Alternative splicing programs in tumour development and progression

Larisa Morales Soto
Department of Human Genetics, Faculty of Medicine and Health Sciences
McGill University, Montreal, Quebec, Canada

December 2023

A thesis submitted to the McGill University in partial fulfillment of the requirements of the degree of Master of Science

DEDICATION

To my best friend Olive, for giving me a purpose in the darkest days.

To my human best friend Gil, for helping me release the bad energy.

To Noah, for always believing in me.

To Karen, for nourishing my creative self and making coffee all those mornings I couldn't.

To my parents, for laying the foundation of my core values.

To my friends and family who always cared for my wellbeing, even from the distance.

I completed this thesis thanks to all of you.

ABSTRACT

Background: Alternative splicing (AS) is a key step in the expression of nearly all human genes. Perturbations of this tightly regulated process have the potential to result in abnormal phenotypes, such as cancer. Widespread splicing alterations have been detected across all known cancer types. Studies on large patient cohorts have found mutations in core splicing factors and RNA binding proteins (RBPs), as well as dysregulated expression of pro-oncogenic transcript isoforms. Despite the numerous AS alterations described so far, previous studies are impacted by unwanted sources of variation, such as tumour purity and patient demographics. Most available methods cannot incorporate additional variables apart from two-group comparisons, and/or oversimplify their splicing model to enable multi-variable contrasts. Thus, there is a strong need for tools that can account for confounding factors and maintain an accurate definition of events when measuring changes across groups.

Methodology: Here, we present TRex, a novel computational tool to model the impact of multiple experimental variables when estimating differential AS from RNA-seq data. TRex derives event quantifications by collapsing the abundances of isoforms supporting each AS outcome and uses a generalized linear model to disentangle the effects of experimental variables of interest. To our knowledge, TRex is until now the only tool that supports all seven types of AS events, allows complex experimental designs, and removes the effect of confounding factors. Using TRex, we studied AS in 8,633 RNA-seq samples from The Cancer Genome Atlas (TCGA) spanning 24 cancer types, in order to identify AS events that are differentially regulated between normal tissue and tumour, as well as AS events associated with tumour stage, followed by identification of potential RBPs that mediate these changes.

Key results: Based on extensive simulation experiments, TRex outperformed two state-of-the-art methods, SUPPA2 and rMATS, by an average AUROC increase of 20% across various cut-offs used to define the ground truth. In addition, TRex effectively separated the effects of variables of interest and confounding factors. Application of TRex to RNA-seq data from TCGA showed the profound impact that sample-level confounders—such as age, sex, and tumour purity—have on the quantification of AS across experimental conditions; purity emerged as the most influential one of the three. In addition, the pan-cancer comparison of tumour vs normal samples revealed that intron retention is the most frequently dysregulated AS mechanism across cancers. We

specifically found strong enrichment of hallmark signature pathways, such as Myogenesis and Mitotic spindle, in genes that were differentially spliced in at least two-thirds of the cancer types studied. We also gained insights on how immune cell populations contribute to apparent tumour-associated splicing changes by analyzing single-cell transcriptomes of KIRC primary tumours. We found a set of events that were only associated with tumors after removing the effect of impurity. This group of events were strongly enriched for genes from the allograft rejection pathway—a process with completely opposite activities between cancer and immune cells. On the contrary, tumor-associated events were enriched in Hallmark pathways specific to cancer cell population, supporting their proposed neoplastic role. Lastly, an unbiased analysis of RBP regulons revealed a set of upstream splicing regulators whose binding to specific splice regions was associated with differential splicing of their targets between normal and tumour tissues. Our findings indicate that RBPs could have an important role in sustaining oncogenic pathways via splicing-mediated regulation of gene expression. In summary, this work represents the first and most comprehensive unbiased compendium of AS programs contributing to different axes of tumour biology at a pan-cancer scale.

RÉSUMÉ

Contexte: L'épissage alternatif (EA) constitue une étape cruciale dans l'expression des gènes humains, et les perturbations de ce processus hautement régulé peuvent engendrer des phénotypes anormaux, dont le cancer. Des altérations majeures de l'épissage sont observées dans tous les types de cancer connus. Les études sur de vastes cohortes de patients ont révélé des mutations dans les facteurs d'épissage centraux et les protéines de liaison à l'ARN, ainsi qu'une expression dérégulée d'isoformes de gènes pro-oncogéniques. Malgré la caractérisation de nombreuses altérations d'EA, les études antérieures sont affectées par des sources non désirées de variation, telles que la pureté tumourale et les caractéristiques démographiques des patients. La plupart des méthodes disponibles ne parviennent pas à intégrer des variables supplémentaires au-delà de la condition (maladie vs normal) ou simplifient excessivement leur modèle d'épissage pour permettre des contrastes multi-variables. Ainsi, il est impératif de disposer d'outils capables de tenir compte des facteurs de confusion tout en maintenant une définition précise des événements lors de la mesure des changements entre les groupes expérimentaux.

Méthodologie: Nous présentons TRex, un nouvel outil informatique permettant de modéliser l'impact de multiples variables expérimentales lors de l'estimation des changements différentiels d'EA à partir de données de RNAseq. TRex dérive les quantifications d'événements en regroupant les abondances des isoformes soutenant chaque résultat d'EA, et utilise un modèle linéaire généralisé pour démêler les effets des variables expérimentales d'intérêt. TRex est actuellement le seul outil qui prend en charge les sept types d'événements d'EA, permet des conceptions expérimentales complexes, et élimine l'effet des facteurs de confusion. En utilisant des expériences de simulation approfondies, TRex surpasse deux méthodes de pointe, SUPPA2 et rMATS, avec une augmentation moyenne de l'AUROC de 20 % à travers des seuils utilisés pour définir la positivité réelle. De plus, TRex sépare efficacement les effets de la condition et du lot en présence d'effets confondants simulés.

Résultats: L'analyse de plus de 10 000 échantillons de RNAseq issus de 31 types de cancer du Cancer Genome Atlas (TCGA) a apporté des éclairages essentiels sur la dérégulation de l'EA. La pureté tumourale s'est révélée être le facteur le plus influent, expliquant plus de 75 % de la variance dans certains cas, notamment dans le carcinome à cellules rénales (KIRC). La rétention intronique s'est avérée être le mécanisme d'EA le plus fréquemment dérégulé à travers les cancers, avec des

enrichissements dans des voies telles que la myogenèse et la transition épithélio-mésenchymateuse. Des protéines de liaison à l'ARN, dont SRSF12, ont été identifiées comme régulateurs potentiels d'événements spécifiques dans plusieurs types de cancer. En outre, l'analyse transcriptomique sur cellules uniques de tumeurs primitives de KIRC a permis d'identifier des contributions significatives des populations immunitaires. Les événements confondus liés au cancer, révélés après élimination de l'effet de l'impureté, ont montré un enrichissement dans la voie de rejet de greffe, avec des activités opposées dans les cellules cancéreuses et immunitaires. Les voies Hallmark principalement actives dans les cellules cancéreuses ont été enrichies en événements véritablement cancéreux, soutenant leur rôle néoplastique. En conclusion, ce travail représente la première compilation impartiale et complète d'événements d'EA associés au cancer, offrant des contributions significatives aux différents axes de la biologie tumourale à l'échelle pan-cancéreuse. L'introduction de TRex comme outil novateur dans l’amélioration substantielle des analyses de données de RNAseq renforce la fiabilité de ces résultats et ouvre la voie à une compréhension plus approfondie des mécanismes régissant l'EA dans le contexte du cancer.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

A3	Alternative 3’ Splice Site
ΔPSI	Delta Psi
A5	Alternative 5’ Splice Site
ACC	Adrenocortical Carcinoma
AF	Alternative First Exon
AL	Alternative Last Exon
AS	Alternative Splicing
AUROC	Area Under the Receiver-Operator Curve
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast Invasive Carcinoma
ccRCC	Clear-Cell Renal Cell Carcinoma
CD44	Cluster Of Differentiation 44
CDF	Cumulative Distribution Function
CDNA	Complementary DNA
CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon Adenocarcinoma
DAS	Differential Alternative Splicing
DAS	Differentially Alternatively Spliced
DLBC	Lymphoid Neoplasm Diffuse Large B-Cell Lymphoma
DPSI	Delta PSI
DS	Differential Splicing
DSR	Differential Splicing Regulation
E	End
EMT	Epithelial-Mesenchymal Transition
ENCODE	Encyclopedia of DNA Elements
ESCA	Esophageal Carcinoma

ESEs	Exonic Splicing Enhancers
ESSs	Exonic Splicing Silencers
ESTIMATE	Estimation Of Stromal and Immune Cells In Malignant Tumours Using Expression Data
FDR	False Discovery Rate
GBM	Glioblastoma Multiforme
GEDI	Gene Expression Decomposition and Integration
GLM	Generalized Linear Model
GORA	Gene Overrepresentation Analysis
HNRNPs	Heterogeneous Nuclear Ribonucleoproteins
HNSC	Head And Neck Squamous Cell Carcinoma
ISEs	Intronic Splicing Enhancers
ISSs	Intronic Splicing Silencers
KD	Knockdown
KICH	Kidney Chromophobe
KIRC	Kidney Renal Clear Cell Carcinoma
KIRP	Kidney Renal Papillary Cell Carcinoma
LAML	Acute Myeloid Leukemia
LFC	Log-2 Fold Change
LGG	Low-Grade Glioma
LIHC	Liver Hepatocellular Carcinoma
lncRNAs	long non-coding RNAs
LOF	Local Outlier Factor
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
MESO	Mesothelioma
MRNA	Messenger RNA
MX	Mutually Exclusive Exons
NK	Natural Killer
OV	Ovarian Serous Cystadenocarcinoma

PAAD	Pancreatic Adenocarcinoma
PCA	Principal Component Analysis
PCPG	Pheochromocytoma And Paraganglioma
PCR	Polymerase Chain Reaction
PCR	Polymerase Chain Reaction
PCs	Principal Components
PFMS	Position Frequency Matrices
PK	Pyruvate Kinase
PRAD	Prostate Adenocarcinoma
pre-MRNA	Premature Messenger RNA
PSI	Percent Spliced In
Py	Polypyrimidine
qPCR	Quantitative Polymerase Chain Reaction
RBFOX1	RNA Binding Fox-1
RBPs	RNA Bind Proteins
RE	Regulatory Element
READ	Rectum Adenocarcinoma
RI	Retained Intron
rMATS	Replicate Multivariate Analysis of Transcript Splicing
RNA	Ribonucleic Acid
RNA-seq	RNA Sequencing
S	Start
SARC	Sarcoma
SE	Skipped Exon
SF	Splicing Factor
SFS	Splicing Factors
SKCM	Skin Cutaneous Melanoma
snoRNAs	Small Nucleolar RNAs
SNRNA	Small Nuclear RNA
SNRNPs	Small Nuclear Ribonuclear Proteins

SNRPA	U1 Small Nuclear Ribonucleoprotein A
SR	Serine/Arginine-Rich
SS	Splice Site
SSEs	Structural Splicing Enhancers
STAD	Stomach Adenocarcinoma
SUPPA2	Sequencing Unified Patch-Based Pipeline for Alternative Splicing Analysis
TC	Technical Confounder
TCGA	The Cancer Genome Atlas
TGCT	Testicular Germ Cell Tumours
THCA	Thyroid Carcinoma
TPM	Transcripts Per Million
tSNE	T-Stochastic Neighbor Embedding
U2AF	U2 Auxiliary Factor
UCEC	Uterine Corpus Endometrial Carcinoma
UTR	Untranslated. Region
UVM	Uveal Melanoma
VEGF	Vascular Endothelial Growth Factor
VOI	Variable Of Interest
WGS	Whole-Genome Sequencing

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis supervisor, **Prof. Hamed Shateri Najafabadi**, for their invaluable guidance and mentorship throughout the entire journey of completing this thesis. Prof. Najafabadi has been an exceptional mentor, whose dedication and commitment to my academic and personal development were instrumental for the completion of this work. I extend my thanks to the members of my Supervisory Committee, professors **Celia Greenwood** and **Logan Walsh**, whose contributions have enriched my understanding of the scientific process and shaped my critical thinking.

I am truly grateful for the unwavering support of my colleagues **Ariel Madrigal-Aguirre**, **Aldo Hernandez-Corchado**, whose work and expertise were critical to perform the single-cell validations and the analysis of upstream regulators, respectively. I would also like to thank my colleague **Rached Alkallas** for providing technical guidance throughout the course of this project. Lastly, my gratitude extends to all the members of the **Computational and Statistical Genomics Lab** for the critical feedback and patient encouragement they provided over the course of my degree.

I am profoundly grateful to some of my dearest friends and family—your unconditional love and support served as a guiding light, keeping me from getting lost amidst the challenges of graduate school. Specially **Olive**, **Gil**, **Noah** and **Karen**, your presence was my anchor during some of the most difficult times of my life. Words cannot capture the depth of my appreciation, so I'll simply say that I would not be here if it wasn't for all of you.

This work was possible thanks to the resources generated by large international research consortia and extensive collaborations with groups within McGill University:

The pan-cancer analyses are based on the comprehensive RNA sequencing and whole-genome sequencing compendium of datasets generated by **The Cancer Genome Atlas (TCGA)** Research Network: https://www.cancer.gov/tcga.

The simulation experiments used for benchmarking were generated from a Serine-Arginine Splicing Factor 9 (SRSF9) knock-down dataset from the **ENCODE Consortium** and the ENCODE production laboratories.

The single-cell clear cell Renal Cell Carcinoma (ccRCC) dataset was generated as part of the McGill **RCC Atlas Project**, led by the CSG Lab and the Cancer Genomics Lab, and processed by Ariel Madrigal-Aguirre. This data is unpublished and was used with permission from the authors.

This project was funded by the **Canadian Institutes of Health Research** (CIHR) grant PJT-173317 and the **Mitacs Globalink Graduate Fellowship**.

FORMAT OF THE THESIS

This thesis was prepared in adherence to the traditional thesis format outlined by McGill University's Faculty of Graduate and Postdoctoral Studies. This thesis comprises six main chapters. **Chapter 1** defines the scientific foundation of the research project and gives a comprehensive introduction to the pertinent fields. **Chapter 2** describes the methodology followed and the materials used in this research. **Chapter 3** presents the results supporting the hypothesis of this thesis; work that is being adapted into a manuscript format for submission to a peer-reviewed journal. **Chapter 4** provides a thorough discussion of the main findings in terms of the limitations of this study and their impact on the field. **Chapter 5** summarizes the key takeaways of this thesis and outlines future areas of inquiry. **Chapter 6** lists all the existing work used to support the arguments of this thesis. **Chapter 7** encompasses supplementary tables and figures, as well as the necessary copyright clearances.

CONTRIBUTION OF THE AUTHORS

Explicit author contributions are listed below:

Larisa Morales Soto (L.M.S) built the computational pipelines to process and analyze all the datasets described, developed the main software package, conducted the experiments, interpreted the results, generated the figures, and wrote this thesis.

Ariel Madrigal-Aguirre (A.M.A) pre-processed the single cell data of clear-cell Renal Cell Carcinoma patients and provided support with running GEDI for the pathway activity analysis.

Aldo Hernandez Corchado (A.H.C) assisted with the identification of transcriptome-wide RBP binding sites using AffiMx.

Rached Alkallas (R.A) downloaded the raw RNA-seq data required for the TCGA analysis and provided guidance to develop the pipeline for transcript quantification.

Hamed Shateri Najafabadi (H.S.N) and L.M.S designed the statistical models, developed, and optimized the computational methods.

H.S.N. conceived the core idea for this project, directed the study, and reviewed this thesis.

CHAPTER 1 PROJECT FOUNDATION

1.1 Introduction

From the emergence of a malignant cell mass to the acquisition of drug resistance, cancerous cells undergo a series of molecular transformations that allow the tumour to grow, migrate to other tissues and evade the immune system. Dysregulation of gene expression is one of the main drivers of cellular reprogramming during oncogenesis and cancer progression [1]. Alternative splicing (AS) is one of the most important mechanisms involved in the regulation of gene expression. Nearly all protein-coding genes undergo a form of AS to produce functional products [2]. Thus, alterations at any stage of the splicing process have the potential to result in transformations that impair cellular homeostasis [3-5]. In fact, studies on large cohorts have found extensive alterations in splicing programs at two levels: (i) mutations and/or expression changes of core splicing factors [6, 7]; and (ii) changes in the relative abundance of pro-oncogenic transcript isoforms [3, 4, 8]. Some of these patterns can be generalized [9], while others are specific to certain cancer types, subtypes, and even cellular subpopulations [10, 11] [12], suggesting the existence of multiple mechanisms driving splicing dysregulation in cancer.

One of the largest efforts to characterize the AS landscape of primary tumours reported major changes in the inclusion/exclusion rates of cassette exons (a type of AS) in more than 1250 genes across 14 cancer types [12] from The Cancer Genome Atlas (TCGA) [13]. In other comparisons of primary tumours and normal tissues, significant expression changes were found in more than 70% of core splicing factors and over 80% of all known RNA-binding proteins (RBPs) [14-16], many of which are involved in regulation of differential RNA splicing [17].

Splicing programs can also distinguish metastatic from non-metastatic conditions. A recent study involving our team reported previously unknown RNA structural elements that regulate the inclusion of thousands of cassette exons promoting metastasis in breast cancer [18]. We demonstrated that this pro-metastatic regulation of splicing is mediated by RNA-protein interactions dependent on secondary structures of the messenger RNA (mRNA) [18]. This work, among others [1, 9, 11, 19], suggests that RBP-mediated regulation of splicing has an integral role in the development of oncogenic phenotypes such as metastasis, and highlights the need for systematic characterization of their potential to drive tumour evolution.

For cancerous cells to survive under strong selective pressures, such as antigen-based therapies, a fast shift in their phenotype is critical. AS is often used by cancerous cells to shift gene expression profiles that enable/disable key cellular features [20]. Thanks to the flexibility conferred by AS, cells can quickly modify their transcriptome and increase their fitness in the presence of a selective pressure, for example, by skipping the exon that includes the downstream signaling domain of the protein targeted by a drug [21]. The importance of AS regulation in cancer is further underlined by the large number of ongoing clinical trials for drugs that target components of the spliceosome or modulate splicing [1]. Despite the extensive splicing rearrangements that have been described so far, many questions remain unanswered. For instance, the upstream regulatory pathways and the tissue specificity of pro-oncogenic splicing programs are still largely unknown.

1.2 Hypothesis

Alternative splicing is a mechanism exploited by cancerous cells to promote the development of oncogenic phenotypes.

1.3 Objectives

1. Develop and validate a robust computational method to study AS changes associated with experimental variables using short-read RNA sequencing data.
    a. Assess the method's performance against comparable methods in a systematic manner.
2. Characterize the alternative splicing landscape of tumour development and progression.
    a. Detect changes in AS between primary tumours and normal tissues and validate their neoplastic role using single cell data.
    b. Identify stage-associated AS programs at a pan-cancer scale.
3. Nominate *trans*-acting factors that may drive the dysregulation of alternative splicing programs of interest.

1.4 Literature review

This section contains a literature review that starts by describing the mechanisms of RNA splicing, the importance of AS, the different types of AS, and their homeostatic regulation. It continues by placing AS in the context of cancer, highlighting what is known about its role in disease progression and how it has been exploited in recent therapeutics. It concludes with a discussion of the pivotal experimental methods and current computational methods used to study AS.

1.4.1 RNA splicing

Eukaryotic genes are made of alternating functional blocks called exons and introns. Exons are defined as any region that is present in the mature RNA, and introns are the stretches of sequence between them that are only present in the premature mRNA (pre-mRNA) (***Figure 1***). [22]. Splicing is a critical step required for gene expression in eukaryotes [23], and one of the largest sources of functional diversity in the cell [24]. It consists of removing introns and ligating exons to form a mature RNA This seemingly simple process is accomplished by one of the most exquisite molecular machines in the cell: the spliceosome [23]. The RNA-protein complex is formed by up to 200 proteins [25] and five small nuclear RNAs (snRNAs) that interact with each other to form five distinct subunits (U1, U2, U4, U5, and U6) called small nuclear ribonuclear proteins (snRNPs) [26]. The components of the spliceosome interact with other RBPs, also called splicing factors (SFs) [27], as well as elements found within or nearby introns and exons to determine the splicing outcome of a given gene [28-30].

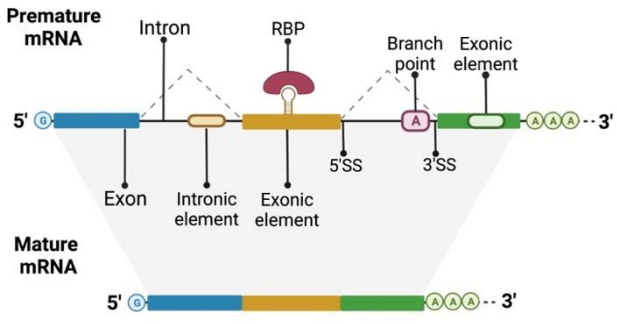


Figure 1 Overview of mRNA splicing

Example of constitutive splicing showing the elements required for the process.

An overview of the splicing mechanism

The molecular mechanisms of splicing have been extensively characterized [31] [32] [33] and its detailed description goes beyond the scope of this thesis. In summary, pre-mRNA splicing (**Figure 1**) proceeds as follows:

1. ***Splice site recognition:*** To begin, the spliceosome determines where splicing will take place. The precise distinction of introns and exons depends on several conserved sequences. The exon-intron boundary at the 5' end of the intron contains a sequence called the 5' splice site (SS). Similarly, there is a 3' SS sequence located at the intron-exon boundary on the 3' end of the intron [32]. These two SS surround a third critical sequence called the branch point site, which is distinguished by its subsequent polypyrimidine tract (Py tract) [34]. Thus, the first stage of the process is the recognition of the 5'SS by the snRNP U1 [35], and the binding of the two subunits of U2 auxiliary factor (U2AF) protein to the Py tract and the 3' SS [36, 37], respectively.
2. ***Spliceosome assembly:*** Once the SSs have been recognized, the remaining snRNPs U2, U4, U5, and U6 join the complex and induce a series of rearrangements that ultimately bring the three SSs closer together [38-41]. Further series of rearrangements trigger the downstream catalytical steps where two successive transesterification reactions occur.
3. ***Intron cleavage:*** The first reaction occurs between the branch site and the 5'SS, which cleaves the 5' exon-intron junction and forms an intron lariat intermediate [40].
4. ***Exon ligation:*** The second transesterification proceeds between the liberated 5' end of the exon and the 3' SS, releasing the intron and joining the 5' and 3' ends of the two subsequent exons [42].

