57 n=57 n=57 n=57 n=57 n=57
Immortalized
Tumorigenic
Metastatic
BJ Fibroblast
Cancer Splicing Program
Oncogenic-like
Tumor suppressor-like

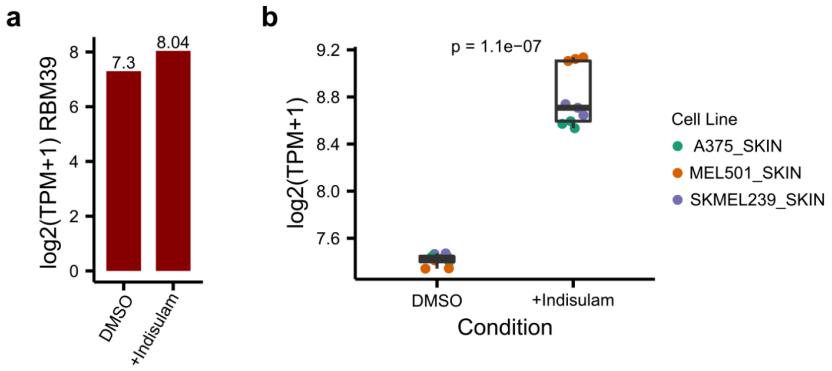
Supplementary Figure 1



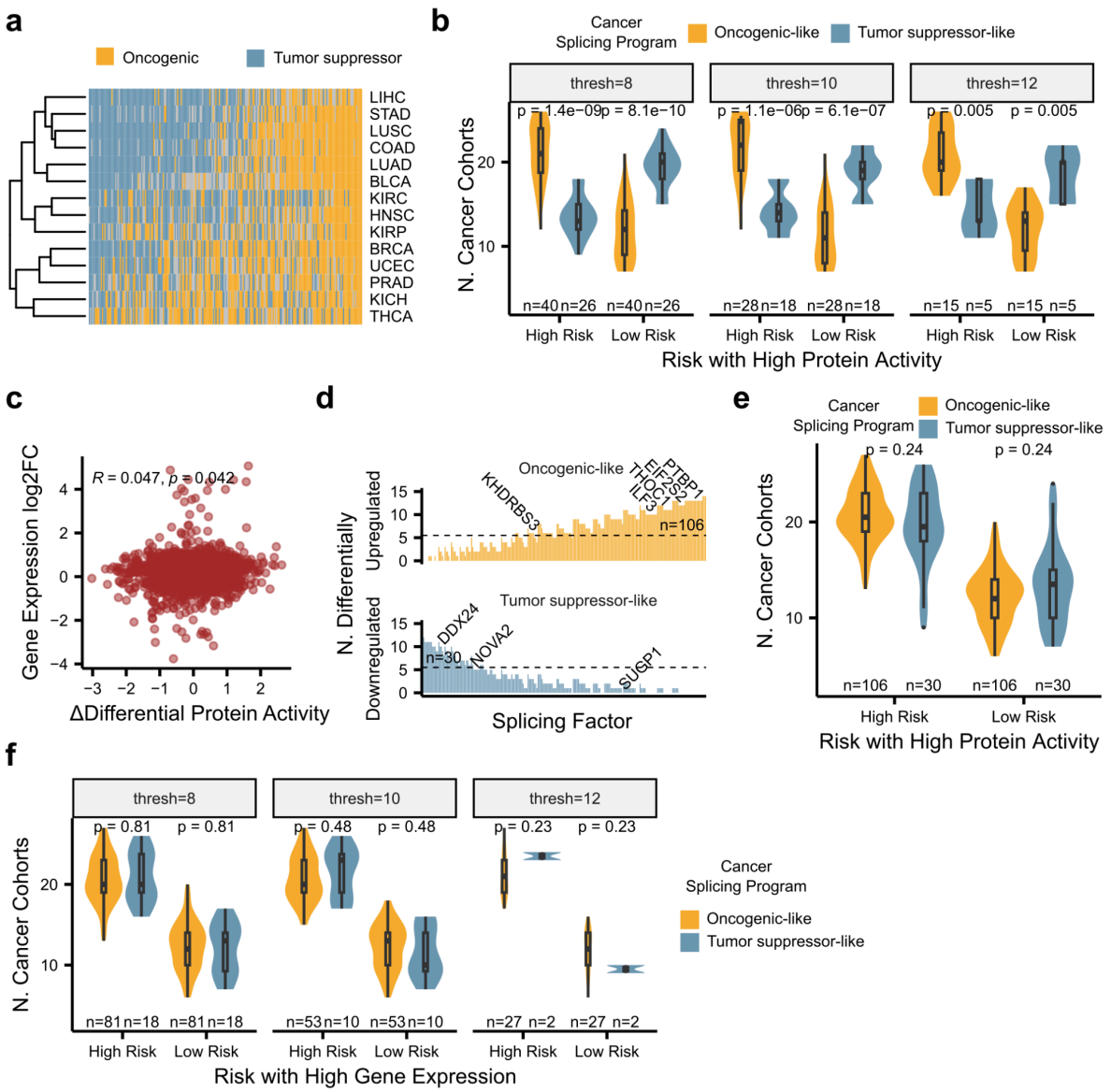
Supplementary Figure 2



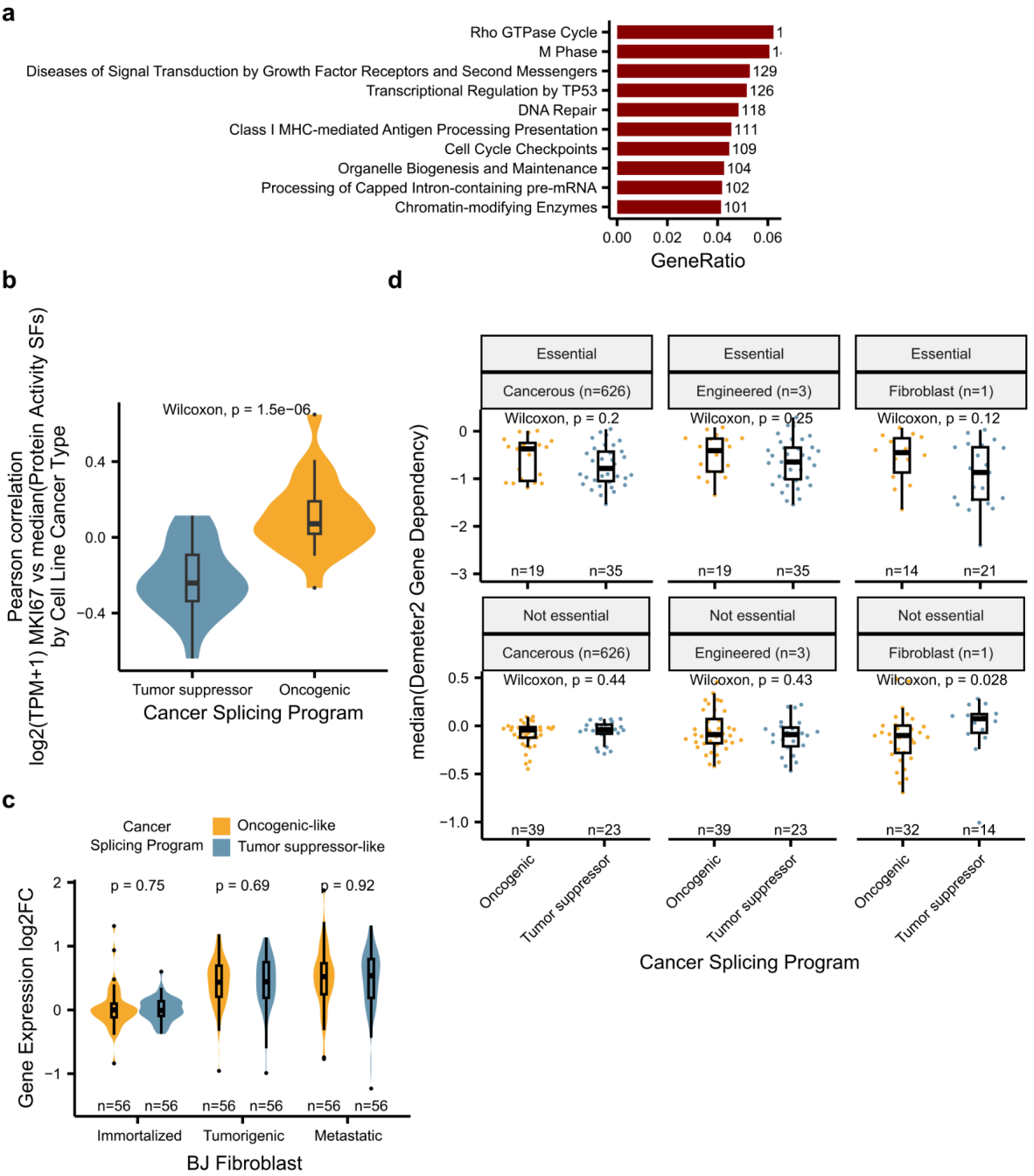
Supplementary Figure 3



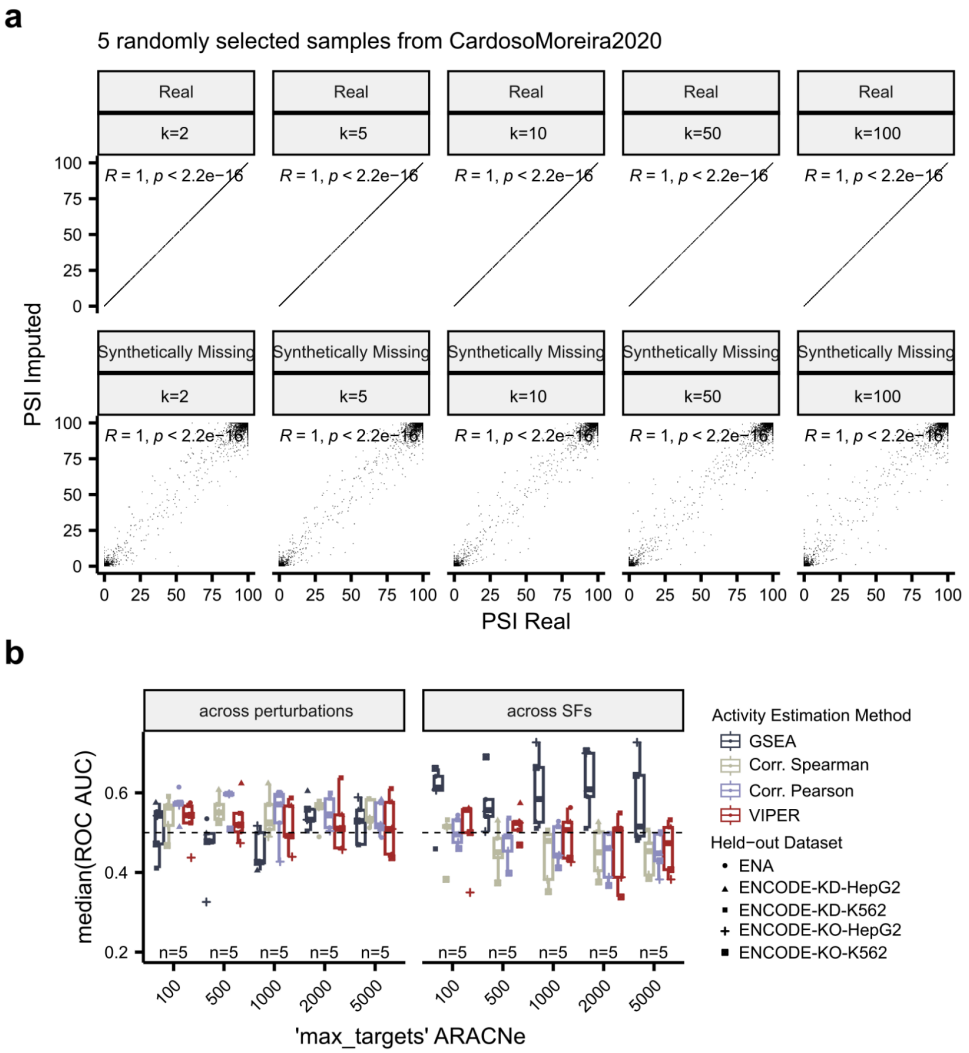
Supplementary Figure 4



Supplementary Figure 5



Supplementary Figure 6



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.

Supplementary Table 4. Overrepresentation analysis results for the genes corresponding to target exons of each defined cancer splicing program.

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. 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).
9. Bonnal, S. C., López-Oreja, I. & Valcárcel, J. Roles and mechanisms of alternative splicing in cancer — implications for care. *Nat. Rev. Clin. Oncol.* **17**, 457–474 (2020).
10. 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).
11. 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).
12. 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).