Once the fourth stage is completed, the mature mRNA is released [43], and the spliceosome disassembled [44]. The newly formed mRNA molecule consists only of exons and can now be further processed and transported.

Alternative splicing of the mRNA

Alternative splicing enables the generation of multiple variants of the same gene through the selective inclusion or exclusion the introns and exons (**Figure 1**). AS is the main source of transcriptional diversity, producing 253,000 different transcripts [45, 46] from only 63,000 protein-coding and non-coding genes [47]—the true number of transcripts encoded by these gene

is likely even higher, as recent isoform sequencing advances have revealed an increasing number of novel isoforms [48, 49].

Although AS of each gene can result in many isoforms through complex combinations of exon inclusion/exclusions or intron retention events, this complexity can often be simplified by studying the splicing locally, by dividing the process into a series of AS "events" at specific introns and exons. These AS events are often defined in a binary manner, so that there are two possible outcomes, each characterized by a distinct composition of introns and exons. Based on the mechanistic pathways that yield such specific intron/exon arrangements, AS events is classified in seven categories, also called *event types* (**Figure 2**). The most common ones are cassette exon skipping (SE), intron retention (IR), mutually exclusive exons (MX), alternative 5' SS (A5), and alternative 3' SS (A3) [22], in addition to alternative first exon (AF) and alternative last exon (AL) events [50, 51].

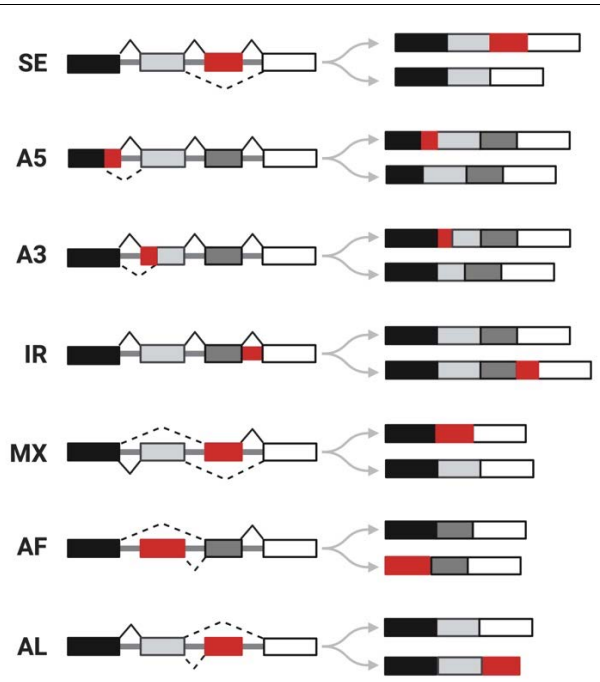


Figure 2 Types of alternative splicing events

Dashed lines in the pre-mRNA diagrams represent the junctions that give rise to the two possible outcomes shown to the right of each event. Figure adapted from the "*mRNA Splicing Types*" template in BioRender (www.biorender.com). SE = exon skipping; A5 = alternative 5' SS; A3 = alternative 3' SS; RI = intron retention; MX = mutually exclusive exons; AF = alternative first exon; AL = alternative last exon.

These AS events work together to determine the relative abundance of alternative isoforms, which can have drastic effects on the cell function [22]. For example, different isoforms of a single gene can create a binary switch that drives a developmental transition [52] or enables the establishment of tissue-specific transcriptomic profiles [53]. These examples are indeed extreme cases, and it is likely that most AS products have more subtle contributions. However, such a diverse compendium of functional products is a powerful substrate for evolution to act on, bypassing the need for new genes to emerge from scratch [22] . In fact,
AS has served as a mechanism for molecular innovation, allowing the generation of novel functional elements [52] and the rewiring of regulatory networks [2].

Regulation of alternative splicing

The process of alternative splicing is intricately controlled to ensure the correct outcome in any given cellular context (Reviewed in [23]). Its regulatory cascade involves a combination of *cis* regulatory elements within the pre-mRNA and *trans*-acting factors (**Figure 3**) that interact both with the pre-mRNA and with each other. This results in a high degree of complexity that allows splicing landscapes to vary across tissues, developmental stages, and physiological conditions even when all cells have identical genomes. In the following sections, I will describe the different regulatory elements and how they contribute to this molecular marvel.

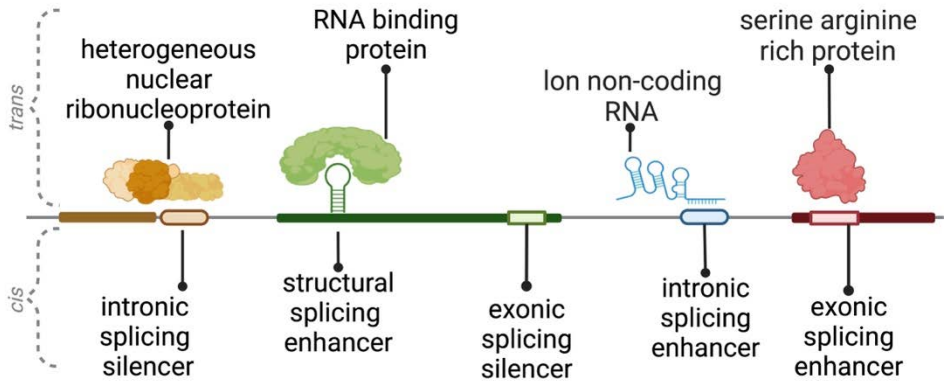


Figure 3 Splicing regulatory elements.
Summary of the main *cis* and *trans* elements described in the text.

Cis-regulatory elements: These elements are specific nucleotide stretches located within the pre-mRNA that serve as landmarks for *trans*-acting factors to determine where splicing will occur [54]. The two main groups are:

1. Exonic Splicing Enhancers (ESEs) and Silencers (ESS): ESEs and ESSs are short RNA sequences located within exons that promote SS recognition and enhance or block exon inclusion, respectively [55].
2. Intronic Splicing Enhancers (ISEs) and Silencers (ISSs): ISEs and ISSs are located within intronic sequences and carry out the similar functions to their exonic counterparts [55].

There is a particular kind of *cis* regulatory elements, that can act both as silencers or enhancers, whose regulatory function does not depend on sequence motifs: the Structural Splicing Elements (SSEs). SSEs can be located in both exonic and intronic regions and are characterized by the formation of secondary RNA structures that can be recognized by RBPs in a sequence-independent manner [18].

Trans-regulatory Elements: These factors can be both protein and/or RNA-based elements located in the nucleus that recognize *cis* regulatory elements and interact with multiple components of the spliceosome to modulate SS selection (reviewed in [56] [54]). Some of these factors are:

1. RBPs: RBPs play a crucial role in alternative splicing regulation. It has been estimated that there are at least 1,500 human genes that can encode for RBPs [57], but clear functional roles and binding specificities have only been established for a fraction ~~handful~~ of them [58, 59]. They bind to splicing regulatory elements and interact or compete with the spliceosome, modulating splice site selection. RBPs can function as splicing enhancers or repressors depending on their binding location and associated protein complexes. Two classes of RBPs of particular interest are:

   a. Serine/arginine-rich (SR) proteins: SR proteins are a family of RBPs that generally promote exon inclusion by interacting with exonic splicing enhancers and facilitating spliceosome assembly [60].
   b. Heterogeneous nuclear ribonucleoproteins (hnRNPs): hnRNPs are another class of RBPs that can have both positive and negative effects on splicing regulation [61]. They often bind to intronic splicing silencers and repress exon inclusion but can also display context-dependent effects [54].

2. Non-coding RNAs: Emerging evidence suggests that non-coding RNAs, including long non-coding RNAs (lncRNAs) and small nucleolar RNAs (snoRNAs), can participate in alternative splicing regulation through various mechanisms, such as DNA-RNA or RNA-RNA interactions and recruitment of splicing factors (SFs) [62] .

Alternative splicing is also affected by other molecular processes happening simultaneously in the cell. Cellular signaling pathways can modulate AS through kinase-mediated phosphorylation events [63]. Kinases can target SFs, RBPs, or components of the spliceosome, altering their activity [64], localization [65], or interactions [66]. Moreover, since splicing occurs co-transcriptionally, factors involved in transcription elongation, such as RNA polymerase II speed and pausing, can also impact SS recognition and splicing efficiency [67]. Moreover, epigenetic modifications, including DNA methylation and histone marks, can contribute to the complex regulation of AS [68]. In addition, changes in spliceosome composition, such as the presence or absence of specific snRNPs, as well as dynamic rearrangements and conformational changes, strongly contribute to shaping the AS landscape [69].

1.4.2 The role of alternative splicing in tumour development and progression

As a once-normal-cell progresses through a series of oncogenic transformations, it undergoes major rearrangements to its genome, epigenome, and transcriptome [70]. Such changes can arise through a variety of mechanisms, including AS [71]. Given its role in generating functional diversity, the dysregulation of AS can result in selective advantages for cancerous cells over time [1]. In fact, various studies have shown how specific isoforms can provide enhanced growth capacity, improve cell migration and metastasis, enable escape from cell death, rewire cell metabolism, alter immune response, or enable drug resistance [1, 71, 72]. The following sections provide a summary of the current knowledge regarding pro-oncogenic splicing alterations and how they favor cellular programs that impact every axis of cancer biology.

Alternative splicing aberrations found across cancer types

Several studies have found extensive dysregulation of both SFs and alternative isoforms using large patient cohorts. A recent pan-cancer study of 14 cancer types from the TCGA reported that tumours harbor many splicing junctions not frequently encountered in normal samples [11].

Similarly, an analysis of three cancer types with the largest number of normal samples in the TCGA revealed a set of cancer specific AS events that are dysregulated in tumours and harbored strong prognostic value. Others have also relied on the data generated from this large patient cohort to find malignant splicing programs, and their results support the general conclusion that AS is extensively dysregulated in tumours [73-75].

Further in-depth inspection of AS dysregulation in tumours with different developmental origins suggest the existence of both general and cancer-specific pro-oncogenic AS programs. The core SFs U2AF1, SRSF2, SF3B1 and ZRSR2 are frequently mutated in myeloid and lymphoid neoplasms [76, 77]. In contrast, these SFs have a much lower mutation rate in solid tumours [78], where they seem to be affected mainly at the expression level [79]. Despite the strong genetic evidence supporting their importance in tumour development and progression [78], there is only speculative information regarding the mechanisms leading to the perturbation of SFs and whether they have a causal role in driving oncogenic pathways [78].

To facilitate the interpretation of widespread AS changes found across cancers, several web-based visualization tools and databases have emerged. The ASCancer Atlas compiles 2006 AS events reported in the literature, and over 2 million events derived from different computational tools [80]. Another example is OncoSplicing, which connects results from previous pan-cancer projects with reference annotations of clinical significance of event-associated transcripts [81]. Such resources not only serve as repositories for existing data, but also enable the generation of new hypothesis and their exploration by scientists specialized in particular cellular processes.

Cancer hallmarks impacted by splicing perturbations

Tumours of various cancer types often share a set of core cellular processes and characteristics, commonly referred to as *cancer hallmarks* [82]. Cancer hallmarks are the result of perturbations to different molecular processes, often caused by genetic instability or environmental factors [70]. As described by Hanahan and Weinberg, there are eight cancer hallmarks: sustaining proliferative signaling, evading growth suppressors, activating invasion and metastasis, enabling replicative immortality, inducing angiogenesis, resisting cell death, deregulating cellular energetics and avoiding immune destruction [70, 83]. Interestingly, all of them have been associated with at least one form of alternative splicing dysregulation [1, 71, 78] [72].

A striking example of a pro-oncogenic change in AS programs lies behind the well-characterized Warburg effect [84, 85]. In cancer cells, a high rate of glucose uptake is paralleled by low activity of the oxidative phosphorylation pathway, and glucose is selectively processed through aerobic glycolysis [84, 85] . The last step of the glycolytic pathway requires the enzyme pyruvate kinase (PK) [86]. The *PKM* gene produces two splicing variants from a pair of mutually exclusive exons: the M1 and M2 isoforms [87]. *PKM1* is expressed in most adult tissues, whereas *PKM2* is only expressed in embryonic stages [88]. It has been reported that tumours selectively express the *PKM2* variant [88, 89], where it contributes to **tumour growth** by **altering cellular energetics** to enable proliferative biosynthesis [85, 88], although the mechanistic routes upstream and downstream of this splicing switch are still poorly understood [85].

Complex AS programs also affect the biology of one of the most studied tumour suppressors, p53 [90]. The canonical p53 is the most abundant protein of twelve isoforms produced by *TP53* gene [91, 92], all of which show differential expression between tumour and normal tissues [93]. p53 isoforms can mediate both cell death and survival [92, 94, 95] to enable **replicative immortality** of cancer cells. But a mechanistic understanding of their interplay in different contexts is still needed to exploit them as a therapeutic strategy [90].

A similar case is the *BCL-X* gene, which produces two isoforms with directly opposite functions. Splicing of the canonical 5' SS of exon 2 yields the anti-apoptotic isoform BCL-XL, whereas an alternative 5' SS results in the pro-apoptotic variant BCL-XS [96]. Several hnRNPs [96, 97] and other RBPs [98] have been identified as direct regulators of BCL-X AS. The balance between these upstream factors has a strong impact on whether BCL-X confers **resistance to cell death** or enables proliferation.

The AS landscape of genes with known implications in cancer biology has expanded our understanding of how factors that are key for cell-survival can have context-dependant effects. The well-characterized regulator of **angiogenesis**, vascular endothelial growth factor (*VEGF*), can generate pro-angiogenic and anti-angiogenic variants through alternative 3' and 5' SS selection [99]. As another example, a switch between two isoforms of the cluster of differentiation 44 (*CD44*) gene can induce epithelial-mesenchymal transition (EMT) [100], thus promoting **cell invasion and metastasis**.

These are only a handful of examples that showcase the breadth of AS dysregulation in cancer; a more comprehensive set has been compiled in detailed reviews on the matter [1, 71, 72].

Despite the limited number of examples presented here, they serve to showcase the impact that dysregulated splicing can have on every axis of tumour biology. The contribution of AS to all cancer hallmarks is highlighted by the increasing attention to splicing modulation as a promising therapeutic strategy [101, 102].

1.4.3 Methods to quantify alternative splicing

Alternative splicing is commonly measured with a metric called Percent Spliced In (PSI), often denoted with the Greek letter ψ. In the case of SE (skipped exon) events, this metric is essentially the ratio between the abundance of inclusion isoforms (isoforms that include the cassette exon of interest) and the total isoform abundance (**Figure 4**). However, what is often the focus is not the just the absolute inclusion level of an exon in each sample, but its relative change across biological conditions. Thus, the actual metric of interest is the difference in PSI (ΔPSI) between conditions. This type of analysis is called differential alternative splicing (DAS).

Both PCR and qPCR have been widely used to study AS [103] [104] . They involve reverse transcription of RNA into complementary DNA (cDNA) followed by PCR amplification of specific splicing variants using primers designed to span exon-exon junctions. By comparing the amplification products, the relative abundances of different splicing variants can be determined. Nonetheless, these methods require specific primers designed to amplify splice variants of interest, which limits their application to cases where the alternative isoforms are well characterized. In addition, they are considered low-to-mid throughput technologies due to the cost and time consumption.

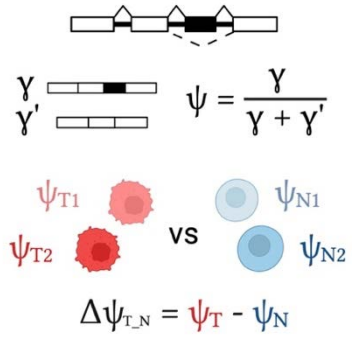


Figure 4 Estimation of PSI values.

Calculation of PSI values for an exon skipping event. γ denotes the cassette exon inclusion counts of event *k*, γ' its skipping counts, and γ'' the total event counts. T = tumour; N = normal

On the other hand, RNA-sequencing provides the throughput required to quantify thousands of splicing events in different samples. However, analysis of data obtained from RNA-seq can pose challenges, requiring a series of methods to quantify AS from RNA-seq data and perform statistical testing strategies to detect changes between experimental groups. More than a dozen computational tools are currently available to perform these tasks. Even though there are several differences and similarities between them, it is worth mentioning that a crucial feature is the definition of AS events themselves. A reference set of AS events becomes the basis of the entire quantification and downstream statistical testing pipelines. Hence, we can classify existing methods in three main groups based on their approach to define AS: (1) intron-based methods such as LeafCutter [105], which define AS based on intronic regions; (2) exon-centric tools like rMATS [103], SUPPA2 [104], MISO [106] and SplAdder [107], which use the coordinates of specific splice junctions to define classical AS events, or like DEXSeq [108] and limma [109] that use the entire exon as basis for quantification of differential exon usage; and (3) other approaches based on splicing graphs to derive non-canonical events like MAJIQ [110] and MAJIQ_v2 [111], which define local splice variants from splits in the splicing graph.

Despite differences in their implementation and modelling approach, all methods follow a similar logic in the progression of steps. First, they build the annotation of splicing events to be analyzed, either from a reference genome or directly from the RNA-seq libraries. Then, they quantify the abundance of a given unit (exon, intron, inclusion isoforms, etc.) by mapping the RNA-seq reads to a reference genome/transcriptome and assigning them to the previously annotated splicing events. Finally, the event counts are fed into a statistical model that estimates an effect size and a P-value for the difference across experimental groups.

Table 1 shows a non-exhaustive comparison of existing methods to quantify AS. Extensive performance benchmarking of these and other methods have been performed by others [112, 113]. However, given their relevance in the community and the present work, two exon-centric methods are worth mentioning in more detail: rMATS (Replicate Multivariate Analysis of Transcript Splicing) [103] and SUPPA2 (Sequencing Unified Patch-based Pipeline for Alternative Splicing Analysis) [104].

Table 1 Comparison of tools used for AS quantification and differential testing.

Method	*Designs supported*	*Detect novel events*	*AS definition*	*Counts used*	*Differential testing*	*Effect size estimator of differential AS*	*Reference*
rMATS	two groups	yes	exon centric events	junctions	likelihood ratio test	ΔPSI	[103]
SUPPA2	two groups	no	exon centric events	transcripts	empirical distribution	ΔPSI	[104]
SplAdder	two groups	yes	Splice graph based	junctions	-	-	[107]
LeafCutter	two groups	yes	intron based	junctions	likelihood ratio test	ΔPSI	[105]
MAJIQ, MAJIQ_v2	two groups	yes	Splice graph based	junctions	multiple tests in v2	ΔPSI	[110, 111]
MISO	two groups	yes	exon centric events	transcripts	Bayes factors	ΔPSI	[106]
DEXSeq	complex	no	exonic bins	exonic	Wald test	Log 2-fold change	[108]
limma	complex	no	exonic bins	exonic	T-test	Log 2-fold change	[109]

rMATS is a count-based method that relies on exon-exon junction reads to quantify five types of AS events (SE, A3, A5, R1 and MX). This modular tool internally derives AS events from the evidence present in the data, which can make it difficult to compare samples from different experiments that were analyzed separately. rMATS uses a hierarchical statistical model to simultaneously account for sampling uncertainty in individual replicates and technical variability among replicates. Additionally, it includes a model specifically designed for paired replicates between sample groups, which improves the statistical power by introducing a bivariate normal distribution with a correlation parameter to model the correlation among matched pairs. For its differential testing, rMATS uses a likelihood-ratio test to assess the statistical significance over a user-defined magnitude of splicing change. While rMATS is a powerful tool for detecting differential AS from RNA-seq data, it is computationally intensive, especially when analyzing large datasets with many replicates. Moreover, rMATS was designed to perform pairwise comparisons between two groups and cannot be used to analyze datasets with complex experimental designs.

SUPPA2 is another computational tool designed for the analysis of AS patterns using RNA-seq data. First, it constructs a comprehensive set of seven AS models (SE, A3, A5, R1, MX, AF, and AL) from a user-supplied reference annotation. SUPPA2 aggregates the counts of the two types of outcomes of each event in question (see **Figure 2**) to calculate PSI values per sample, which are then aggregated into a ΔPSI between groups. The significance level of ΔPSI values is estimated from an empirical distribution of ΔPSI between replicates. Similar to rMATS, this method only supports the comparison of two groups at a time.

Challenges and limitations

As shown in the case of rMATS and SUPPA2, one of the main shortcomings of existing methods is that they only support two-group comparisons. Although there are a handful of methods that incorporate complex designs, they do so at the cost of over-simplifying the definition of AS events into simply quantifying the usage of exonic bins (DEXSeq and limma). This becomes a major limitation when it comes to analyzing tumour samples from large cohort studies, a context in which several external factors add to the complexity of the problem.

It is generally understood that when conducting cancer studies, additional variables such as patient demographics, experimental interventions, sample collection procedures, and tissue

composition, can hinder our ability to accurately associate changes between patient groups to a single factor (i.e., condition). For instance, it is known that tumour samples contain not only malignant cells but also normal epithelial cells, stromal, vascular, and immune cells [114]. This means that virtually any molecular readout obtained from tumour biopsies will reflect a mixture of malignant and non-malignant patterns. Variations in cellular composition of tumours, herein referred to as tumour purity, have been shown to strongly bias genomic analyses, including RNA-seq-based approaches [115-117]. The confounding effect of tumour purity on RNA-seq-based analyses has been extensively explored in the context of differential gene expression analysis [118-121].

Although the influence of confounding factors on RNA-seq analyses is well-established, it has not been systematically addressed in the context of alternative splicing quantification. Slaff et al. [102] introduce a method aimed at removing the effect of confounding factors from RNA-seq data. This method offers a convenient solution when the primary objective is the generation of a "corrected" RNA-seq dataset, without delving into the variability introduced by other experimental variables. However, in many instances, we are interested in disentangling the contributions of different experimental variables to the observed changes across sample groups (i.e., conditions such as tumour vs normal). It is essential to note that, to our knowledge, none of the existing methods has been able to achieve this level of specificity, particularly at the level of individual alternative splicing events.

1.5 Motivation

The previous sections highlight how dysregulated AS programs can have pervasive and causative roles in tumorigenesis. This dysregulation often involves changes in the relative abundance of oncogenic and tumour-suppressor splicing factors, which further promote the expression of pro-oncogenic isoforms. For example, some isoforms correspond to the embryonic version of the gene, enabling a reversion to a stem-like cell state, a phenomenon commonly observed in cancer [122]. Nonetheless, the specific contributions of individual splicing events to tumour initiation, growth, and metastasis remain largely elusive. Moreover, the mechanistic origins of most of these splicing aberrations are still poorly understood. Thus, the combination of the existing evidence and the numerous missing pieces underscores the pressing need for a

systematic characterization of the splicing programs that drive tumour development and progression.

While there have been significant methodological advancements in the study of AS, translating large-scale transcriptomic data from tumours into actionable splicing-based therapeutic strategies remains challenging. To put it differently, although several computational tools have identified nearly 2 million cancer-associated splicing events [80], only a handful of them turned into promising drug targets. This disparity is in part due to the "needle in a haystack" issue, stemming from several factors, one of them being the substantial confounding effects present in RNA-seq datasets [123]. While addressing these biases is now standard in gene expression analyses [124-127], the same attention has not yet been extended to AS, which also relies on RNA-seq data. This technical limitation has significantly hindered our understanding of the splicing programs exploited by cancer cells during tumour development and progression.

In summary, the field still lacks tools that can incorporate all the necessary components to (a) model independent AS mechanisms; and (b) separate differential AS rates accompanying oncogenesis from those explained by variations in cellular composition or biological differences between sexes and age groups. Here, we sought to develop a computational method that can address these issues and apply it to characterize the splicing landscape of tumours to ultimately unravel the splicing programs accompanying the transitions between tumour evolutionary stages.

1.6 Significance

This project explores the role of AS in driving tumour development and progression by leveraging robust statistical approaches to conduct systematic analysis of transcriptomes. Such knowledge has the potential to result in novel prognostic markers, understand the patient-level factors that contribute to tumour metastasis, and derive actionable targets along the regulatory cascade of AS events. In addition, our methodological innovation will strongly benefit the community since it addresses a previously unresolved challenge: dissecting the effect of experimental variables on different types of AS changes in datasets with complex experimental designs.

CHAPTER 2 MATERIALS AND METHODS

2.1 The TRex method

To enable the inclusion of sample-level covariates into the analysis of differential AS, we developed TRex, a novel computational method that leverages **TR**anscript abundances to derive **ex**on-centric quantifications of seven types of AS events (shown in **Figure 2**). We divided the problem into four main stages: (1) quantifying isoform abundances, (2) building a reference annotation of splicing events, (3) converting transcript quantifications into AS event counts, and (4) modelling the effect of experimental variables on AS changes observed between conditions.

The first two steps are performed using existing methods, since both tasks have been extensively addressed by others [128-130]. This modular design gives TRex the flexibility to analyze RNA-seq data from different technologies (short-read and long-read), that can be quantified using the tool that best suits each case. This also applies for the annotation of AS events, which can be generated from a reference annotation or derived from the RNA-seq data itself to detect novel events using external software. It is worth mentioning that even though in this study we limited our analyses to known transcripts, TRex is designed to support the analysis of novel events if provided in the event annotation.

2.1.1 Preparing TRex inputs

All the short-read RNAseq datasets used in this thesis were processed using Salmon (v.1.10.2) [128] in mapping-based mode using the GRCh38 genome reference. Salmon quantifications were aggregated into transcript and gene expression matrices using the R package *tximport* [131]. The annotation of all types of AS events were obtained using the *generateEvents* command from SUPPA2 v.2.3 [104] using the GRCh38.p13 transcript annotations from GENCODE v37 [132].

Once the transcript abundance matrix and the annotations were obtained, a custom script is used to aggregate both the counts and the abundance of all transcripts supporting each of the two possible outcomes of every event (shown in **Figure 2**). This procedure yields two abundance matrices, which we call A and C. Each abundance matrix is then corrected for the average length of all transcripts in each group (A or C) and then scaled back to the corresponding library size of

each sample. This procedure facilitates the downstream application of count-based statistical methods in an unbiased manner. The A and C matrices of estimated counts are then combined into a single augmented matrix that has two entries (rows) for every event, each containing one of the two count types (A or C), both measured in the same sample (columns) (described in more detail in 7.2.1).

2.1.2 TRex statistical model

For each event k in sample n, we assume that both A and C counts are sampled from a negative binomial distribution:

$$\mathrm{m}_{kn} \sim NB(\mu_{kn}, \sigma^2_{kn}) \tag{1}$$

$$\mathrm{m'}_{kn} \sim NB(\mu'_{kn}, \sigma'^2_{kn}) \tag{2}$$

In the above, m_{kn} is the observed count of outcome A of event k in sample n, and $\mathrm{m'}_{kn}$ is the observed count of outcome C of event k in sample n. Then we can approximate μ_{kn} and μ'_{kn}, the means of the distributions of m_{kn} and $\mathrm{m'}_{kn}$, as:

$$\mu_{kn} = \gamma_{kn} \times q_k \times s_n \tag{3}$$

$$\mu'_{kn} = \gamma'_{kn} \times q'_k \times s'_n \tag{4}$$

Here, γ_{kn} and γ'_{kn} are the sum of abundances of isoforms that belong to A and C counts, respectively, of event k in sample n. q_{kn} and q'_{kn} are event-specific scaling factors that are shared across samples, and s_n and s'_n are sample-specific scaling factors shared across events. Using such definitions, we can express the ratio of the abundances of A and C isoforms as a function of an experimental condition ρ:

$$r_{k,\rho(n)} = \frac{\gamma_{kn}}{\gamma'_{kn}} \tag{5}$$

Note that $r_{k,\rho(n)}$ can be interpreted as the odds of exon inclusion. In other words, $r_{k,\rho(n)} = \psi_{k,n}/(1-\psi_{k,n})$, where ψ is the percent-spliced-in (PSI).

We can rewrite Eqn. 4 as:

$$\mu'_{kn} = \gamma_{kn} \times r_{k,\rho(n)} \times q'_{kn} \times s_n \tag{6}$$

And then apply a log transformation to Eqn. 3 and 6 so that:

$$\log(\mu_{kn}) = \log(\gamma_{kn}) + \log(q_{kn}) + \log(s_n) \tag{7}$$

$$\log(\mu'_{kn}) = \log(\gamma_{kn}) + \log(r_{k,\rho(n)}) + \log(q'_{kn}) + \log(s_n) \tag{8}$$

Such formulation allows us to apply a generalized linear model (GLM) with a log-link function to estimate the unknown model parameters. TRex internally uses DESeq2 v.1.32.0 [133] for the model fit, since it was designed to deal with RNA-seq count data. The details of how the above model can be implemented as a design matrix in DESeq2 can be found in Supplementary Note 1.

2.2 Benchmarking TRex using simulated RNA-seq datasets

To benchmark the performance of TRex in a systematic manner we opted for using a series of simulation experiments, modeled after effect size distributions expected from disruption of a splicing factor. In order to ensure the presence of strong splicing changes, we selected a dataset from the ENCODE consortia [134] where the serine and arginine rich splicing factor 9 (SRSF9) was knocked down (KD) in K562 cells, using untreated K562 cells as the control group (C). The FASTQ files were downloaded from the ENCODE data portal (https://www.encodeproject.org) [135] using the following accessions: ENCSR972AZD and ENCSR341TTW. The data was processed as described in Section 2.1.1, and used as a template to generate the ground truth transcript abundances needed for simulation experiments, as described below.

We used the SRSF9 dataset as a basis to simulate an experiment that had five replicates in each group. The simulation consisted of the following steps:

1. **Fit two models around the observed mean** using the real data: one for the relationship between the mean expression of a transcript and its log2 fold change between conditions (mean-LFC model); and another one for the mean expression and the standard deviation of log2 fold changes across transcripts (mean-variance model).
2. **Simulate the ground truth mean of the abundance of each transcript in the control group** by taking the ranks of the mean observed TPM (transcripts per million) values to derive a cumulative distribution function (CDF). Then we use this CDF to sample the baseline logTPM values for the simulated control from a normal distribution parametrized by the observed mean and standard deviation of the SRSF9 control group.
3. **Generate the simulated ground truth difference between groups** KD and control (condition effect) by sampling from a Laplace distribution parametrized by the baseline logTPM and the coefficients of the mean-LFC model.
4. **Obtain the final simulated abundances** using the previous components. First, the simulated logTPM values are calculated by adding the observed mean logTPM of the control group (step 2) and the ground truth condition effect (step3). Then, noise is added to the simulated logTPM values of each replicate using a Laplace distribution parametrized by the coefficients of the mean-variance model (step 1). Finally, the log transformation is reverted to obtain transcript abundances as TPM values.
5. **Use the simulated abundances to generate FASTQ files** using the RSEM v.1.3.3 module *rsem-simulate-reads* [136]. These fastq files were used as input to rMATS v.4.1.1, SUPPA2 v.2.3 and TRex.

We repeated the above procedure 25 times to obtain 25 simulated datasets; the results reported correspond to the average of these simulations. Note that, since the ground truth TC effects were generated for transcripts, they had to be converted to event-level effects in order to obtain a ground truth $\Delta logitPSI$ and ΔPSI for each event, which are the metrics reported by various exon-centric splicing tools, including TRex. This was accomplished by calculating a ground truth PSI for each event from the ground truth transcript abundances (**Figure 4**), followed by calculation of $\Delta logitPSI$ and ΔPSI.

2.2.1 Modeling confounding effects

To model the effect of a technical confounder (TC) on the resulting RNA-seq reads, we modified the Step 3 of the standard simulation framework described in the previous section by also generating a simulated ground truth TC effect. We used a similar approach to what we did for the condition effect and sampled from a Laplace distribution parametrized by the simulated ground truth logTPM and the coefficients of the mean-LFC model but added a strength parameter ranging from 0 to 1, where 1 means that the TC effect is as strong as the difference between conditions. Then, to generate the final simulated abundances, we simply added the ground truth TC effects to the other components during Step 4.

In this case, we generated a total of 25 RNA-seq data sets by modulating the strength parameter from 0 to 1 with 0.25 increments. For every strength value, 5 random simulations were generated and then processed independently with TRex as described in Section 2.1.1. To convert the ground truth isoform abundances to $\Delta logitPSI$ values for each event and each variable, logitPSI was calculated for each event and each combination of variables, similar to the previous sections. Then, a simple linear model was fitted to logitPSI values, to obtain coefficients that corresponded to the association of each variable with the change in ground truth logitPSI.

2.2.2 Computing performance metrics

The previous steps yielded a series of RNA-seq datasets that were then analyzed using rMATS, SUPPA2 and TRex to estimate differential splicing changes between the KD condition and the control. In the case where no TC effects were added, we treated the performance evaluation like a classification problem. The nominal P-values of the predictions from each method were used as ranking score to calculate the area under the receiver-operator curve (AUROC) for distinguishing true differential AS events from background (reference classes). To define the reference classes, we used a combination of cut-offs on the ground truth ΔPSI (from 0.1 to 0.5) and $\Delta logitPSI$ (from 0.5 to 3). Events with effect sizes above both cut-offs were classified as *positive* (differentially alternatively spliced*),* and the rest as *negative.* Including both cut-offs was necessary to ensure a fair comparison across methods, given that rMATS and SUPPA2 measure effect sizes as the difference in PSI between groups, whereas TRex's effect sizes are measured as the difference in logitPSI.

In the analysis of TC effects, we assessed the performance of TRex by directly comparing the effect size estimates from TRex against the ground truth effects used to generate transcript abundances in Step 4.

2.3 TCGA analysis

2.3.1 Processing the TCGA dataset

Raw FASTQ files of 10,247 samples from 31 cancer types from the TCGA along with their corresponding metadata were downloaded using the Genomic Data Commons Automatic Programming Interface (https://gdc.cancer.gov/developers/gdc-application-programming-interface-api). Additional clinical metadata tables were downloaded using the R package *TCGAbiolinks* v.2.22.4 [137]. RNA-seq reads were quantified using Salmon v.1.10.2 [128] in mapping-based mode against the GRCh38 transcriptome. Salmon quantifications were aggregated into transcript and gene expression matrices using the R package *tximport* v.1.20.0 [131].

After quantification, further inspection of the data was performed to detect tumour outliers. First, a Principal Component Analysis (PCA) was performed on the gene expression matrix of each cancer type separately. Using the first three principal components, a local outlier factor (LOF) was calculated for every sample. After applying threshold on this metric, a total of 145 tumour outliers were removed in downstream analysis. This procedure was done using the R package *bigutilsr* v.0.3.4 [138]. Further processing of this dataset follows the same procedure described in the section *TRex inputs*.

2.3.2 Estimating tumour purity and impurity

The preferred approach to estimate tumour purities relies on DNA copy number alterations measured using whole-genome sequencing (WGS) data [139]. However, not all the samples in the TCGA have both RNA-seq and WGS data available. Thus, in order to include a larger number of samples in the analysis, we opted for estimating tumour purity from RNA-seq data using the R package *Estimation of STromal and Immune cells in MAlignant Tumours using Expression data* (ESTIMATE) v.1.0.13 with a modified formula [140]. Since many of the TCGA RNA-seq samples were not available at the time this method was developed, directly applying their empirical formula to convert ESTIMATE scores to tumour purity yielded values outside of the 0-1 range

(**Figure S1**). Thus, we recalibrated the parameters of the empirical formula $\cos(\alpha + \beta \times ESTIMATEscore)$ to accommodate the new samples. We implemented an optimization algorithm to find the values of α and β that minimized the squared error between predicted purity and DNA-based purity from ABSOLUTE. This procedure resulted in the following formula:

$$purity = \cos(4.97499329410793 + 5.34321675550045\mathrm{e}^{-5} \times \mathrm{ESTIMATEscore})$$

Tumour purities derived from the recalibrated formula showed a better correlation with ABSOLUTE purities than what was previously reported (**Figure S1**).

For modeling purposes, we defined tumour "impurity" as 1–purity, so that we could assign an impurity of 0 to normal samples. Even though this assumes that normal tissues contain only normal epithelial cells and no immune or stromal cells, it vastly facilitates the downstream interpretation of the results. We are aware that this assumption might not hold in all samples, as we have further discussed in later sections.

2.3.3 Fitting models of differential AS

Transcript counts were aggregated into event counts as described in the methods section 2.1.1. We then used TRex to fit the models on the event counts. A series of models were fitted to each cancer type and each AS event type separately using the DESeq2 framework. Prior to fitting the models, the following pre-filtering steps were applied to discard: (a) cancer types with less than 5 samples in each group of interest depending on the analysis (condition or stage), (b) samples where impurity could not be inferred by ESTIMATE, (c) tumour samples flagged as outliers, (d) samples in which the number of events with exactly zero counts exceeded the 95th percentile of all samples, and (e) events that had less than 10 counts in more than the number of normal samples. A final shrinkage step was performed on all model coefficients used for downstream analysis using DESeq2's *normal* estimator.

Supplementary Data Table 1 contains all the information regarding the number of filtered samples and events. Following these quality control steps, we fitted a series of models in each cancer type separately using the following designs:

Table 2 Models fitted in the TCGA dataset		
Analysis	***Model type***	***Design***
tumour	simple	~ condition
	complex	~ condition + impurity + sex + age
stage	simple	~ stage
	complex	~ stage + impurity + sex + age

Here, *condition* is a binary variable denoting tumour and normal status, where normal is the reference group. *Stage* is represented as an integer from 0 to 4. *Age* is reported in years. *Sex* is a binary variable indicating male or female, where male is the reference group. *Impurity* is a continuous variable, described in more detail in the previous section. Functional analysis of AS events of interest.

2.4 Functional analysis of AS events of interest

2.4.1 Identifying potential events of interest

In downstream analyses, we discarderd events with $\Delta logit$PSIs that were unlikely to be of biological relevance based on their per-group (tumour or normal) PSI values. For instance, a tumour PSI of 0.9888 and a normal PSI of 0.9900 would result in a $\Delta logit$PSI of 0.11. If the tumour PSI where to be 0.385 and the normal 0.45, we would also get the same $\Delta logit$PSI. However, a change of 0.11 is more likely to be biologically meaningful in the later case, as it translates to a larger ΔPSI. Hence, we implemented the following filters on the per-sample PSIs to remove events where: (a) the mean event PSI in normal and in tumour were both less than 0.01 or greater than 0.99, (b) the per-sample PSIs fell under the previous thresholds in more than 95% of the samples in each group (tumour or normal), or (c) the event was not measured (NA value) in more than 95% of the samples. In total, the discarded events amounted to ~ 20% of the events measured across canccer types (**Figure S5** and **Figure S6**). The exact number of events filtered per event type and cancer type in each model are shown in **Supplementary Data Table 1**.

2.4.2 Gene overrepresentation analysis

Gene overrepresentation analyses (GORA) were conducted using the R package *fgsea* (v.1.18.0) [141]. Hallmark pathway gene sets were retrieved using the R package *msigdb* (v.7.5.1) [142]. *CellMarker* 2021 [143] gene sets were downloaded from *Enrichr* library [144] [145]. Genes selected for testing in each analysis group (i.e., cancer and/or event type) were those that had at least one event significantly associated with the variable of interest (adjusted P value <0.05) with an absolute $\Delta logitPSI$ greater or equal to one (unless specified otherwise), which we will hereafter call differentially spliced genes.

2.4.3 Single cell pathway activity analysis

Pathway activities were estimated using GEDI [146]. The gene sets for this analysis comprised of the overlap between the differentially spliced genes and the significant pathways in the overrepresentation analysis above. The remaining pathway genes were included in the gene set matrix (C matrix) for GEDI with a weight of zero, so that they could still contribute to learning the low dimensional representations.

2.5 Analysis of AS regulators

2.5.1 Selecting representative RBP binding motifs

We obtained 175 position frequency matrices (PFMs) corresponding to experimentally derived binding motifs of 154 RBPs from CisBP-RNA [17]. However, many of these motifs are highly similar to each other, as they either correspond to motifs obtained for the same RBP in different experiments, or to motifs of highly similar homologous RBPs (**Figure S17**). To remove this redundancy and find a set of representative RBP binding motifs, we computed the pairwise similarities of all the motifs using MoSBAT [147]. Then, we applied K-means clustering on the similarity matrix and selected the best value for k ($k=70$) based on average silhouette values (**Figure S17**). In each cluster, we selected as representative motif the one closest to its centroid in Euclidean space. This procedure resulted in a set of 70 motifs, hereafter called the set of representative motifs.

2.5.2 Fetching sequences around splice junctions

Every event type is defined by a unique set of genomic coordinates that indicate the position of all the splice junctions involved in it (detailed in [104]). To obtain the sequence surrounding every position of all annotated events, hereafter called *splice regions*, we selected 50, 100 or 200 nucleotides upstream or downstream of each splice junction. Each set of sequences was treated independently, and the optimal window size was selected at the model fitting stage (described below). The resulting ranges were processed using the R package *GenomicRanges* (v.1.44.0) [148], and then used to fetch the corresponding DNA sequences using the R package *BSgenome.Hsapiens.UCSC.hg38* (v.1.4.3) [149]. Events in the reverse strand were converted to their resulting 5' to 3' mRNA sequence using the function *reverseComplement* from the R package *Biostrings* (v.2.60.2) [150]. This procedure yielded a total of 2,721,368 sequences per window size.

2.5.3 Determining RBP binding near splice junctions

To determine whether a given RBP could potentially bind to each sequence, we calculated its affinity using AffiMx [147]. Sequences were scanned in the 5' to 3' direction. All the remaining parameters were kept as default. The resulting motif scores corresponding to each scanned sequence were then used as a proxy for RBP binding affinity to such region.

2.5.4 Associating RBP binding with differential alternative splicing

For all events of the same type, we modeled the tumour-associated $\Delta logit(PSI)$ values as a function of the binding affinities of representative RBP motifs near splice junctions. We tested the upstream and downstream regions of all splice junctions in windows of 50, 100 and 200 nucleotides (splice regions). The number of sequences scanned varied depending on the number of splice coordinates needed to identify the event. For instance, SE events are defined with 4 coordinates denoting the start (S) and end (E) positions of the exons involved: E1, S2, E2, and S3. Description of the splice sites of all events can be found in [104].

We fit a series of linear models for every cancer, event, and window size, separately, using Ordinary Least Squares regression. Each model is of the form $y = \mathbf{A}\beta + \varepsilon$, where y is the vector of $\Delta logitPSI$ values between tumour and normal of all events; the matrix $\mathbf{A}$ contains the affinity of all 70 representative motifs over all the corresponding splice regions of each event in y; ε is the

vector of residuals; and β is the vector of coefficients corresponding to each RBP and splice region in **A**. The β coefficients are hereafter called the differential splicing regulation (DSR) coefficients.

Model performance was assessed using 10-fold cross validation from the R package *caret* (6.0-93) [151-153]. The optimal window size (n=200) was selected as it yielded the highest Pearson correlation in the largest number of cancer and event types (based on in 10-fold cross validation; **Figure S18**).

2.5.5 Finding RBP target genes

A procedure similar to that proposed by Ray et al. [17] was followed to assign target genes to each RBP based on its binding affinity at a given splice region. An important difference is that here we have multiple events of the same type within the same gene, as opposed to a single measurement per gene region. Thus, to obtain a gene-level measurement per gene and splice region, we calculated the mean affinity of each representative motif over all events of the same type within the same gene. Then we calculate the Z score of each motif's binding affinity at each splice region over all annotated splicing events. The Z scores were normally distributed; hence we calculated their corresponding P value under a normal distribution and then performed an FDR correction for multiple testing. Targets were selected per splice region using and FDR cut off of 0.2. This procedure was followed for each event type separately.

2.5.6 Detecting RBPs associated with tumour development

The DSR coefficients of representative motifs were compared against the differential expression of their corresponding RBPs (including RBPs associated with other motifs represented by the motif that was included in our model). In cases were the same RBP was associated with more than one representative motif (e.g., when dissimilar motifs were obtained in different experiments), the analysis was conducted for each representative motif separately. We measured the log2 fold change in expression of 148 RBPs in tumours with respect to normal samples using DEseq2. The model accounted for the confounding effects of impurity, sex, and age. We used Pearson correlation to measure the relationship between each RBPs differential expression and DSR over a given splice region across cancer types. The resulting P values were corrected for multiple testing in each event type using FDR adjustment.

2.6 Single-cell RNAseq of clear-cell Renal Cell Carcinoma tumours

This dataset is part of an ongoing project in our group and is not publicly available, it was included with permission from the authors [154].

2.7 Software, code, and data availability

All the codes required to reproduce the results in this thesis are available at the GitHub companion repository https://github.com/csglab/hgthesis_lms.git). Raw event quantifications and TRex predictions of all datasets used in this thesis, along with supplementary data tables were uploaded to various Zenodo collections. The individual DOIs can also be found in the companion GitHub. Detailed versions of additional software used in this thesis are included in **Table S1**.