13. Garcia-Alonso, L. *et al.* Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer. *Cancer Res.* **78**, 769–780 (2018).
14. 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).
15. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
16. 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).
17. Badia-I-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer

biological activities from omics data. *Bioinform Adv* **2**, vbac016 (2022).
18. Zhang, C. *et al.* Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science* **329**, 439–443 (2010).
19. Aghamirzaie, D., Collakova, E., Li, S. & Grene, R. CoSpliceNet: a framework for co-splicing network inference from transcriptomics data. *BMC Genomics* **17**, 845 (2016).
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. Head, S. A. *et al.* Silencing of SRRM4 suppresses microexon inclusion and promotes tumor growth across cancers. *PLoS Biol.* **19**, e3001138 (2021).
23. Hegele, A. *et al.* Dynamic protein-protein interaction wiring of the human spliceosome. *Mol. Cell* **45**, 567–580 (2012).
24. Chen, L. & Zheng, S. Studying alternative splicing regulatory networks through partial correlation analysis. *Genome Biol.* **10**, R3 (2009).
25. Mazin, P. V., Khaitovich, P., Cardoso-Moreira, M. & Kaessmann, H. Alternative splicing during mammalian organ development. *Nature Genetics* **53**, 925–934 (2021).
26. Fu, X.-D. & Ares, M. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics* **15**, 689–701 (2014).
27. Zhao, W. *et al.* POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res.* **50**, D287–D294 (2021).
28. 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).
29. Lu, S. X. *et al.* Pharmacologic modulation of RNA splicing enhances anti-tumor immunity. *Cell* **184**, 4032–4047.e31 (2021).
30. Funnell, T. *et al.* CLK-dependent exon recognition and conjoined gene formation revealed with a novel small molecule inhibitor. *Nat. Commun.* **8**, 7 (2017).
31. Zhang, D. *et al.* Intron retention is a hallmark and spliceosome represents a therapeutic vulnerability in aggressive prostate cancer. *Nat. Commun.* **11**, 2089 (2020).
32. Gregersen, L. H. *et al.* SCAF4 and SCAF8, mRNA Anti-Terminator Proteins. *Cell* **177**, 1797–1813.e18 (2019).
33. 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).
34. Sun, C. The SF3b complex: splicing and beyond. *Cell. Mol. Life Sci.* **77**, 3583–3595 (2020).
35. 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).
36. Cretu, C. *et al.* Structural Basis of Splicing Modulation by Antitumor Macrolide Compounds. *Mol. Cell* **70**, 265–273.e8 (2018).
37. 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).
38. Roybal, G. A. & Jurica, M. S. Spliceostatin A inhibits spliceosome assembly subsequent to prespliceosome formation. *Nucleic Acids Res.* **38**, 6664–6672 (2010).
39. 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).
40. Dvinge, H., Kim, E., Abdel-Wahab, O. & Bradley, R. K. RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer* **16**, 413–430 (2016).
41. Koedoot, E. *et al.* Co-regulated gene expression of splicing factors as drivers of cancer progression. *Scientific Reports* **9**, 1–13 (2019).
42. Zhang, Q., Ai, Y. & Abdel-Wahab, O. Molecular impact of mutations in RNA splicing factors in cancer. *Molecular Cell* **84**, 3667–3680 (2024).
43. Zhu, W. *et al.* Roles of PTBP1 in alternative splicing, glycolysis, and oncogensis. *Journal*