CHAPTER 3 RESULTS

3.1 Study overview

To characterize the AS programs associated with the development and progression of tumours, we analyzed RNA-seq data from 31 cancer types from The Cancer Genome Atlas (TCGA) [14]. **Figure 5a** shows the number of tumour and normal samples included in the analysis, as well as the number of tumour samples per cancer stage across cancer types (exact numbers available in **Supplementary Data Table 1**). Furhter details regarding the pre-processing of this dataset can be found in Chapter 2. In total, we measured 206,928 events resulting from seven types of exon centric AS (**Figure 5b**) across 10,247 samples.

As it was thouroughly discussed in the first chapter, patient samples from large cohort studies can display substantial variability due to various measured and unmeasured sources of heterogeneity. Here, we aim to explore and model biases introduced by tumour purity (the proportion of malignant cells), the patients' reported sex, and their age in years. We found tumour purity especially intriguing because its impact on AS quantification had not been previously investigated. Moreover, from a technical standpoint, it was reasonable to assume that conclusions drawn from bulk RNA-seq measuments may not accurately reflect the tumour's biology due to non-malignant cells present in tumour samples (**Figure 5c**).

3.2 Tumour purity has a strong cancer-specific effect on the quantification of alternative splicing

To systematically assess the impact of tumour purity on the downstream analysis of RNA-seq-data, we contrasted its effect on the quantification of classical gene expression and AS events in the same datasets. Tumour purity was measured using ESTIMATE [140] with a recalibrated empirical formula (see Chapter 2) to calculate purity in samples outside the range of the original dataset (Supplementary **Figure S1**). These values were then transformed into *impurity* (fraction of non-malignant cells), which is simply calculated as $1 - purity,$ to simplify the interpretation of downstream modelling steps.

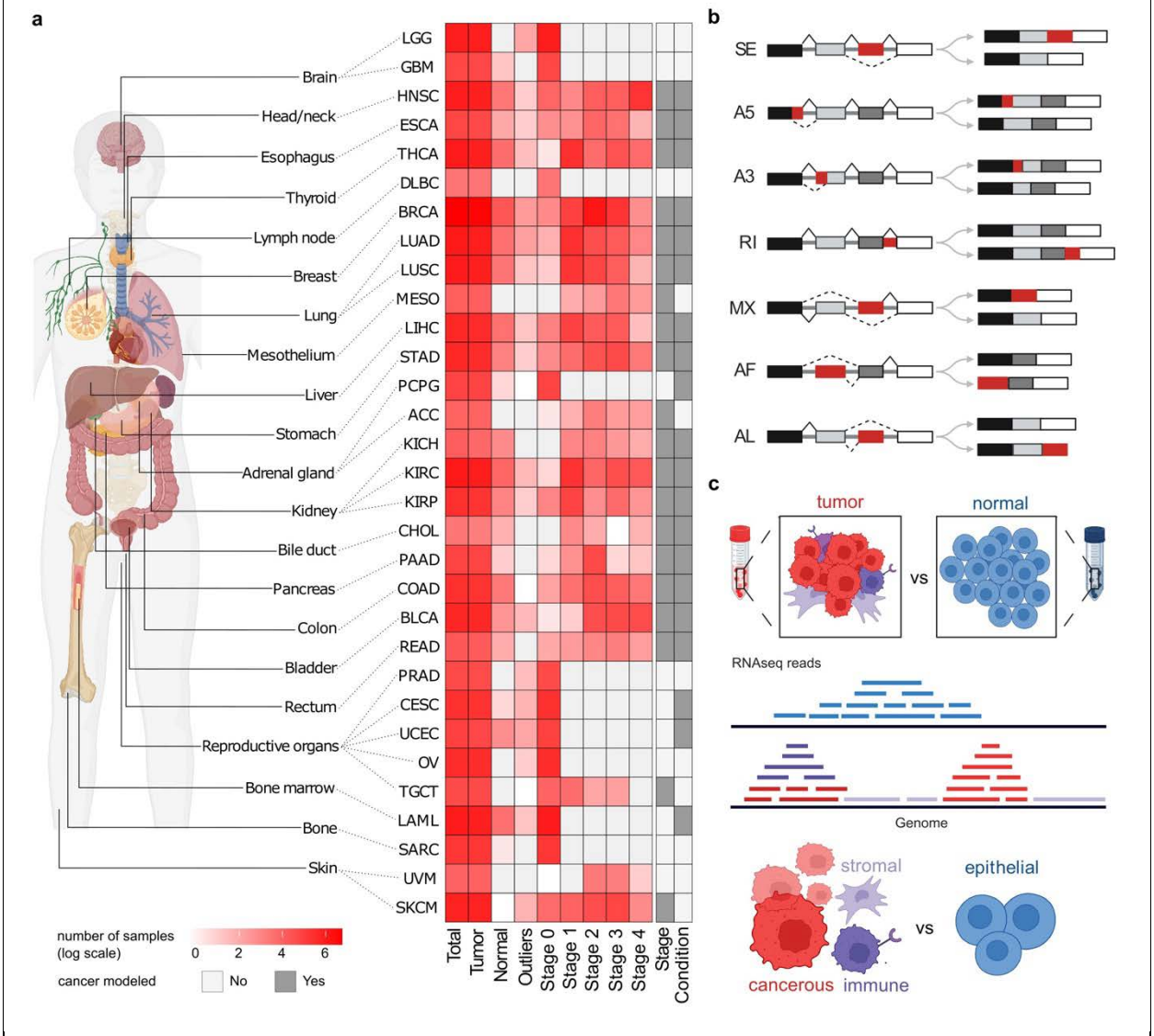


Figure 5 Pan-cancer study overview

(a) Summary of TCGA samples included in the study, ordered by the tissue of origin. 'Total' = total number of samples; 'Tumour' = number of tumour samples; 'Normal' = number of normal samples; 'Outliers' = number of tumour outlier samples detected by a local outlier factor (see Section 2.3). All 'Stage N' columns indicate the number of tumour samples of each corresponding stage. The number of cases is displayed in logarithmic scale. The last two columns indicate if the cancer type was included (dark grey) in the stage or tumour models, respectively. **(b)** diagram of exon-centric alternative splicing events, abbreviations described in the main text; **(c)** technical rationale behind the need to address tumour purity. The diagram

schematically shows how reads coming from non-malignant cells are found in tumour biopsies. As a result, a tumour vs normal contrast becomes a comparison between normal (epithelial) cells vs a mixture of malignant and immune/stromal cells. LGG = low-grade glioma; GBM = Glioblastoma multiforme; HNSC = Head and Neck squamous cell carcinoma; ESCA = Esophageal carcinoma; THCA = Thyroid carcinoma; DLBC = Lymphoid Neoplasm Diffuse Large B-cell Lymphoma; BRCA = Breast invasive carcinoma; LUAD = Lung adenocarcinoma; LUSC = Lung squamous cell carcinoma; MESO = Mesothelioma; LIHC = Liver hepatocellular carcinoma; STAD = Stomach adenocarcinoma; PCPG = Pheochromocytoma and Paraganglioma; ACC = Adrenocortical carcinoma; KICH = Kidney Chromophobe; KIRC = Kidney renal clear cell carcinoma; KIRP = Kidney renal papillary cell carcinoma; CHOL = Cholangiocarcinoma; PAAD = Pancreatic adenocarcinoma; COAD = Colon adenocarcinoma; BLCA = Bladder Urothelial Carcinoma; READ = Rectum adenocarcinoma; PRAD = Prostate adenocarcinoma; CESC = Cervical squamous cell carcinoma and endocervical adenocarcinoma; UCEC = Uterine Corpus Endometrial Carcinoma; OV = Ovarian serous cystadenocarcinoma; TGCT = Testicular Germ Cell Tumours; LAML = Acute Myeloid Leukemia; SARC = Sarcoma; UVM = Uveal Melanoma; and SKCM = Skin Cutaneous Melanoma.

To obtain transcript quantifications of each cancer type, we used Salmon in mapping-based mode. Transcript abundances were then used to generate a gene expression matrix of transcripts per million (TPM) and a splicing event matrix of PSI values (as described in **Figure 4**). Samples from all cancer types were aggregated into pan-cancer gene expression and event PSI matrices, hereafter referred to as *expression* and *splicing* matrices, (see Chapter 2). For each pan-cancer matrix, we performed two-dimensional t-Stochastic Neighbor Embedding (tSNE) using its first 20 principal components (PCs) (**Figure 6**, see Chapter 2 for detailed procedure).

When all cancer types were analyzed simultaneously, for both splicing (**Figure 6a**) and expression (**Figure 6b**), the largest source of variation captured by the tSNE embeddings is driven by the tissue of origin in most, but not all, cancer types. Nonetheless, we still observe a large correlation with impurity in the top ten PCs in both analyses (**Figure S2**), e.g., PC8 from the splicing matrix ($R = 0.67$, $p < 2e\text{-}16$) and PC6 from its expression counterpart ($R = -0.72$; $p < 2e\text{-}16$), suggesting that, despite the visual grouping by cancer type in the tSNE embeddings, impurity

is still captured by the top 10 axes of variation in both cases. The contribution of age and sex was observed at a local scale, driving differences within cancer types only (**Figure S3**).

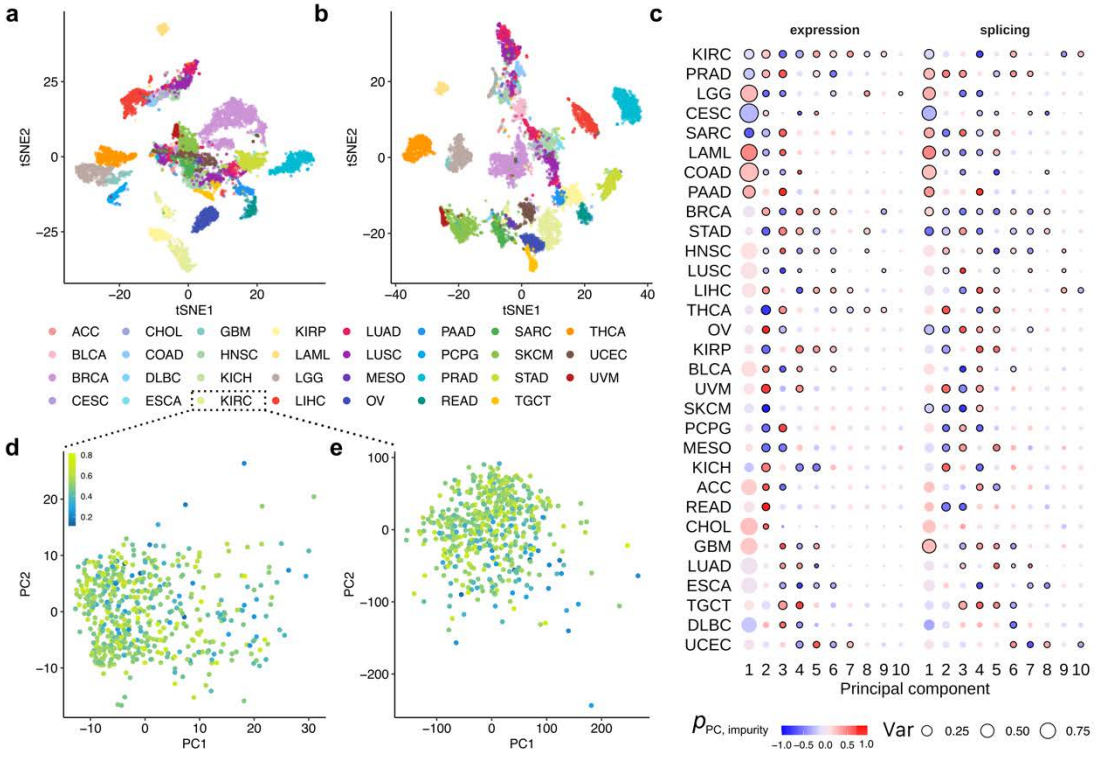


Figure 6 Impact of tumour purity on downstream RNA-seq analyses

Pan-cancer tSNE embedding on **(a)** splicing matrix of PSI values and **(b)** gene expression matrix. **(c)** Matrix of correlations between impurity and PC scores derived from tumour samples of each cancer type and analysis (expression and splicing). The size of the points reflects the proportion of variance explained by the corresponding principal component, color shows the strength and direction of the correlation with impurity, and solid black outlines indicate a significant correlation with a cancer-wise Bonferroni-adjusted P value < 0.05. **(d)** Splicing-based and **(e)** expression-based PCA embeddings of tumour samples of KIRC. Cancer abbreviations remain the same as **Figure 5**.

To better assess the impact of impurity in the absence of large tissue differences, we calculated a series of pairwise correlations between impurity and each of the top 10 PCs from both expression and splicing in each cancer type. **Figure 6c** shows that tumour impurity impacts the

quantification of splicing as much as it impacts gene expression quantification. We found that the contribution of impurity to the variance across samples was not consistent across cancers. For example, several cancer types show extreme patterns such as clear cell renal cell carcinoma (KIRC), low-grade glioma (LGG), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) where the impurity-correlated PCs explain almost 75% of their variance harbor strong correlations with impurity. To demonstrate this visually, we examined the first two PCs of KIRC from both splicing-based (**Figure 6d**) and expression-based (**Figure 6e**) analyses. In both cases, we observed a clear gradient of impurity on the PCA embedding of both splicing and expression measurements.

Interestingly, cancer types where a large proportion of their variance is explained by impurity-correlated PCs are not necessarily the most impure ones. In fact, there is no significant association between the proportion of variance explained by impurity PCs from either expression or splicing and the median impurity across cancer types (**Figure S2**). In summary, our findings suggest that event-level PSIs derived from RNAseq data are influenced by tumor impurity. Moreover, this bias is often comparable in strength to that observed in classical gene expression measurements.

3.3 TRex outperforms state-of-the-art and removes confounders of differential alternative splicing

As shown in the previous section, alternative splicing quantification is biased by tumour purity (and potentially other factors). Motivated by the need to incorporate such confounding factors in exon-centric analysis of AS, we developed TRex. This novel computational method can model the effect of multiple experimental variables on seven types of exon centric AS events. As described in Chapter 2, TRex models the read counts observed for opposing splicing outcomes as a function of experimental variables using a negative binomial generalized linear model. This approach effectively results in a contrast between each variable of interest and the abundance ratio of one outcome (e.g., exon inclusion) relative to the other (e.g., exon exclusion).

To systematically assess the performance of TRex, we compared it against two state-of-the-art tools, rMATS and SUPPA2, using simulations. As described in Chapter 2, we used a generative model to simulate RNA-seq datasets with known ground truth transcript abundances across conditions (knockdown and control) while incorporating multiple sources of variation at

different stages of the process. The resulting datasets were then processed with the default parametrization of each tool (see Chapter 2 for further details).

In our datasets, the log-fold changes in isoform abundances are simulated from a continuous distribution of values. Thus, by extension, the change in PSI of splicing events follows a continuous distribution. Therefore, to measure each method's true positive (TP) and false positive (FP) rates, a set of ground truth differentially alternatively spliced (DAS) events need to be defined based on some threshold. It is common practice to define DAS events by setting an arbitrary threshold on their PSI and its associated metric of statistical significance. However, this approach is not suitable in our case because TRex derives a different effect size estimator ($\Delta logit$PSI) than SUPPA2 and rMATS (ΔPSI). Defining the ground truth based on a single metric, either ΔPSI or $\Delta logit$PSI, would strongly bias the performance to either method(s). Instead, we decided to use a series of thresholds on both ΔPSI and $\Delta logit$PSI to generate an array of ground truth DAS events. Then, we determined each method's performance as the area under the receiver operator curve (AUROC) for the classification of true DAS events defined based on each pair of ΔPSI and $\Delta logit$PSI thresholds. This procedure is explained in more detail in Chapter 2.

To determine the performance of TRex where no confounding effects were present, we ran 25 simulations in which the ground truth reflects PSI differences between two conditions, as well as random noise, in the absence of any confounding variables. The differences in performance between TRex and rMATS or between TRex and SUPPA2 are shown in **Figure 7c-d**. TRex demonstrated superior or equivalent performance in virtually all settings, with a global average increase in AUROC of 0.2 across methods and cut-offs. The performance metrics per method along the grid of cut-offs is available in **Figure S4**. Altogether, these results demonstrate that TRex outperforms the current state-of-the-art methods, even when there are no confounding factors are present in the data.

To determine whether TRex could successfully separate two distinct sources of variation—experimental condition and a confounding factor—we added a confounder of increasing strength to the simulations (see Chapter 2). The resulting *confounder effect* was parametrized in a way that a strength of 1 would result in differences equivalent to those found between conditions. First, we determined the agreement between the ground truth and predicted condition effects upon addition this confounder. TRex had the highest correlation between the true condition effects and its mean predicted effects across all strengths of the confounder (**Figure 7e** and **Figure S4**).

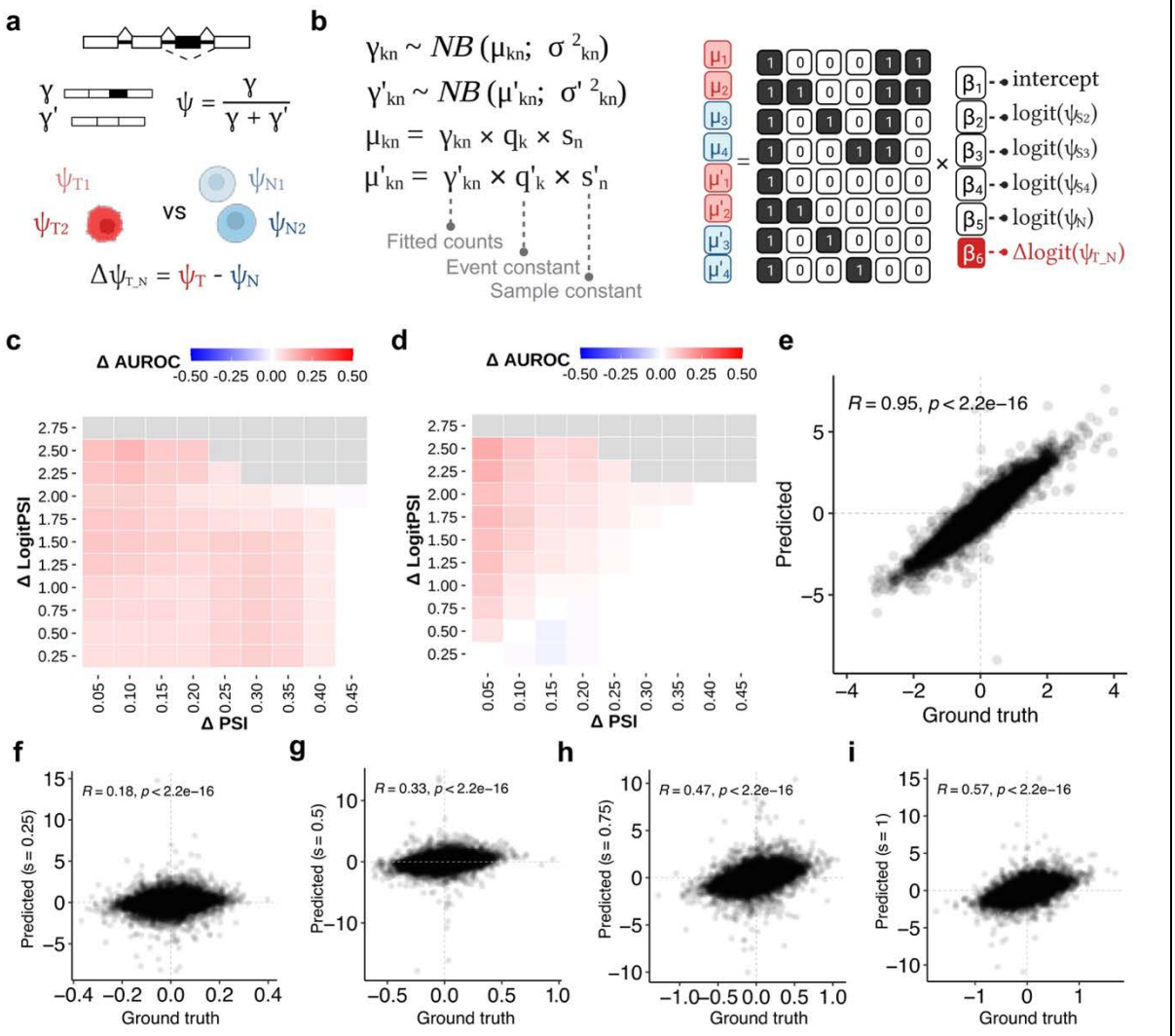


Figure 7 Benchmarking TRex in simulation experiments

(a) Classical calculation of PSI values illustrated in terms of SE events. **(b)** Overview of the formulation behind the TRex method; **(c)** Grids of differential performance in terms of AUROC between TRex and rMATS **(d)** and TRex and SUPPA2 (right) for varying PSI (x-axis) and logitPSI (y-axis) thresholds used to define the ground truth. **(e)** Comparison of ground truth vs TRex-predicted condition effects in the presence of simulated batch effects. (**f-i**) Ground truth vs TRex-predicted batch effects at a batch effect strength of **(f)** 0.25, **(g)** 0.5, **(h)** 0.75, and **(i)** 1. T = tumour, N = normal.

Similarly, we then compared the predicted confounder effects against the ground truth introduced to the simulated data at each strength (**Figure 7f-i**). TRex's capacity to detect real confounding effects markedly improved as their intensity increased. This aligns with our expectations, as the simulations also contained random noise, making it more challenging for the model to distinguish milder confounder effects from stochastic fluctuations. Overall, these findings confirm TRex's ability to disentangle unwanted sources of variation from differential AS changes linked to experimental groups of interest.

3.4 A pan-cancer map of unbiased AS programs reveals context-dependant and global patterns of dysregulation in cancer

To determine the AS changes that accompany tumour development, we contrasted tumour and normal samples from a total of 24 cancer types from the TCGA in two types of models, while accounting for the confounding effect of impurity, age and sex. **Figure 8** summarizes the set of DAS events that we identified in association with tumours and stage progression (detailed in Section 2.4.1, **Figure S5** and **Figure S6**). **Figure 8a** shows the proportion of events of each type that are significantly associated (FDR < 0.05) with each variable in the tumour and stage models. This high level summary revealed more DAS events between tumour and normal than between low and high stage tumours (**Figure 8a**). We also observe comparable numbers of tumour-associated and impurity-associated DAS events. This was not the case for age and sex-associated effects. Hence, we consider impurity to be the confounding variable with the largest impact on the observed AS differences between tumour and normal samples.

3.4.1 Exploring the individual effects of different variables

To further dissect the contribution of individual factors, we focused on cassette-exon skipping events. We chose this event given its clinical relevance observed across cancer types by our group and others [11, 18, 75]. In line with previous observations, we observed more than 13,000 events DAS in tumour vs normal (also called *tumour-associated* events) (**Figure 8b**) and ~8,000 associated with stage (referred to as *stage-associated* events) (**Figure 8c**) in at least one cancer type. A hierarchical clustering of tumor-associated SE events indicate that these programs often reflect the tumour's tissue of origin (**Figure 8b**). In contrast, a similar analysis of stage-associated

SE events suggest that these do not capture tissue-specific patterns (**Figure 8c**). Similar observations can be drawn from the remaining event types (**Figure S9** and **Figure S10**).

A gene overrepresentation analysis (GORA) on the Cancer Hallmark pathways [155] from MSigDB [156] revealed numerous pathways enriched in the genes with DAS events in tumours compared to normal ($|\Delta logitPSI| > 1|$; FDR < 0.05) and across tumor stages ($||\Delta logitPSI| > 0.05|$; FDR < 0.05) (**Figure S7**). The top three most frequently enriched pathways across cancer types were *Mitotic spindle*, *Myogenesis,* and *UV response*. In addition, we found 18 more Hallmarks overrepresented in at least three cancer types for SE events (**Figure 8d**) and several more in other event types (**Figure S7**). Thus, our findings suggest the existance of splicing-mediated dysregulation of genes in Hallmark pathways accompanying tumor development and progression.

To further validate the separation of effects driven by each experimental variable, we examined the set of impurity-associated changes resulting from both models (**Figure S8**). When we applied hierarchical clustering to the impurity-associated changes of SE events, we no longer captured the tumors' tissue of origin (**Figure S8a**), indicating that the impurity-associated effects are likely capturing changes driven by immune and stromal populations, which are distributed across different tissues. We then tested for the enrichment of gene signatures of immune and stromal cells in genes with impurity-associated events resulting from both models (tumor and stage). The analysis based on events from the tumour models (**Figure S8a-c**) indicated a general enrichment of several immune and stromal signatures across cancer types in impurity-associated genes (**Figure S8b**). As a negative control, we repeated the analysis using genes with tumor-associated events; in line with our expectations, most of the enrichment signals of immune and stromal gene signatures dissipated but did not disappear completely in all cancer types (**Figure S8b)**. One explanation for this could be that in some cancer types the normal samples vary substantially in cell type composition, which is currently not considered in our model. In fact, when we performed the analogous comparison with the results from stage-models, in which the variability of cell composition is modeled in both comparison groups of interest (high- vs. low-stage), the immune and stromal signals (**Figure S8d**) were nearly abolished in 19 out of 20 cancer types (**Figure S8e**). Altogether, these results suggest that TRex successfully separates the effects of confounders from the estimation of tumour- and stage-associated AS changes for SE events. Similar findings were observed in other event types (**Figure S11** and **Figure S12**).

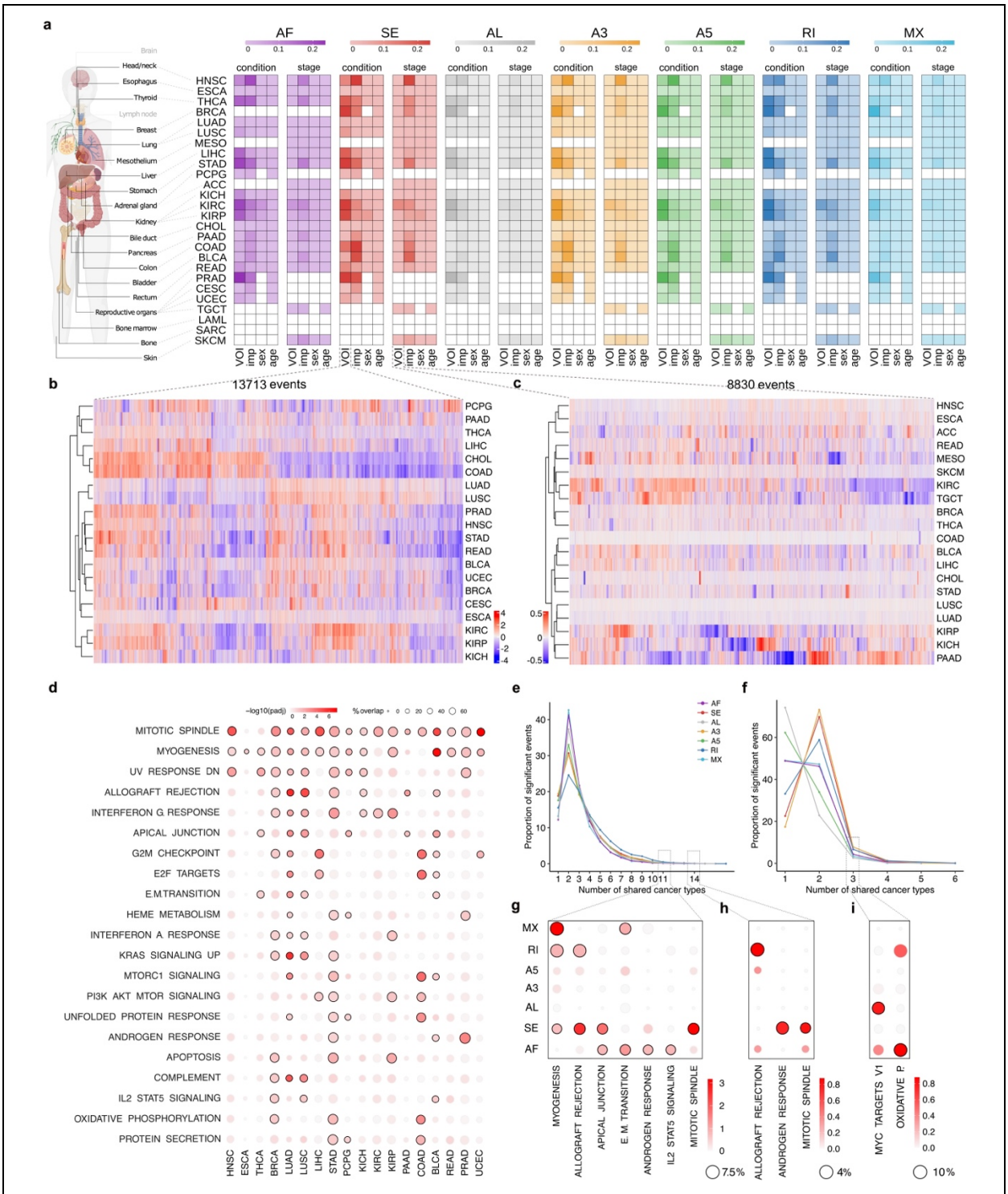


Figure 8 Pan-cancer view of AS events associated with tumour development and progression

(a) Proportion of events significantly associated with each variable at an adjusted P-value of 0.05; **(b)** heatmap of tumour-associated $\Delta logit$PSI values for SE events significant (adjusted p-value <0.05) in at least one cancer type (n=13,713); **(c)** heatmap of stage-associated $\Delta logit$PSI values for SE events significant (adjusted p-value <0.05) in at least one cancer type (n=8,830). **(d)** Overrepresented cancer Hallmark pathways in the set of genes with at least one SE events significantly associated with tumours ($|\Delta logit\text{PSI}|>1$; adjusted p-value<0.05). Proportion of significant events shared per number of cancer types in association with **(e)** tumour and **(f)** stage. Cancer Hallmark pathways overrepresented in the sets of genes with events significantly associated with **(g)** tumours in 11 cancer types, **(h)** tumours in 14 cancer types, and **(i)** stage in 3 cancer types. Heatmaps were clustered using the Ward method with a Pearson correlation distance metric. Solid black circles in (d), (g), (h), and (i) show significant pathways at an adjusted P. value < 0.2. Circle sizes reflect the number of overlapping genes. VOI = variable of interest.

3.4.2 Global cancer-associated AS programs

To identify pan-cancer splicing programs linked with tumour development and progression in multiple tissues, we looked for general trends in terms of AS mechanisms and cellular pathways. First, we defined *shared* events as those that were significantly associated with tumours (**Figure 8e**) or with stage (**Figure 8f**) in increasing number of cancer types. To our surprise, tumour-associated RI events showed a distinctive trend (**Figure 8e**). We observe a larger fraction of RI events that are shared by more cancer types than any of the other AS mechanism after the group size reaches n=3. These results suggest a potential role for RI as a pan-cancer mechanism of splicing dysregulation. However, stage-associated events of all AS types were shared by a maximum of three cancer types only, suggesting that AS programs associated with disease progression tend to be more cancer-specific than AS programs associated with tumour development. However, we cannot rule out the possibility that this difference reflects the reduction in sample sizes in stage models.

Motivated by the idea of finding AS programs frequently dysregulated across multiple cancer types, we focused on AS events with significant tumor associations in varying numbers of cancers. First, for each event, we determined the number of cancer types where it had an FDR<0.05

for the association with tumours. We observed that genes with SE events shared by at least 11 cancer types had a strong enrichment of the *mitotic spindle* and *myogenesis* Hallmark pathways (**Figure 8g**), consistent with our observations at the level of individual cancer types, where both pathways were significantly enriched in DAS SE events in 19 out of 20 cancer types. When we pushed this analysis to the largest number of shared cancer types (n=14), only three pathways remained significantly enriched, including the *allograft rejection* (**Figure 8h**). Similarly, the stage associated events shared in three cancer types revealed an enrichment of cell proliferation pathways (**Figure 8i**). In summary, our findings support the contribution of splicing-mediated dysregulation of genes involved in Hallmark pathways at a pan-cancer scale.

3.5 TRex detects neoplastic splicing programs in a case study of clear-cell Renal Cell Carcinoma

To further validate the ability of TRex in isolating the confounding effect of cell composition from AS dysregulation in neoplastic cells, we conducted a series of targeted analysis using kidney clear cell renal cell carcinoma (KIRC) as a case study. This model was of interest because of the large proportion of variance likely explained by impurity, and its extensive dysregulation of AS in tumours, as shown by the proportion of events significantly associated with tumours (**Figure 8a**). Nonetheless, we anticipate this model to be reflective of the performance of TRex in other cancer types.

3.5.1 Understanding the impact of different variables to the observed AS differences

To assess the influence of confounding factors on the estimation of DAS of SE events, we compared the coefficients obtained from two models: a *simple* one that explains DAS effects solely as a function of the experimental group (tumour or normal); and a *complex* model that accounts for impurity, sex, and age (see Chapter 2). By contrasting the tumor-associated effects resulting from both models, we classified events in four categories—cancer, cancer confounded, cancer flipped and potential false positives—that reflect the possible outcomes for events affected by confounders (**Figure 9a**). The *cancer* group consisted of the events that remain significant in both models (FDR<0.05). The *cancer confounded* category includes events whose association with tumours is significant only when accounting for confounding factors. The *cancer flipped* group captures cases in which the direction of the effect size changes upon the removal of confounders.

Finally, the *potential false positive* events are those with significant associations only when confounders are not considered. In addition, we defined a set of *impurity* events by selecting the ones significantly associated with impurity at an FDR<0.05. We did not group events based on age or sex, since the effects associated with those two confounders were less prominent compared to those of impurity (**Figure S15**).

To visually demonstrate the differences between tumor-associated DAS events vs those linked to impurity, we directly estimated the mean logitPSI values in normal samples and tumours stratified into five groups ranging from low to high impurity (**Figure S15**). As an example, **Figure 9b-c** shows two distinct SE events within the CD46 gene that belong to two separate categories. The *cancer* event pertains a small exon (~45 bp) in the middle of the gene, whereas the *impurity* event occurs near the 3' UTR of the gene. As the event category suggests, the mean $logit$PSI of the cancer event was clearly different from normal in all impurity bins (**Figure 9b**). In contrast, the mean $logit$PSI of the *impurity* event increased from low- to high-impurity groups (**Figure 9b**), with mid-to-high impurity tumours appearing different compared to normal. However, the difference observed between low impurity tumours and normal is negligible. This pattern is also evident from the sashimi plots of raw junction reads spanning the regions of each event in normal samples and tumour bins (**Figure 9c**), showcasing a scenario in which the wrong conclusion would be reached if impurity had not been accounted for.

3.5.2 Molecular pathways affected upon incorporation of confounders

To gain insights into the molecular pathways whose inferred differential splicing is affected by the addition of confounding factors to the model, we performed a gene overrepresentation of cancer Hallmark pathways across gene categories (**Figure 9d**). The genes from the cancer flipped group were of particular interest to us, since this set represents genes that are associated with tumours in both models but behave in completely opposite ways depending on whether we accounted for confounders or not. To our surprise, we observed a significant enrichment of the allograft rejection pathways in both sets of genes from the cancer flipped category that had either positive (red) or negative (blue) $\Delta logitPSI$. The same pathway also had a strong enrichment in the upregulated gens from the impurity category (**Figure 9d**). In addition, other Hallmark pathways such as *Mitotic Spindle*, *Interferon Gamma Response* and *Il2-STAT Signaling* were significantly enriched in different gene categories. Interestingly, we did not identify any

significantly enriched hallmark pathways in the genes we defined as potential false positives. These results collectively suggest a complex interplay between malignant and non-malignant AS programs across cancer Hallmark pathways.

To further validate the influence of different cell populations on the observed AS dysregulation of genes in Hallmark pathways, we analyzed single-cell RNA-seq profiles from primary tumours of clear-cell Renal Cell Carcinoma (ccRCC) [154]. **Figure 9e** shows the location of cancerous and non-cancerous cells for reference. Similarly, **Figure 9f** indicates the specific malignant, immune, stromal, and endothelial cell types present in these tumors. Standard quality control metrics for this dataset are shown in **Figure S16**.

We used GEDI [146] to reconstruct a shared low-dimensional cell representations across patients and conduct the downstream analysis of Hallmark pathway activities. When we overlayed the GEDI-inferred activity of each Hallmark pathway on the cell embedding, we observed a strong agreement with our findings from TRex analysis of bulk RNA-seq data of KIRC samples (**Figure 9g-i** and **Figure S16**). For instance, the IL2-STAT5 pathway that was overrepresented among *impurity* genes, showed a strong upregulation in the T-cell cluster (**Figure 9g**). The allograft rejection pathway, enriched in genes from the *cancer flipped* category, was downregulated in cancerous cells and upregulated in the population of natural killer (NK) cells and CD8+ T cells (**Figure 9h**). This can explain why the effect sizes of such genes are inverted once impurity is included in the model. Finally, the activity of the mitotic spindle pathway in single cells also explained the complex pattern observed in bulk data: as shown in **Figure 9d,** there were strong overlaps between mitotic spindle genes and events from the *cancer*, *cancer confounded*, and *cancer flipped* categories; measurements from single cells (**Figure 9i**) reflect how this pathway is active at varying degrees in subpopulations of cancer, immune and normal endothelial cells.

In summary, we showed that TRex successfully disentangles the effect of multiple factors underlying AS changes, even within the same gene. We revealed splicing-driven programs impacting cancer Hallmark pathways and, using single cell data, showcased their localization to different cell types. Importantly, the results from this case study highlight the imperative need to account for confounding factors, especially tumour purity, when measuring DAS using RNA-seq data from cancer cohorts.

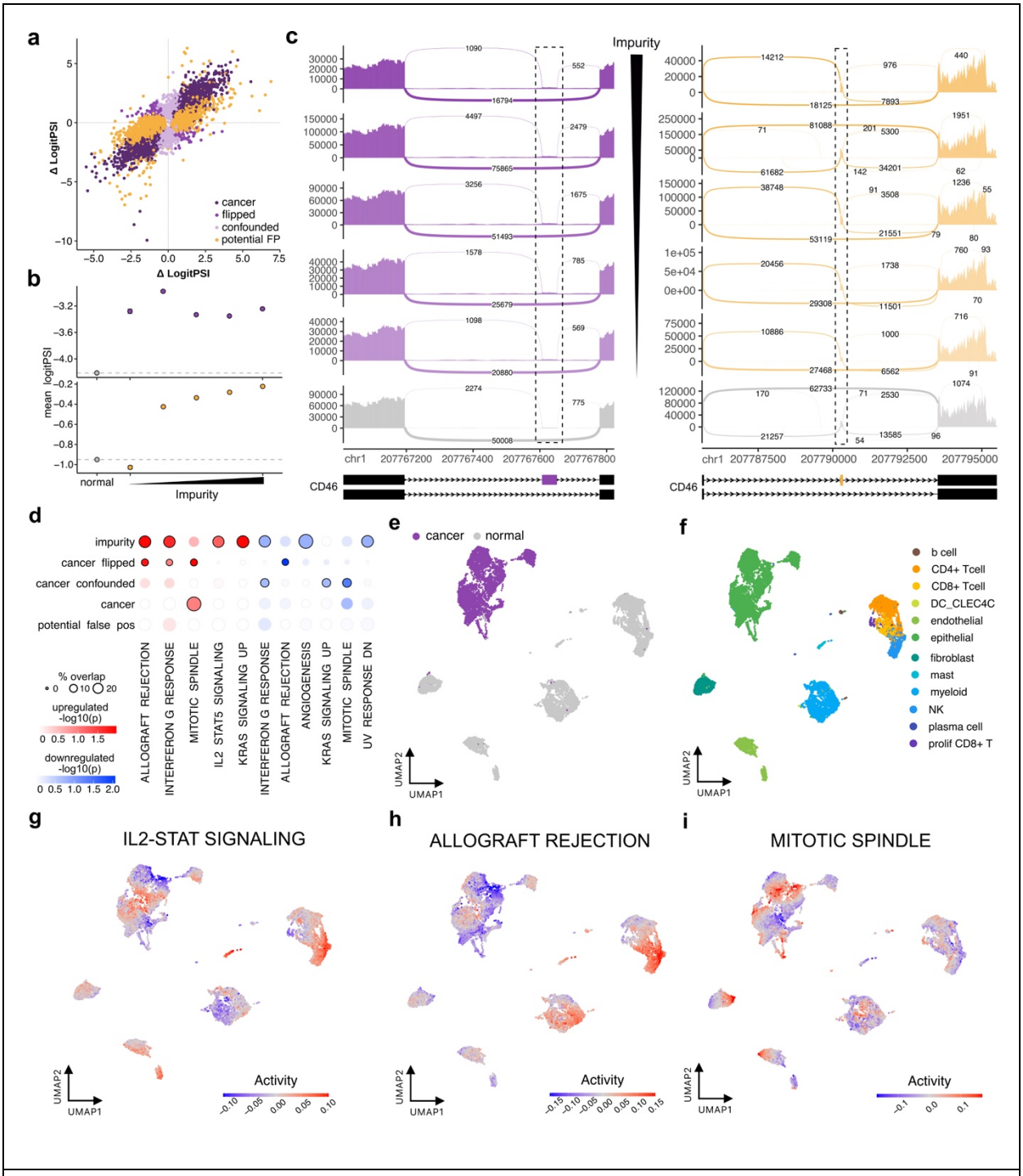


Figure 9 KIRC case study of tumour associated AS programs

(a) Comparison of tumour-associated changes derived from a simple (x axis) vs a complex (y axis) model, legend shows abbreviations of gene category names explained in the main text. **(b)** Mean logitPSI of two different SE events in the CD46 gene in normal samples and across tumour samples binned into five groups of low to high impurity. Top panel shows an event from the

cancer category (hg38 coordinates chr1:207767195-207767607:207767651-207767779:+); bottom panel shows an *impurity* event (hg38 coordinates chr1:207785682-207790253:207790345-207793519:+). **(c)** Sashimi plots of cancer event (in purple) and impurity event (in yellow) from **b** in normal samples and tumour samples grouped by impurity. **(d)** Gene overrepresentation analysis of cancer Hallmark pathways across genes with events in each category split by the direction of event effect sizes; solid black circles indicate significant pathways (adjusted p-value<0.05). UMAP embedding of single cells from a dataset of primary ccRCC tumours colored by **(e)** their inferCNV status, **(f)** cell type labels, **(g)** activity of the *Hallmark IL2-STAT Signaling* pathway, **(h)** activity of the *Allograft Rejection* pathway, and **(i)** activity of the *Mitotic Spindle* pathway.

3.6 RBPs as regulators of splicing programs associated with tumour development

To identify the upstream factors underlying the dysregulation of AS programs in cancer, we conducted a comprehensive analysis of all the RBPs with known binding motifs reported in CisBP-RNA [17]. Here, we sought to associate the binding of a given RBP to specific splice regions with downstream changes in splicing outcomes. First, we derived a set of representative RBP motifs by clustering all CisBP-RNA motifs based on their similarity (**Figure S17**). This procedure yielded 70 representative RBP motifs (**Figure S17**). We measured their affinity across sequences surrounding all the splice junctions (herein called *splice regions*) for each event type.

For each event type and each cancer, we modeled the tumor-associated $\Delta logitPSI$s as a linear function of the binding of RBPs to different splice regions. From these models, we obtained a series of differential splicing regulation (DSR) coefficients that reflect the extent to which the presence or absence of a certain RBP motif can explain the observed splicing differences between normal and tumour tissues. The optimal splice region size was determined based on the 10-fold cross-validation performance of models using different sizes. The size that showed consistently high performance across events and cancer types was 200 (**Figure S18**). A detailed description of the procedure outlined above can be found in Section 2.5.

3.6.1 Finding RBPs with differential splicing regulation effects

The DSR coefficients offer insights on the relationship between RBP binding and AS dysregulation within each cancer type. We leveraged observations across cancer types to rule out potential false positive RBPs by contrasting the DSR coefficients with differential expression changes of every RBP across cancer types. The rationale behind this is that an RBP whose DSR coefficients match the trend of its differential expression across cancer types is more likely to be a true positive regulator of AS. Hence, we tested the correlation between the DSR coefficient of each RBP—at every splice region of all seven AS types—with its tumour-associated differential expression across all the cancer types (**Figure 10a and Figure S19**).

Our results indicate that the splicing regulatory activities of RBPs are strongly dependent on the splice region they bind to. This is the case not only with regions of different types of events (**Figure S19**) but also for splice regions of the same event type (**Figure 10a**). When we examined more closely the regulation of SE events, we observed cases in which considering binding of the same RBP to a different region either yielded a stronger correlation or suggested completely opposite activities(**Figure 10a-e, Figure S19 and Figure S20**).