of Zhejiang University. Science. B **21**, 122 (2020).
44. Cook, K. B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* **39**, D301–8 (2011).
45. Cvitkovic, I. & Jurica, M. S. Spliceosome database: a tool for tracking components of the spliceosome. *Nucleic Acids Res* **41**, D132–41 (2013).
46. Chowell, D. *et al.* Improved prediction of immune checkpoint blockade efficacy across multiple cancer types. *Nat. Biotechnol.* **40**, 499–506 (2022).
47. 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).
48. Riaz, N. *et al.* Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* **171**, 934–949.e16 (2017).
49. Sun, X. & Kaufman, P. D. Ki-67: more than a proliferation marker. *Chromosoma* **127**, 175–186 (2018).
50. Pitot, H. C. The molecular biology of carcinogenesis. *Cancer* **72**, 962–970 (1993).
51. 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).
52. Hahn, W. C. *et al.* Creation of human tumour cells with defined genetic elements. *Nature* **400**, 464–468 (1999).
53. McFarland, J. M. *et al.* Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat. Commun.* **9**, 4610 (2018).
54. A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer Cell* **42**, 301–316.e9 (2024).
55. Michael Krill-Burger, J. *et al.* Partial gene suppression improves identification of cancer vulnerabilities when CRISPR-Cas9 knockout is pan-lethal. *Genome Biology* **24**, 192 (2023).
56. Lenoir, W. F. *et al.* Discovery of putative tumor suppressors from CRISPR screens reveals rewired lipid metabolism in acute myeloid leukemia cells. *Nature Communications* **12**, 1–15 (2021).
57. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
58. Cilloni, D., Itri, F., Bonuomo, V. & Petiti, J. Mutations in Hematological Malignancies. *Cancers* **14**, (2022).
59. Maurin, M. *et al.* RBFOX2 deregulation promotes pancreatic cancer progression and metastasis through alternative splicing. *Nat. Commun.* **14**, 8444 (2023).
60. 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).
61. Moakley, D. F. *et al.* Reverse engineering neuron type-specific and type-orthogonal splicing-regulatory networks using single-cell transcriptomes. *Submitted*.
62. 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).
63. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
64. Cardoso-Moreira, M. *et al.* Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).
65. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
66. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS* **16**, 284–287 (2012).
67. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021) doi:10.1101/060012.
68. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare

ROC curves. *BMC Bioinformatics* **12**, 1–8 (2011).
69. 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).
70. 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).
71. 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).
72. 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).
73. Tseng, C.-K. *et al.* Human Telomerase RNA Processing and Quality Control. *Cell Rep.* **13**, 2232–2243 (2015).
74. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 1–21 (2014).
75. Dubrot, J. *et al.* In vivo CRISPR screens reveal the landscape of immune evasion pathways across cancer. *Nat. Immunol.* **23**, 1495–1506 (2022).
76. 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).
77. 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.