Interestingly, the effect of RBP binding on AS outcome appears to depend on the region to which the RBP binds. For example, binding of RNA Binding fox-1 (RBFOX1) downstream of the start position of the cassette exon (S2), but not upstream of it, is positively associated with cassette exon inclusion, since the inferred DSR coefficient for RBFOX1 binding to this region has a strong positive correlation with RBFOX1 differential expression (R=0.58; nominal P=0.00013, **Figure 10b**). A similar case is the Serine-Arginine Rich Splicing Factor 4 (SRSF4) (**Figure 10c**), where only binding to the downstream region of S2 is significantly associated with exon inclusion (R= –0.65). On the other hand, the effect of U1 small nuclear ribonucleoprotein A (SNRPA) binding ranges from strongly negative to positive in the upstream and the downstream regions of S2 (**Figure 10d**).

Figure 10 Upstream regulators of pan-cancer splicing programs

(a) Pearson correlations between RBP's differential expression (log2 fold-change in tumour vs normal) and DSR coefficients across cancers, separately calculated for each 200 nucleotide regions neighboring splice sites in SE events. Point sizes reflect number of cancer types used,

ranging from 15 (smallest) to 20 (largest). The number of cancer types included for each RBP varied depending on whether they expressed the RBP or not. Solid black outlines represent significant correlations (FDR<0.05). (**b-d**) Examples of the relationship between differential expression and DSR downstream of the S2 site of **(b)** RBFOX1, **(c)** SRSF4, and **(d)** SNRPA. **(e)** DSR coefficients of RBPs with at least one significant correlation in **(a).** Black outlines represent significant DSR coefficients (nominal P value <0.01). **(f)** Pathways significantly overrepresented (black outline circles, FDR<0.2) in RBP target genes with splicing changes that behave in an opposite way or similar to expression changes of the RBP across cancer types. Black outlined squares indicate significant correlations in **a**.

To further explore the region specificity of RBP activity, we systematically compared the individual DSR coefficients for all splice regions of each event type in every cancer (**Figure 10e** and **Figure S20**). In all even types, the DSR coefficients show region-specific activities, as opposed to cancer-specific patterns that describe all regions (**Figure S20**). In each event type, we could pinpoint the most informative region for differential RBP activity based on the number of RBPs with significant DSR coefficients (nominal P<0.01) across cancers. In the case of SE events, RBP binding upstream of E2 appears to be the most important splice region when predicting differential splicing regulation activities of RBPs (**Figure 10**). Such context-specific effects are consistent with previous reports showing that the RBP effect on splicing depends on the location to which they bind (Reviewed in [54]).

3.6.2 Molecular processes affected by RBP-mediated regulation of splicing

To gain insights into whether RBPs with DSR activities could underlie the splicing dysregulation of genes from Hallmark pathways across cancers, we conducted a GORA among targets of RBPs. The targets of RBPs were classified into two groups based on the sign of the correlation of the event's $\Delta logitPSI$ and the RBP's log2 fold change across cancers: RBP target events whose $\Delta logitPSI$ was positively correlated with RBP log2 fold expression across cancers are labeled as "similar", and those with negative correlation are labeled as "opposite". We found significant enrichment (FDR<0.2) of 19 Hallmark pathways among different groups of target genes of 14 RBPs (**Figure 10f**).

A recurring pathway appearing significantly enriched throughout this thesis is *Allograft rejection.* We detected an enrichment of genes from this pathway among genes with SE events targeted by Musash-2 (MSI2). Based on the relationships between MSI2 differential expression, DSR coefficients upstream of S3, and the $\Delta logitPSI$ of its targets across cancer types, we can conclude that MSI2 binding upstream of S3 promotes the inclusion of its target cassette exons (**Figure 10e**). This case is simply highlighting an example of the results obtained through this analysis, but further analysis is needed to systematically explore these results and confirm the most striking observations, as discussed in the next chapter. Overall, our findings support the association between RBPs-mediated splicing dysregulation across cancers and Hallmark gene signatures.

CHAPTER 4 DISCUSSION

Here, we conducted a comprehensive analysis of seven types of AS programs and their upstream regulators across 31 cancer types from the TCGA (**Figure 5**). Our methodological innovation filled a major gap in the field and enabled us to address previously unappreciated sample-level confounders impacting the quantification of AS from RNA-seq data. For the first time, we provide an atlas of AS regulatory programs associated with tumour development and progression across cancer types which considers the variability in cell composition and other confounding factors. In this chapter, I will explore some of the implications as well as limitations of this study.

To build this resource, we focused on two types of models. The first one aimed at characterizing the dysregulation of AS that occurs when a normal cell develops into a tumour mass, we refer to these as *tumour models*. The second series of models, which we will call *stage models*, focused on characterizing the programs that accompany tumour stage progression as a proxy for detecting pro-metastatic changes. In both analyses, we included three sample-level confounders: tumour purity (the proportion of non-cancerous cells present in a tumour sample), age (in years), and sex (male or female).

In a pivotal study published in 2021, Slaff et al. [123] emphasized the challenge of confounding factors in splicing analysis, particularly when comparing the effects of sample batch processing on gene expression and alternative splicing (AS) quantification using RNA-seq data from the ENCODE consortium. The study introduced the Modeling Confounding Factors Affecting Splicing Quantification (MOCCASIN) tool, which serves as a pre-processing step in the second version of MAJIQ [111]. While MOCCASIN effectively removes confounding effects from RNA-seq count data for downstream AS analysis, it functions primarily as a normalization tool, lacking the ability to directly interpret the impact of each factor on observed differential splicing changes across different experimental groups, as demonstrated by TRex in our study. TRex, stands out as the only method capable of measuring seven types of AS with experimental variables in large datasets featuring complex designs.

Addressing a significant methodological gap, TRex stands out as the only method capable of measuring seven types of alternative splicing with experimental variables in large datasets featuring complex designs. Applied to 8,633 transcriptomes of tumor and normal samples of 24

cancer types from TCGA to detect associations with tumor development and progression, TRex successfully removed the effect of confounders from these. Consistent across all seven alternative splicing (AS) event types, we observed fewer stage-associated events compared to tumor-associated events, exemplified by 8,485 stage-associated SE events versus 31,873 tumor-associated SE events with FDR<0.05 in at least one cancer type (**Figure S8**). As an additional sanity check, we confirmed that our measurements of DAS were not merely capturing gene expression changes (**Figure S13**, **Supplementary Data Table 2**). In addition, we did not observe any obvious dependency between the proportion of significant DAS and the fraction of normal samples per cancer type (**Figure S14**).

Regarding the effects of confounding factors, impurity was clearly the variable with the largest effect on splicing measurements out of the three tested (**Figure 6**). Given the minor effect that age and sex have compared to impurity, our discussion will be centered on the latter. The confounding effect of tumour purity was already known to bias differential gene expression analyses [117, 118]. Nonetheless, this important matter has been neglected by existing tools for quantification of the seven AS event types studied here, despite relying on the same type of data. Our findings show that tumour purity can bias the quantification of splicing to at least the same degree as gene expression—a systematic comparison of impurity vs the first 10 PCs of each cancer type derived from either gene expression or cassette exon splicing revealed comparable correlations in both settings (**Figure 6**). The proportion of variance explained by the PCs that significantly correlated (FDR<0.05) with tumour purity ranged from 25% to 75% in both cases (expression and splicing), and every cancer type had at least one PC significantly correlated with purity, highlighting the importance of accounting for tumour purity in the quantification of differential AS.

In the ideal scenario, the purity of a sample is estimated from DNA copy number variations that distinguish cancer cells from immune and stromal cells, which is the strategy used by methods like ABSOLUTE [139]. However, whole-genome/exome sequencing data is not always available for RNA-seq samples. Therefore, computational methods have been developed to infer the purity of a sample based on the expression of known immune and stromal marker genes. Therefore, in order to include as many samples as possible in our analysis, we opted for using the method ESTIMATE to infer tumour purities directly from RNA-seq data [140].

Deconvolving the exact cell type composition from bulk RNAseq data is a challenging task, and the resulting inferences may not always be an accurate depiction of the real tumour. In fact, it was necessary to re-calibrate ESTIMATE's empirical formula because ~25% of the current TCGA samples were not included when the method was developed and yielded tumour purities outside the 0–1 range. After recalibration, we achieved a Pearson correlation of 0.69 ($p<2.2e-16$) with ABSOLUTE purities in the subset of samples that had both genome/exome-sequencing and RNA-seq data (**Figure S1**). We recognize that expression-based estimates of impurity do not perfectly correlate with DNA-based estimates; this limitation could be addressed by using DNA-based purity metrics as more cancer genome datasets with joint transcriptome profiles become available.

Another limitation of our approach is that we assume that normal samples have an impurity of zero; in other words, we assume that the cellular composition of normal tissue is the same across all samples, and that the metric of interest is the difference between splicing in neoplastic cells of the tumour and the average of the cells of the normal tissue. We recognize, however, that normal samples consist of various cell types with potentially varying proportions across samples. Furthermore, it will be of interest to precisely identify the differences between neoplastic cells and their cell type of origin in the normal tissue, as opposed to the average of all cell types present in the tissue of origin. This could be addressed in the future by incorporating an approach to quantify and model the cellular composition of normal samples.

The previous paragraphs illustrate one of the most challenging drawbacks we faced in this study: distinguishing the cells underlying the observed splicing patterns. Although our models account for impurity, they operate under the assumption that there is no crosstalk between the tumour and its microenvironment. In other words, a change in cellular composition (such as increase in the non-neoplastic fraction of tumour cells) does not affect the splicing landscape of cancer cells. There is a growing body of evidence suggesting that tumour-microennvrionment interactions help shape the transcriptome of cancerous cells in the tumor (reviewed in [157]). Thus, we acknowledge that some of our of observations may be influenced by the crosstalk between malignant cells and the tumour microenvironment. Nonetheless, addressing this matter when relying on bulk-level measurements is an extremely challenging task, inheret to this measurement technolgy.

Delving into the functional implications, we identified significant enrichment of *Mitotic Spindle* and *Myogenesis* pathways genes in SE cancer events across 18 out of the 20 cancer types studied (**Figure 8**). Notably, these pathways are associated with cell proliferation—a hallmark of cancer—thus suggesting a potential role of splicing in sustaining the proliferative capacity of malignant cells. The same pathways were also overrepresented in stage-associated SE events, albeit in lower numbers than tumor-associated events (**Figure S7**). The varying proportion of significant events shared by multiple cancers revealed intriguing trends across AS types, with intron retention events displaying a distinctive pattern, further emphasizing the diverse mechanisms of splicing dysregulation across cancer types.

Examining events shared by multiple cancers revealed distinct proportions of significance per AS type, with respect to the number of events of each type with significant associations in at least one cancer (n=147,268 across all event types) (**Figure 8**). While six AS types showed similar patterns, with 49-56% of significant events appearing in two or less cancers, RI events demonstrated a unique trend. A smaller fraction of RI events (~40%) was associated with tumours in maximum two cancers, indicating that ~60% of them were shared by at least three cancer types. These differences become more noticeable when interrogating a larger number of cancers (**Figure 8**). This observation aligns with a previous study spanning three TCGA cancer types, emphasizing the recurrent dysregulation of RI events across various cancers [75]. The connection between intron retention and mRNA degradation through nonsense-mediated decay (NMD) suggests a common mechanism of gene silencing in tumors [75], reinforcing the potential significance of this splicing mechanism in cancer cell remodeling. Thus, our findings emphasize the recurrent dysregulation of RI events across diverse cancers and underscore the potential importance of this splicing mechanism in cancer cell remodeling.

The results of this thesis also contribute to the evolving understanding of the role of RBPs in cancer, particularly in the context of AS dysregulation. Recent attention has turned towards the dynamic and intricate regulatory role of RBPs in shaping the cancer transcriptome by influencing processes such as mRNA stability, localization, and alternative splicing [158] [159] [160] [161] [54]. This study takes a systematic approach to dissecting the contribution of RBPs to the dysregulation AS across cancers. By analyzing the binding affinities of RBPs to specific splice regions, we unveiled context-specific RBP activities that persist across cancer types, emphasizing that the impact of RBPs on the outcomes of AS depends on the precise region to which they bind.

Moreover, the correlation analysis between changes in RBP expression and DSR coefficients across various cancers strengthens the connection between RBP dysregulation and AS alterations. The incorporation of Hallmark pathways into the analysis provides functional context, suggesting that identified RBPs with significant DSR activities may serve as key regulators in pathways crucial to cancer progression. The recurring enrichment of the Allograft rejection pathway among genes with splicing events targeted by specific RBPs hints at potential links between RBP-mediated AS dysregulation and modulation of immune-related processes.

Finally, beyond its implications for cancer research, this thesis represents a methodological innovation in the field of splicing quantification. The TRex framework, utilized in this study, not only addresses tumor and stage-associated changes but is also capable of incorporating the effects of various experimental variables into its differential AS testing framework. The adaptability of TRex to quantify both known and novel AS events, if provided with reference annotations, positions it as a versatile tool with applications extending beyond cancer research.

CHAPTER 5 CONCLUDING REMARKS

5.1 Conclusion

In summary, we successfully developed a computational method to associate differential alternative splicing with experimental variables of interest. TRex not only outperforms the state-of-the-art, but also fills a methodological gap in the field. The application of TRex to 24 cancer types from TCGA revealed that intron retention is the most frequently dysregulated form of alternative splicing. Similarly, we discovered that the Hallmark pathways of myogenesis and mitotic spindle are disrupted by the dysregulation of AS in nearly all cancer types. Furthermore, our findings indicate that there is a much larger alteration of splicing programs between tumour and normal samples than across tumours of different stages. Our comprehensive analysis of upstream regulators of splicing revealed 39 RBPs with potential roles in the cancer-associated splicing programs across multiple cancer types. Overall, the results of this thesis underline dysregulation of alternative splicing as a widespread mechanism exploited by cancerous cell throughout the development of oncogenic phenotypes in tumours.

5.2 Future directions

Apart from the potential avenues previously discussed, a direct continuation of this project would be the further validation of stage-associated events. A valuable experiment would be to contrast them with metastasis-associated events obtained from a model that includes primary tumours and their matched metastases. Such comparison would facilitate the interpretation of stage-associated events as pro-metastatic programs across cancer types, without the need for additional matching primary and metastatic tumour pairs from all cancer types. Similarly, performing experimental validations of RBP-mediated splicing programs could result in the identification of promising therapeutic and/or diagnostic targets. In the future, we aim to expand TRex to further characterize alternative splicing programs at the single cell level.

CHAPTER 6 REFERENCES

1. Bonnal, S.C., I. Lopez-Oreja, and J. Valcarcel, *Roles and mechanisms of alternative splicing in cancer - implications for care.* Nat Rev Clin Oncol, 2020. **17**(8): p. 457-474.
2. Ule, J. and B.J. Blencowe, *Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution.* Mol Cell, 2019. **76**(2): p. 329-345.
3. David, C.J. and J.L. Manley, *Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged.* Genes Dev, 2010. **24**(21): p. 2343-64.
4. Ladomery, M., *Aberrant alternative splicing is another hallmark of cancer.* Int J Cell Biol, 2013. **2013**: p. 463786.
5. Oltean, S. and D.O. Bates, *Hallmarks of alternative splicing in cancer.* Oncogene, 2014. **33**(46): p. 5311-8.
6. Seiler, M., et al., *Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types.* Cell Rep, 2018. **23**(1): p. 282-296 e4.
7. Supek, F., et al., *Synonymous mutations frequently act as driver mutations in human cancers.* Cell, 2014. **156**(6): p. 1324-1335.
8. Yoshida, K., et al., *Frequent pathway mutations of splicing machinery in myelodysplasia.* Nature, 2011. **478**(7367): p. 64-9.
9. Climente-Gonzalez, H., E. Porta-Pardo, A. Godzik, and E. Eyras, *The Functional Impact of Alternative Splicing in Cancer.* Cell Rep, 2017. **20**(9): p. 2215-2226.
10. Yang, Q., et al., *Aberrant alternative splicing in breast cancer.* J Mol Cell Biol, 2019. **11**(10): p. 920-929.
11. Kahles, A., et al., *Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients.* Cancer Cell, 2018. **34**(2): p. 211-224 e6.
12. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, 2014. **344**(6190): p. 1396-401.
13. TCGA, *The Cancer Genome Atlas*. TCGA Research Network.
14. Dvinge, H., E. Kim, O. Abdel-Wahab, and R.K. Bradley, *RNA splicing factors as oncoproteins and tumour suppressors.* Nat Rev Cancer, 2016. **16**(7): p. 413-30.

15. Sebestyen, E., et al., *Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks.* Genome Res, 2016. **26**(6): p. 732-44.
16. Sveen, A., et al., *Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes.* Oncogene, 2016. **35**(19): p. 2413-27.
17. Ray, D., et al., *A compendium of RNA-binding motifs for decoding gene regulation.* Nature, 2013. **499**(7457): p. 172-7.
18. Fish, L., et al., *A prometastatic splicing program regulated by SNRPA1 interactions with structured RNA elements.* Science, 2021. **372**(6543).
19. Tavazoie, S.F., et al., *Endogenous human microRNAs that suppress breast cancer metastasis.* Nature, 2008. **451**(7175): p. 147-52.
20. Deng, K., et al., *Abnormal alternative splicing promotes tumor resistance in targeted therapy and immunotherapy.* Transl Oncol, 2021. **14**(6): p. 101077.
21. Isobe, K., et al., *Association of BIM Deletion Polymorphism and BIM-gamma RNA Expression in NSCLC with EGFR Mutation.* Cancer Genomics Proteomics, 2016. **13**(6): p. 475-482.
22. Marasco, L.E. and A.R. Kornblihtt, *The physiology of alternative splicing.* Nat Rev Mol Cell Biol, 2023. **24**(4): p. 242-254.
23. Rogalska, M.E., C. Vivori, and J. Valcarcel, *Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects.* Nat Rev Genet, 2023. **24**(4): p. 251-269.
24. Wright, C.J., C.W.J. Smith, and C.D. Jiggins, *Alternative splicing as a source of phenotypic diversity.* Nat Rev Genet, 2022. **23**(11): p. 697-710.
25. Black, C.S., et al., *Spliceosome assembly and regulation: insights from analysis of highly reduced spliceosomes.* RNA, 2023. **29**(5): p. 531-550.
26. Plaschka, C., P.C. Lin, C. Charenton, and K. Nagai, *Prespliceosome structure provides insights into spliceosome assembly and regulation.* Nature, 2018. **559**(7714): p. 419-422.
27. Fredericks, A.M., K.J. Cygan, B.A. Brown, and W.G. Fairbrother, *RNA-Binding Proteins: Splicing Factors and Disease.* Biomolecules, 2015. **5**(2): p. 893-909.
28. Wang, Y., M. Ma, X. Xiao, and Z. Wang, *Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules.* Nat Struct Mol Biol, 2012. **19**(10): p. 1044-52.

29. Fairbrother, W.G., D. Holste, C.B. Burge, and P.A. Sharp, *Single nucleotide polymorphism-based validation of exonic splicing enhancers.* PLoS Biol, 2004. **2**(9): p. E268.
30. Xiao, X., Z. Wang, M. Jang, and C.B. Burge, *Coevolutionary networks of splicing cis-regulatory elements.* Proc Natl Acad Sci U S A, 2007. **104**(47): p. 18583-8.
31. Wang, Y., et al., *Mechanism of alternative splicing and its regulation.* Biomed Rep, 2015. **3**(2): p. 152-158.
32. Wilkinson, M.E., C. Charenton, and K. Nagai, *RNA Splicing by the Spliceosome.* Annu Rev Biochem, 2020. **89**: p. 359-388.
33. Yan, C., R. Wan, and Y. Shi, *Molecular Mechanisms of pre-mRNA Splicing through Structural Biology of the Spliceosome.* Cold Spring Harb Perspect Biol, 2019. **11**(1).
34. Sheth, N., et al., *Comprehensive splice-site analysis using comparative genomics.* Nucleic Acids Res, 2006. **34**(14): p. 3955-67.
35. Zhuang, Y. and A.M. Weiner, *A compensatory base change in U1 snRNA suppresses a 5' splice site mutation.* Cell, 1986. **46**(6): p. 827-35.
36. Kramer, A., P. Gruter, K. Groning, and B. Kastner, *Combined biochemical and electron microscopic analyses reveal the architecture of the mammalian U2 snRNP.* J Cell Biol, 1999. **145**(7): p. 1355-68.
37. Berglund, J.A., N. Abovich, and M. Rosbash, *A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition.* Genes Dev, 1998. **12**(6): p. 858-67.
38. Bringmann, P., et al., *Evidence for the existence of snRNAs U4 and U6 in a single ribonucleoprotein complex and for their association by intermolecular base pairing.* EMBO J, 1984. **3**(6): p. 1357-63.
39. Agafonov, D.E., et al., *Molecular architecture of the human U4/U6.U5 tri-snRNP.* Science, 2016. **351**(6280): p. 1416-20.
40. Galej, W.P., et al., *Cryo-EM structure of the spliceosome immediately after branching.* Nature, 2016. **537**(7619): p. 197-201.
41. Plaschka, C., P.C. Lin, and K. Nagai, *Structure of a pre-catalytic spliceosome.* Nature, 2017. **546**(7660): p. 617-621.
42. Ohrt, T., et al., *Molecular dissection of step 2 catalysis of yeast pre-mRNA splicing investigated in a purified system.* RNA, 2013. **19**(7): p. 902-15.

43. Schwer, B., *A conformational rearrangement in the spliceosome sets the stage for Prp22-dependent mRNA release.* Mol Cell, 2008. **30**(6): p. 743-54.

44. Arenas, J.E. and J.N. Abelson, *Prp43: An RNA helicase-like factor involved in spliceosome disassembly.* Proc Natl Acad Sci U S A, 1997. **94**(22): p. 11798-802.

45. GENCODE. *Statistics about the current GENCODE Release (version 43)*. 2022 2022 [cited 2023 May 31]; Available from: https://www.gencodegenes.org/human/stats.html.

46. Schneider, V.A., et al., *Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly.* Genome Res, 2017. **27**(5): p. 849-864.

47. Zhang, Z., et al., *Deep-learning augmented RNA-seq analysis of transcript splicing.* Nat Methods, 2019. **16**(4): p. 307-310.

48. Hardwick, S.A., et al., *Getting the Entire Message: Progress in Isoform Sequencing.* Front Genet, 2019. **10**: p. 709.

49. Prjibelski, A.D., et al., *Accurate isoform discovery with IsoQuant using long reads.* Nat Biotechnol, 2023. **41**(7): p. 915-918.

50. Ebrahimie, E., S. Rahimirad, M. Tahsili, and M. Mohammadi-Dehcheshmeh, *Alternative RNA splicing in stem cells and cancer stem cells: Importance of transcript-based expression analysis.* World J Stem Cells, 2021. **13**(10): p. 1394-1416.

51. Hiller, M. and M. Platzer, *Widespread and subtle: alternative splicing at short-distance tandem sites.* Trends Genet, 2008. **24**(5): p. 246-55.

52. Gracheva, E.O., et al., *Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats.* Nature, 2011. **476**(7358): p. 88-91.

53. Penev, A., et al., *Alternative splicing is a developmental switch for hTERT expression.* Mol Cell, 2021. **81**(11): p. 2349-2360 e6.

54. Fu, X.D. and M. Ares, Jr., *Context-dependent control of alternative splicing by RNA-binding proteins.* Nat Rev Genet, 2014. **15**(10): p. 689-701.

55. Black, D.L., *Mechanisms of alternative pre-messenger RNA splicing.* Annu Rev Biochem, 2003. **72**: p. 291-336.

56. Lopez, A.J., *Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation.* Annu Rev Genet, 1998. **32**: p. 279-305.

57. Gerstberger, S., M. Hafner, and T. Tuschl, *A census of human RNA-binding proteins.* Nat Rev Genet, 2014. **15**(12): p. 829-45.

58. Van Nostrand, E.L., et al., *Author Correction: A large-scale binding and functional map of human RNA-binding proteins.* Nature, 2021. **589**(7842): p. E5.

59. Van Nostrand, E.L., et al., *A large-scale binding and functional map of human RNA-binding proteins.* Nature, 2020. **583**(7818): p. 711-719.

60. Fu, X.D., *The superfamily of arginine/serine-rich splicing factors.* RNA, 1995. **1**(7): p. 663-80.

61. Huelga, S.C., et al., *Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins.* Cell Rep, 2012. **1**(2): p. 167-78.

62. Yang, G., X. Lu, and L. Yuan, *LncRNA: a link between RNA and cancer.* Biochim Biophys Acta, 2014. **1839**(11): p. 1097-109.

63. Naro, C. and C. Sette, *Phosphorylation-mediated regulation of alternative splicing in cancer.* Int J Cell Biol, 2013. **2013**: p. 151839.

64. Feng, Y., M. Chen, and J.L. Manley, *Phosphorylation switches the general splicing repressor SRp38 to a sequence-specific activator.* Nat Struct Mol Biol, 2008. **15**(10): p. 1040-8.

65. Lai, M.C., R.I. Lin, and W.Y. Tarn, *Transportin-SR2 mediates nuclear import of phosphorylated SR proteins.* Proc Natl Acad Sci U S A, 2001. **98**(18): p. 10154-9.

66. Manceau, V., et al., *Major phosphorylation of SF1 on adjacent Ser-Pro motifs enhances interaction with U2AF65.* FEBS J, 2006. **273**(3): p. 577-87.

67. Muniz, L., E. Nicolas, and D. Trouche, *RNA polymerase II speed: a key player in controlling and adapting transcriptome composition.* EMBO J, 2021. **40**(15): p. e105740.

68. Shayevitch, R., D. Askayo, I. Keydar, and G. Ast, *The importance of DNA methylation of exons on alternative splicing.* RNA, 2018. **24**(10): p. 1351-1362.

69. Dvinge, H., J. Guenthoer, P.L. Porter, and R.K. Bradley, *RNA components of the spliceosome regulate tissue- and cancer-specific alternative splicing.* Genome Res, 2019. **29**(10): p. 1591-1604.

70. Hanahan, D., *Hallmarks of Cancer: New Dimensions.* Cancer Discov, 2022. **12**(1): p. 31-46.

71. Bradley, R.K. and O. Anczukow, *RNA splicing dysregulation and the hallmarks of cancer.* Nat Rev Cancer, 2023. **23**(3): p. 135-155.

72. Zhang, Y., J. Qian, C. Gu, and Y. Yang, *Alternative splicing and cancer: a systematic review.* Signal Transduct Target Ther, 2021. **6**(1): p. 78.

73. Zhang, S., et al., *A widespread length-dependent splicing dysregulation in cancer.* Sci Adv, 2022. **8**(33): p. eabn9232.

74. Shen, S., et al., *SURVIV for survival analysis of mRNA isoform variation.* Nat Commun, 2016. **7**: p. 11548.

75. Tsai, Y.S., D. Dominguez, S.M. Gomez, and Z. Wang, *Transcriptome-wide identification and study of cancer-specific splicing events across multiple tumors.* Oncotarget, 2015. **6**(9): p. 6825-39.

76. Maciejewski, J.P. and R.A. Padgett, *Defects in spliceosomal machinery: a new pathway of leukaemogenesis.* Br J Haematol, 2012. **158**(2): p. 165-173.

77. Cazzola, M., M. Rossi, L. Malcovati, and M. Associazione Italiana per la Ricerca sul Cancro Gruppo Italiano Malattie, *Biologic and clinical significance of somatic mutations of SF3B1 in myeloid and lymphoid neoplasms.* Blood, 2013. **121**(2): p. 260-9.

78. Yoshida, K. and S. Ogawa, *Splicing factor mutations and cancer.* Wiley Interdiscip Rev RNA, 2014. **5**(4): p. 445-59.

79. Urbanski, L.M., N. Leclair, and O. Anczukow, *Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics.* Wiley Interdiscip Rev RNA, 2018. **9**(4): p. e1476.

80. Wu, S., et al., *ASCancer Atlas: a comprehensive knowledgebase of alternative splicing in human cancers.* Nucleic Acids Res, 2023. **51**(D1): p. D1196-D1204.

81. Zhang, Y., et al., *OncoSplicing: an updated database for clinically relevant alternative splicing in 33 human cancers.* Nucleic Acids Res, 2022. **50**(D1): p. D1340-D1347.

82. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.

83. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.

84. Warburg, O., *On the origin of cancer cells.* Science, 1956. **123**(3191): p. 309-14.

85. Christofk, H.R., et al., *The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth.* Nature, 2008. **452**(7184): p. 230-3.

86. Jurica, M.S., et al., *The allosteric regulation of pyruvate kinase by fructose-1,6-bisphosphate.* Structure, 1998. **6**(2): p. 195-210.

87. Noguchi, T., H. Inoue, and T. Tanaka, *The M1- and M2-type isozymes of rat pyruvate kinase are produced from the same gene by alternative RNA splicing.* J Biol Chem, 1986. **261**(29): p. 13807-12.

88. Dayton, T.L., T. Jacks, and M.G. Vander Heiden, *PKM2, cancer metabolism, and the road ahead.* EMBO Rep, 2016. **17**(12): p. 1721-1730.

89. Mazurek, S., C.B. Boschek, F. Hugo, and E. Eigenbrodt, *Pyruvate kinase type M2 and its role in tumor growth and spreading.* Semin Cancer Biol, 2005. **15**(4): p. 300-8.

90. Surget, S., M.P. Khoury, and J.C. Bourdon, *Uncovering the role of p53 splice variants in human malignancy: a clinical perspective.* Onco Targets Ther, 2013. **7**: p. 57-68.

91. Flaman, J.M., et al., *The human tumour suppressor gene p53 is alternatively spliced in normal cells.* Oncogene, 1996. **12**(4): p. 813-8.

92. Bourdon, J.C., et al., *p53 isoforms can regulate p53 transcriptional activity.* Genes Dev, 2005. **19**(18): p. 2122-37.

93. Khoury, M.P. and J.C. Bourdon, *The isoforms of the p53 protein.* Cold Spring Harb Perspect Biol, 2010. **2**(3): p. a000927.

94. Fujita, K., et al., *p53 isoforms Delta133p53 and p53beta are endogenous regulators of replicative cellular senescence.* Nat Cell Biol, 2009. **11**(9): p. 1135-42.

95. Courtois, S., et al., *DeltaN-p53, a natural isoform of p53 lacking the first transactivation domain, counteracts growth suppression by wild-type p53.* Oncogene, 2002. **21**(44): p. 6722-8.

96. Bielli, P., M. Bordi, V. Di Biasio, and C. Sette, *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, 2014. **42**(19): p. 12070-81.

97. Garneau, D., T. Revil, J.F. Fisette, and B. Chabot, *Heterogeneous nuclear ribonucleoprotein F/H proteins modulate the alternative splicing of the apoptotic mediator Bcl-x.* J Biol Chem, 2005. **280**(24): p. 22641-50.

98. Paronetto, M.P., et al., *The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x.* J Cell Biol, 2007. **176**(7): p. 929-39.

99. Harper, S.J. and D.O. Bates, *VEGF-A splicing: the key to anti-angiogenic therapeutics?* Nat Rev Cancer, 2008. **8**(11): p. 880-7.

100. Brown, R.L., et al., *CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression.* J Clin Invest, 2011. **121**(3): p. 1064-74.

101. Murphy, A.J., A.H. Li, P. Li, and H. Sun, *Therapeutic Targeting of Alternative Splicing: A New Frontier in Cancer Treatment.* Front Oncol, 2022. **12**: p. 868664.

102. Fackenthal, J.D., *Alternative mRNA Splicing and Promising Therapies in Cancer.* Biomolecules, 2023. **13**(3).

103. Shen, S., et al., *rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data.* Proc Natl Acad Sci U S A, 2014. **111**(51): p. E5593-601.

104. Trincado, J.L., et al., *SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions.* Genome Biol, 2018. **19**(1): p. 40.

105. Li, Y.I., et al., *Annotation-free quantification of RNA splicing using LeafCutter.* Nat Genet, 2018. **50**(1): p. 151-158.

106. Katz, Y., E.T. Wang, E.M. Airoldi, and C.B. Burge, *Analysis and design of RNA sequencing experiments for identifying isoform regulation.* Nat Methods, 2010. **7**(12): p. 1009-15.

107. Kahles, A., C.S. Ong, Y. Zhong, and G. Ratsch, *SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data.* Bioinformatics, 2016. **32**(12): p. 1840-7.

108. Anders, S., A. Reyes, and W. Huber, *Detecting differential usage of exons from RNA-seq data.* Genome Res, 2012. **22**(10): p. 2008-17.

109. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

110. Vaquero-Garcia, J., et al., *A new view of transcriptome complexity and regulation through the lens of local splicing variations.* Elife, 2016. **5**: p. e11752.

111. Vaquero-Garcia, J., et al., *RNA splicing analysis using heterogeneous and large RNA-seq datasets.* Nat Commun, 2023. **14**(1): p. 1230.

112. Mehmood, A., et al., *Systematic evaluation of differential splicing tools for RNA-seq studies.* Brief Bioinform, 2020. **21**(6): p. 2052-2065.

113. Liu, R., A.E. Loraine, and J.A. Dickerson, *Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems.* BMC Bioinformatics, 2014. **15**(1): p. 364.

114. Joyce, J.A. and J.W. Pollard, *Microenvironmental regulation of metastasis.* Nat Rev Cancer, 2009. **9**(4): p. 239-52.

115. Aran, D., M. Sirota, and A.J. Butte, *Systematic pan-cancer analysis of tumour purity.* Nat Commun, 2015. **6**: p. 8971.

116. Aran, D., M. Sirota, and A.J. Butte, *Corrigendum: Systematic pan-cancer analysis of tumour purity.* Nat Commun, 2016. **7**: p. 10707.

117. Rhee, J.K., et al., *Impact of Tumor Purity on Immune Gene Expression and Clustering Analyses across Multiple Cancer Types.* Cancer Immunol Res, 2018. **6**(1): p. 87-97.

118. Zhang, W., H. Long, B. He, and J. Yang, *DECtp: Calling Differential Gene Expression Between Cancer and Normal Samples by Integrating Tumor Purity Information.* Front Genet, 2018. **9**: p. 321.

119. Shen, Q., et al., *contamDE: differential expression analysis of RNA-seq data for contaminated tumor samples.* Bioinformatics, 2016. **32**(5): p. 705-12.

120. Molania, R., et al., *Removing unwanted variation from large-scale RNA sequencing data with PRPS.* Nat Biotechnol, 2023. **41**(1): p. 82-95.

121. Ji, Y., C. Yu, and H. Zhang, *contamDE-lm: linear model-based differential gene expression analysis using next-generation RNA-seq data from contaminated tumor samples.* Bioinformatics, 2020. **36**(8): p. 2492-2499.

122. Carvalho, J., *Cell Reversal From a Differentiated to a Stem-Like State at Cancer Initiation.* Front Oncol, 2020. **10**: p. 541.

123. Slaff, B., et al., *MOCCASIN: a method for correcting for known and unknown confounders in RNA splicing analysis.* Nat Commun, 2021. **12**(1): p. 3353.

124. Zhang, Y., G. Parmigiani, and W.E. Johnson, *ComBat-seq: batch effect adjustment for RNA-seq count data.* NAR Genom Bioinform, 2020. **2**(3): p. lqaa078.

125. Stegle, O., L. Parts, R. Durbin, and J. Winn, *A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies.* PLoS Comput Biol, 2010. **6**(5): p. e1000770.

126. Leek, J.T. and J.D. Storey, *Capturing heterogeneity in gene expression studies by surrogate variable analysis.* PLoS Genet, 2007. **3**(9): p. 1724-35.

127. Consortium, G.T., *The GTEx Consortium atlas of genetic regulatory effects across human tissues.* Science, 2020. **369**(6509): p. 1318-1330.

128. Srivastava, A., et al., *Alignment and mapping methodology influence transcript abundance estimation.* Genome Biol, 2020. **21**(1): p. 239.

129. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.

130. Bray, N.L., H. Pimentel, P. Melsted, and L. Pachter, *Near-optimal probabilistic RNA-seq quantification.* Nat Biotechnol, 2016. **34**(5): p. 525-7.

131. Soneson, C., M.I. Love, and M.D. Robinson, *Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences.* F1000Res, 2015. **4**: p. 1521.

132. Frankish, A., et al., *Gencode 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D916-D923.

133. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

134. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

135. Luo, Y., et al., *New developments on the Encyclopedia of DNA Elements (ENCODE) data portal.* Nucleic Acids Res, 2020. **48**(D1): p. D882-D889.

136. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.

137. Colaprico, A., et al., *TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data.* Nucleic Acids Res, 2016. **44**(8): p. e71.

138. Privé, F., *bigutilsr*.

139. Carter, S.L., et al., *Absolute quantification of somatic DNA alterations in human cancer.* Nat Biotechnol, 2012. **30**(5): p. 413-21.

140. Yoshihara, K., et al., *Inferring tumour purity and stromal and immune cell admixture from expression data.* Nat Commun, 2013. **4**: p. 2612.

141. Korotkevich G., S.V., and Sergushichev A., *Fast gene set enrichment analysis.* bioRxiv, 2019.

142. Bhuva D., S.G.a.G.A., *msigdb: An ExperimentHub Package for the Molecular Signatures Database*. Bioconducto: Bioconductor.

143. Zhang, X., et al., *CellMarker: a manually curated resource of cell markers in human and mouse.* Nucleic Acids Res, 2019. **47**(D1): p. D721-D728.

144. Chen, E.Y., et al., *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.* BMC Bioinformatics, 2013. **14**: p. 128.

145. Xie, Z., et al., *Gene Set Knowledge Discovery with Enrichr.* Curr Protoc, 2021. **1**(3): p. e90.

146. Madrigal, A., Lu, T., Soto, L. M. and Najafabadi, H. S., *A unified model for interpretable latent embedding of multi-sample, multi-condition single-cell data.* bioRxiv, 2023.

147. Lambert, S.A., M. Albu, T.R. Hughes, and H.S. Najafabadi, *Motif comparison based on similarity of binding affinity profiles.* Bioinformatics, 2016. **32**(22): p. 3504-3506.

148. Lawrence, M., et al., *Software for computing and annotating genomic ranges.* PLoS Comput Biol, 2013. **9**(8): p. e1003118.

149. TBD, T., *BSgenome.Hsapiens.UCSC.hg38: Full genomic sequences for Homo sapiens (UCSC genome hg38)*. 2023: Bioconductor.

150. Pagès, H., *Biostrings: Efficient manipulation of biological strings*. 2023, Bioconductor.

151. Kuhn, M., *caret: Classification and Regression Training*. 2019: CRAN.

152. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* J Stat Softw, 2010. **33**(1): p. 1-22.

153. Tay, J.K., B. Narasimhan, and T. Hastie, *Elastic Net Regularization Paths for All Generalized Linear Models.* J Stat Softw, 2023. **106**.

154. Madrigal-Aguirre, A., et al., *RCC Atlas Project.* unpublished.

155. Liberzon, A., et al., *The Molecular Signatures Database (MSigDB) hallmark gene set collection.* Cell Syst, 2015. **1**(6): p. 417-425.

156. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

157. Fang, J., et al., *Exploring the crosstalk between endothelial cells, immune cells, and immune checkpoints in the tumor microenvironment: new insights and therapeutic implications.* Cell Death Dis, 2023. **14**(9): p. 586.

158. Perron, G., et al., *Pan-cancer analysis of mRNA stability for decoding tumour post-transcriptional programs.* Commun Biol, 2022. **5**(1): p. 851.
159. Fish, L., et al., *Nuclear TARBP2 Drives Oncogenic Dysregulation of RNA Splicing and Decay.* Mol Cell, 2019. **75**(5): p. 967-981 e9.
160. Perron, G., et al., *A General Framework for Interrogation of mRNA Stability Programs Identifies RNA-Binding Proteins that Govern Cancer Transcriptomes.* Cell Rep, 2018. **23**(6): p. 1639-1650.
161. Engel, K.L., et al., *Mechanisms and consequences of subcellular RNA localization across diverse cell types.* Traffic, 2020. **21**(6): p. 404-418.

CHAPTER 7 APPENDICES

7.1 Supplementary Figures

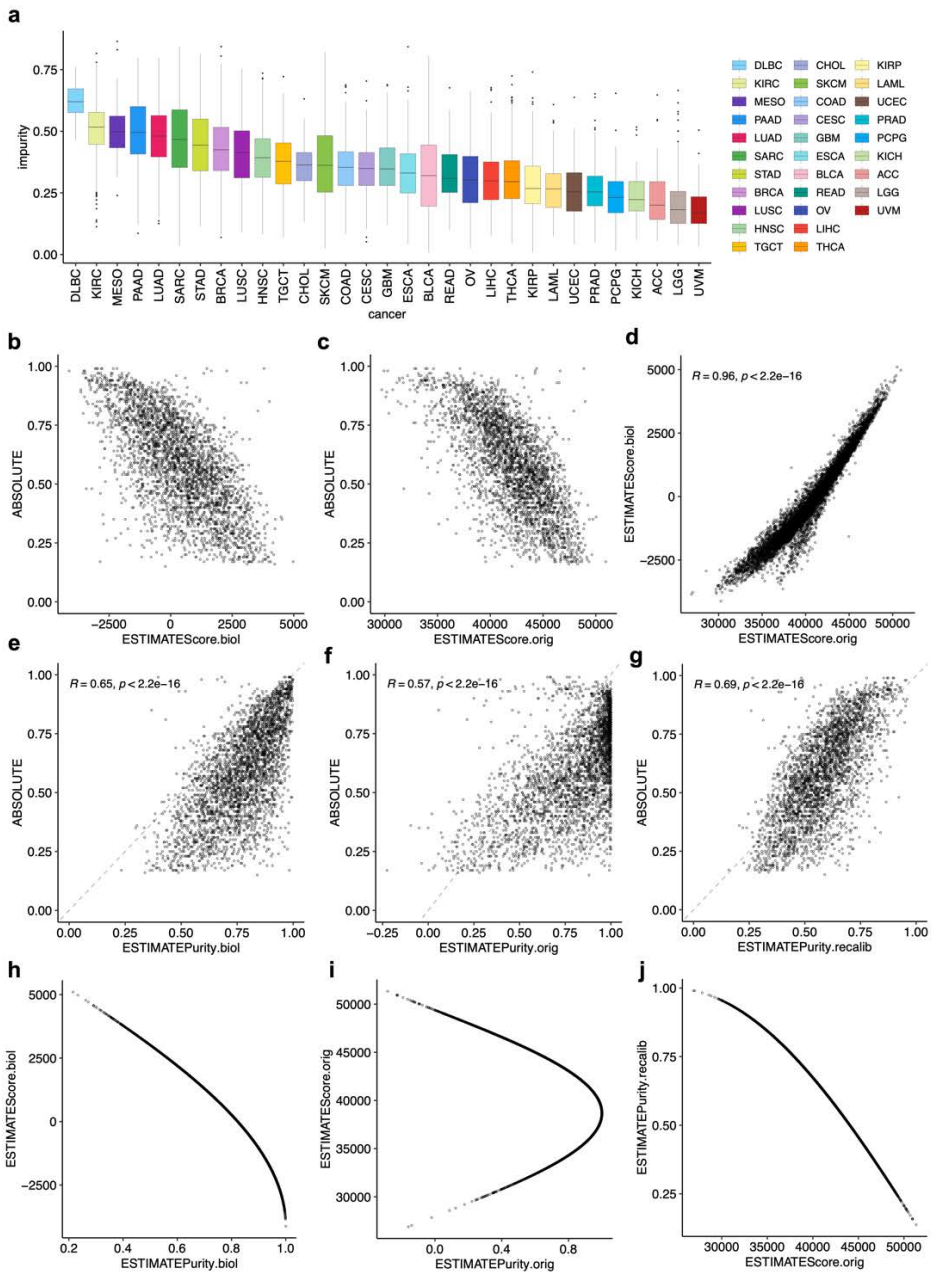


Figure S1 Calculating tumour purity with ESTIMATE

(a) Distribution of impurity of tumour samples across cancer types; **(b)** ESTIMATE Scores retrieved from TCGAbiolinks vs ABSOLUTE purity; **(c)** Re-calculated ESTIMATE Scores using ESTIMATE as implemented vs ABSOLUTE purity; **(d)** Comparison of ESTIMATE Scores from **(c)** and **(b)**. Comparison of ABSOLUTE purities against **(e)** ESTIMATE Tumour purity retrieved from TCGAbiolinks, **(f)** ESTIMATE Tumour purity calculated using the original implementation of ESTIMATE, and **(g)** ESTIMATE Tumour purity obtained using our recalibrated empirical formula. Comparisons of ESTIMATE Tumour purity vs ESTIMATE Scores **(h)** from TCGAbiolinks, **(i)** from the original implementation of ESTIMATE, and **(j)** from ESTIMATE with our recalibrated formula. LGG = low-grade glioma; GBM = Glioblastoma multiforme; HNSC = Head and Neck squamous cell carcinoma; ESCA = Esophageal carcinoma; THCA = Thyroid carcinoma; DLBC = Lymphoid Neoplasm Diffuse Large B-cell Lymphoma; BRCA = Breast invasive carcinoma; LUAD = Lung adenocarcinoma; LUSC = Lung squamous cell carcinoma; MESO = Mesothelioma; LIHC = Liver hepatocellular carcinoma; STAD = Stomach adenocarcinoma; PCPG = Pheochromocytoma and Paraganglioma; ACC = Adrenocortical carcinoma; KICH = Kidney Chromophobe; KIRC = Kidney renal clear cell carcinoma; KIRP = Kidney renal papillary cell carcinoma; CHOL = Cholangiocarcinoma; PAAD = Pancreatic adenocarcinoma; COAD = Colon adenocarcinoma; BLCA = Bladder Urothelial Carcinoma; READ = Rectum adenocarcinoma; PRAD = Prostate adenocarcinoma; CESC = Cervical squamous cell carcinoma and endocervical adenocarcinoma; UCEC = Uterine Corpus Endometrial Carcinoma; OV = Ovarian serous cystadenocarcinoma; TGCT = Testicular Germ Cell Tumours; LAML = Acute Myeloid Leukemia; SARC = Sarcoma; UVM = Uveal Melanoma; and SKCM = Skin Cutaneous Melanoma.

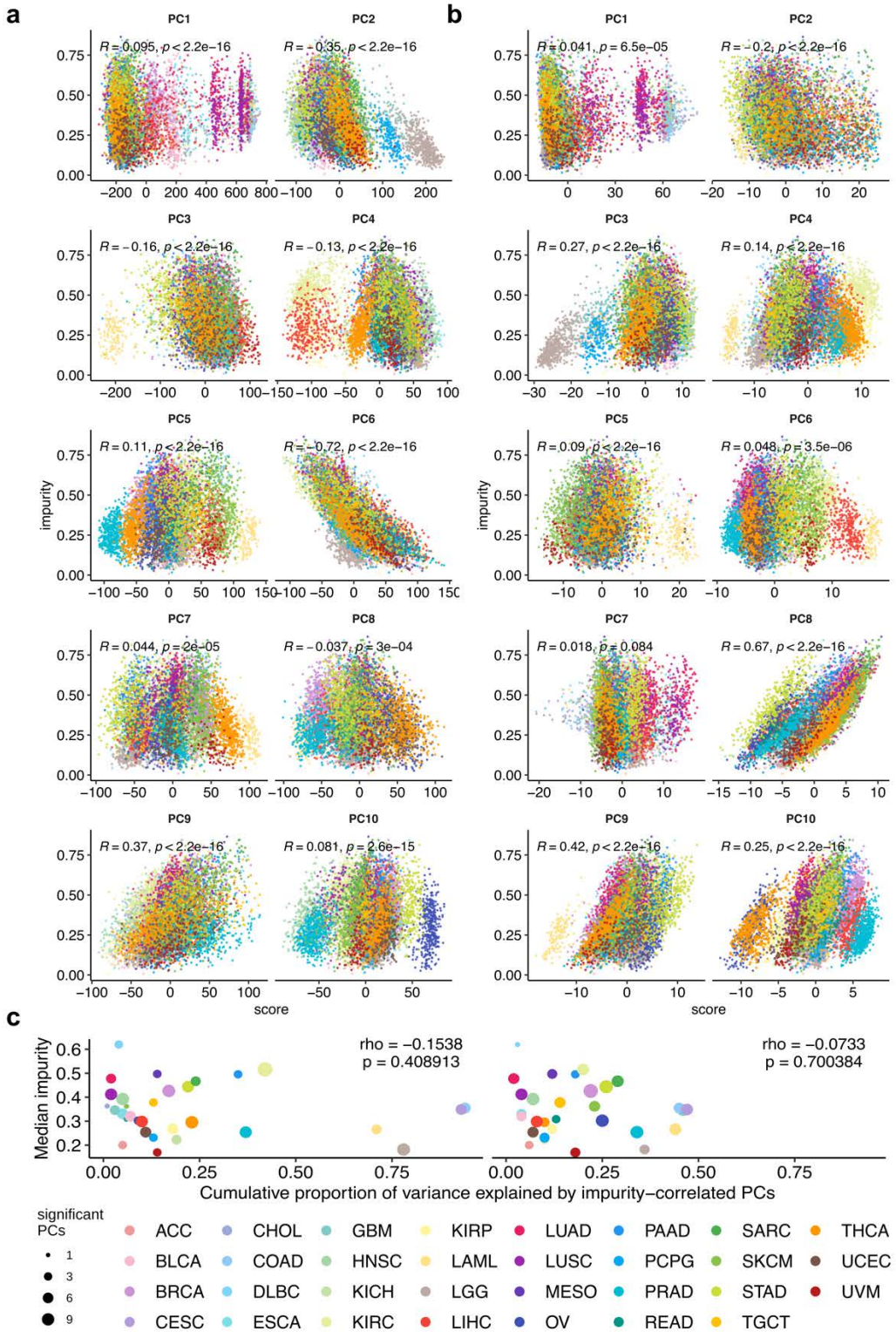


Figure S2 Comparison of pan-cancer principal components with impurity

Pairwise comparison of impurity with the top 10 PC scores from the pan-cancer **(a)** expression matrix and **(b)** splicing matrix. (c) Comparison of the total variance explained by the PCs with significant correlations with impurity (Bonferroni-adjusted P value<0.05) and the median impurity of each cancer type. Top right annotations indicate the P value of a Spearman correlation. Cancer name abbreviations remain the same as in **Figure S1.**

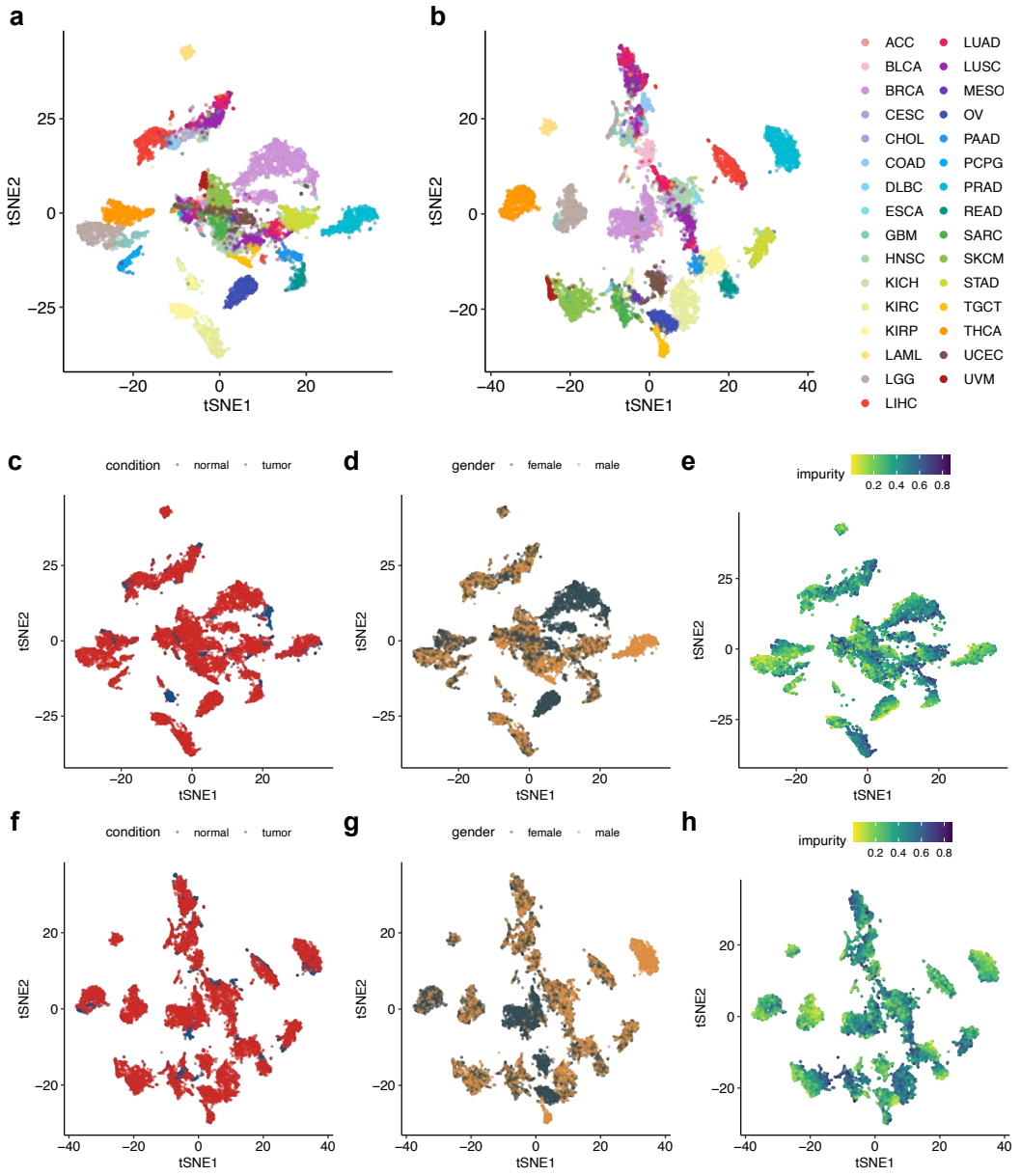


Figure S3 Sample covariates on expression and splicing embeddings

Pan-cancer tSNE embedding on (a) splicing matrix of PSI values and (b) gene expression matrix included for reference. Splicing t-SNE embeddings colored by (c) experimental condition (tumour or normal), (d) reported biological sex, (e) sample impurity. Expression t-SNE embeddings colored by (f) experimental condition (g) reported biological sex, (h) sample impurity. All normal samples have an impurity of zero. Cancer name abbreviations remain the same as in **Figure S1.**

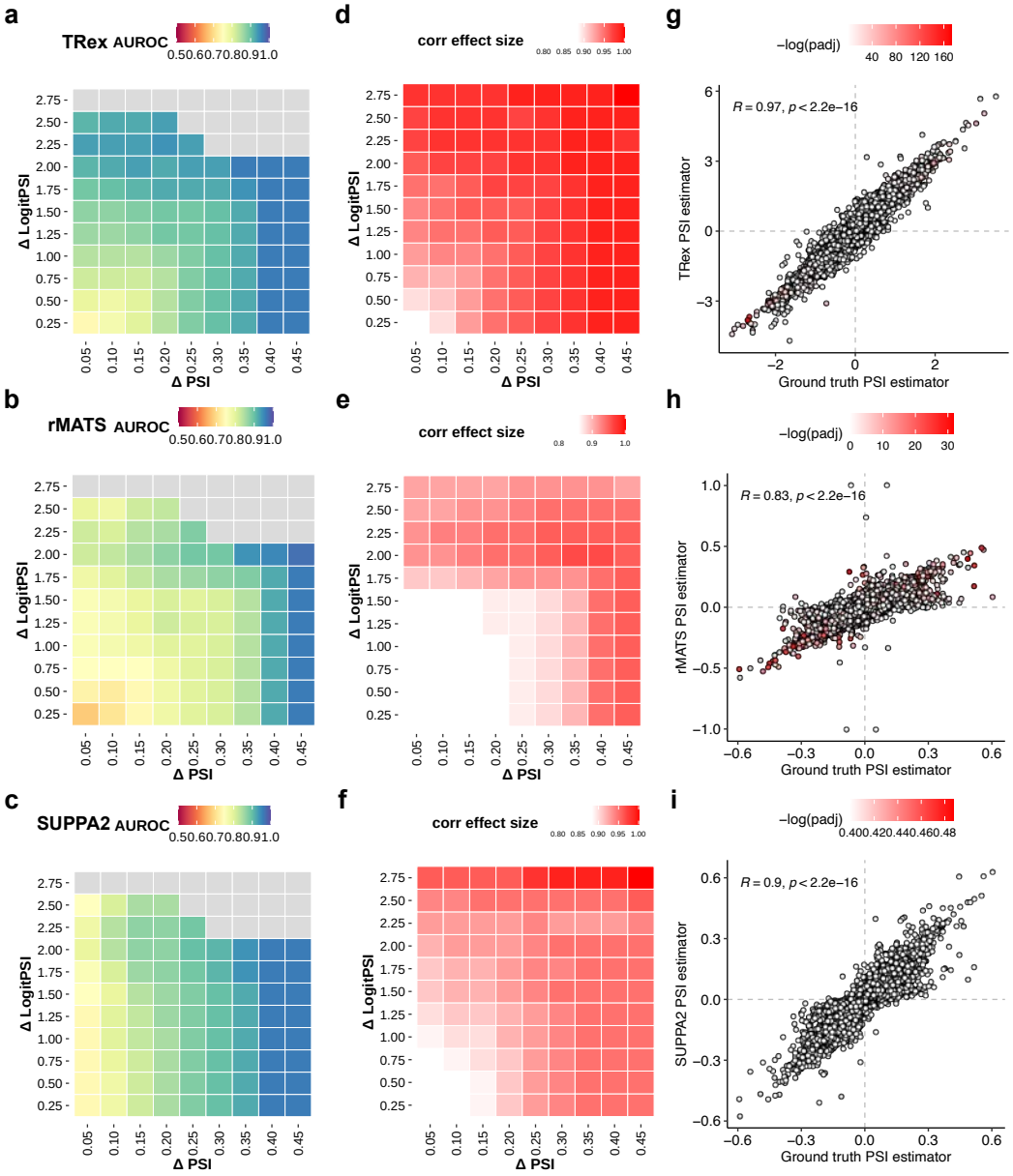


Figure S4 Performance per method in the simulation benchmark

Performance measured as area under the receiver operator curve (AUROC) along a grid of thresholds on PSI and Δ$logit$PSI used to define the ground truth for (a) TRex, (b) rMATS and (c) SUPPA2. Pearson correlation coefficients between the ground truth and predicted effect size estimators of DAS along the same grid of thresholds for (d) TRex, (e) rMATS, and (f) SUPPA2. Global correlation of the ground truth effect sizes and the average predicted effect size over 25 random simulations for (g) TRex, (h) rMATS, and (i) SUPPA2. padj = Bonferroni adjusted P-values.

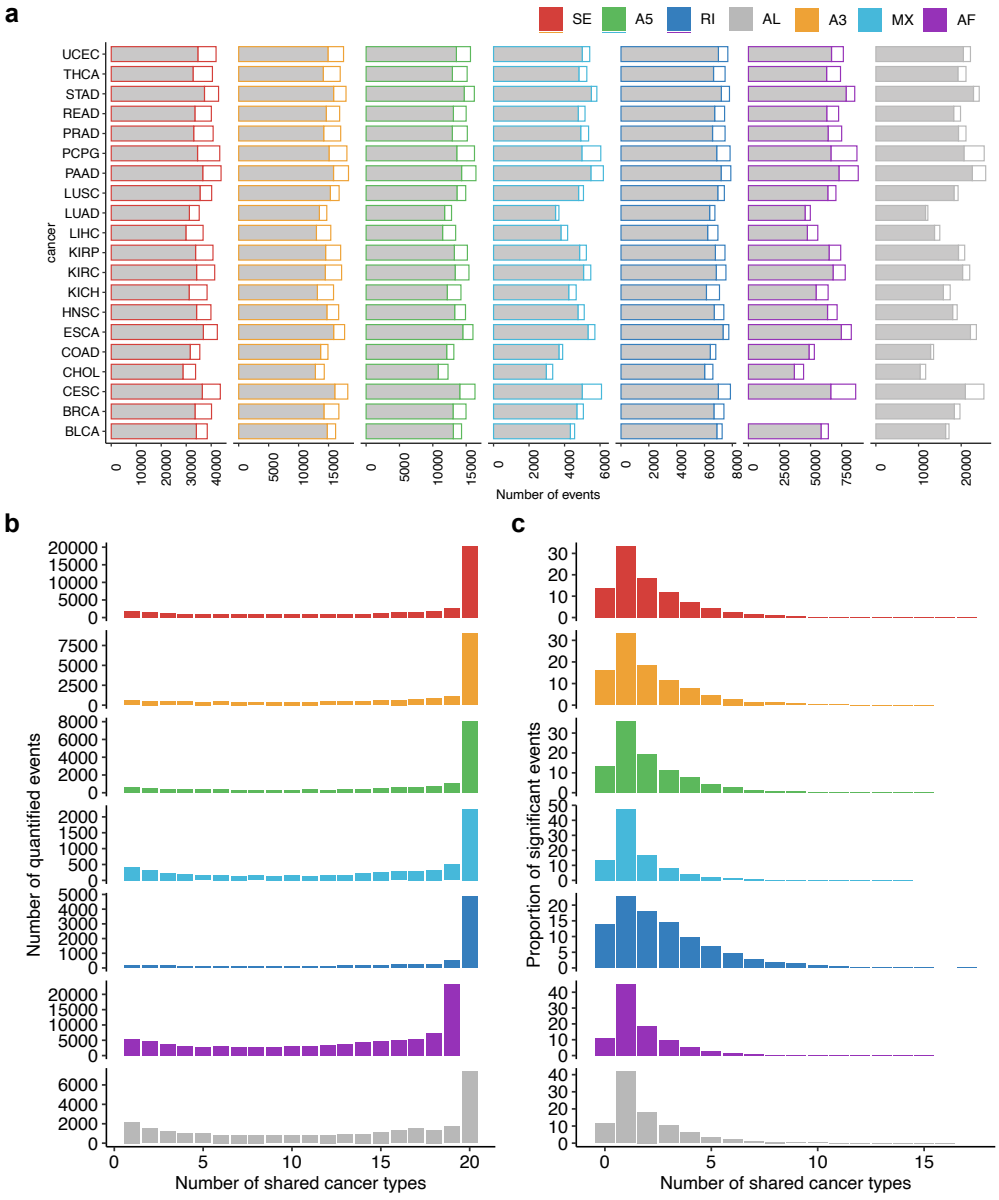


Figure S5 Summary of tumour-associated events

(a) Distribution of retained (solid gray) and discarded (white) events based on sample PSI flags across cancer types; (b) Number of events of each type retained per number of cancers; (c) number of events of each type that were significantly associated with tumours (adjusted p-value<0.05) per number of cancer types. Cancer name abbreviations remain the same as in **Figure S1.**

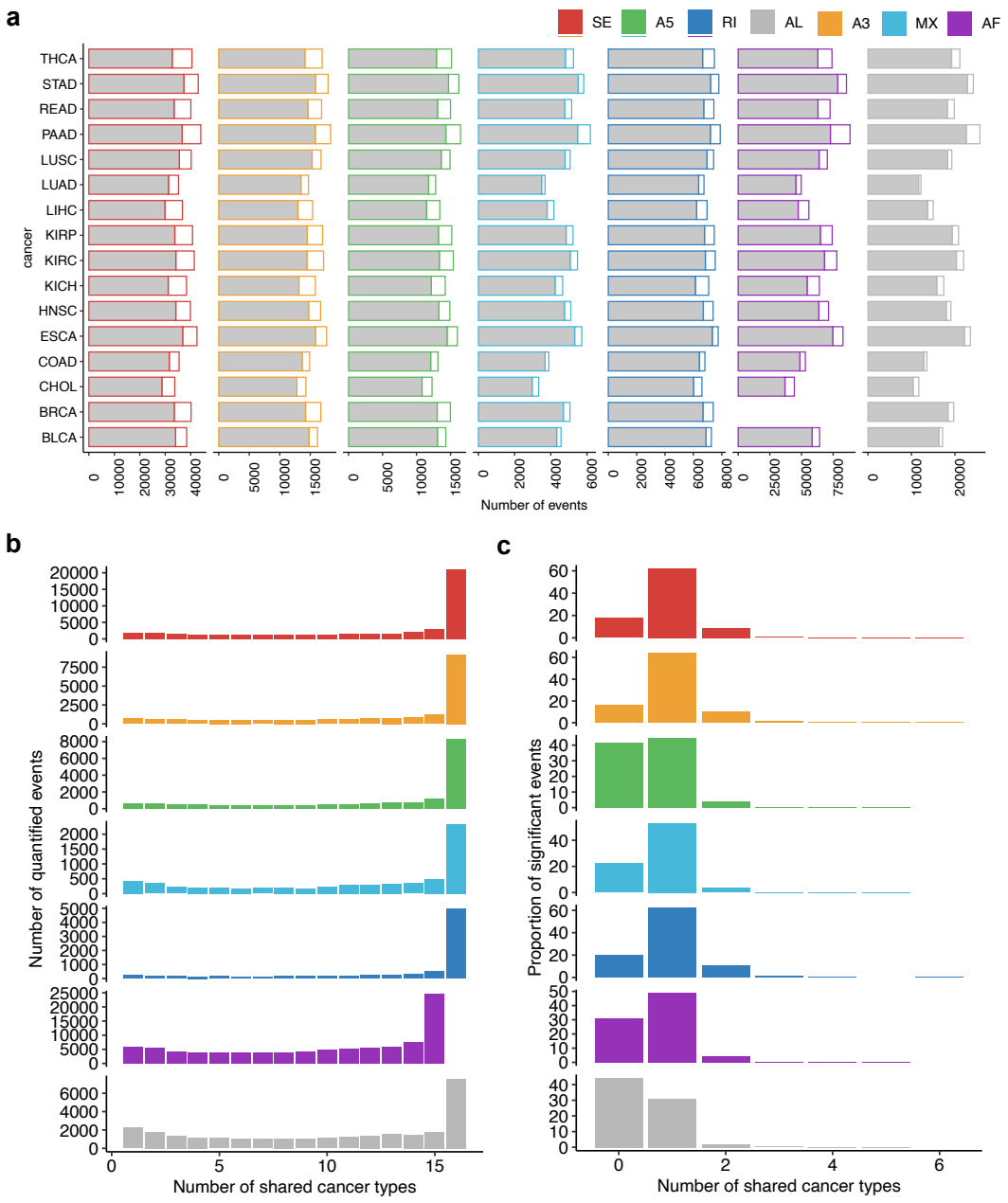


Figure S6 Summary of stage associated events

(a) Distribution of retained (solid gray) and discarded (white) events based on sample PSI flags across cancer types; (b) Number of events of each type retained per number of cancers; (c) number of events of each type that were significantly associated with tumours (adjusted p-value<0.05) per number of cancer types. Cancer name abbreviations remain the same as in **Figure S1.**

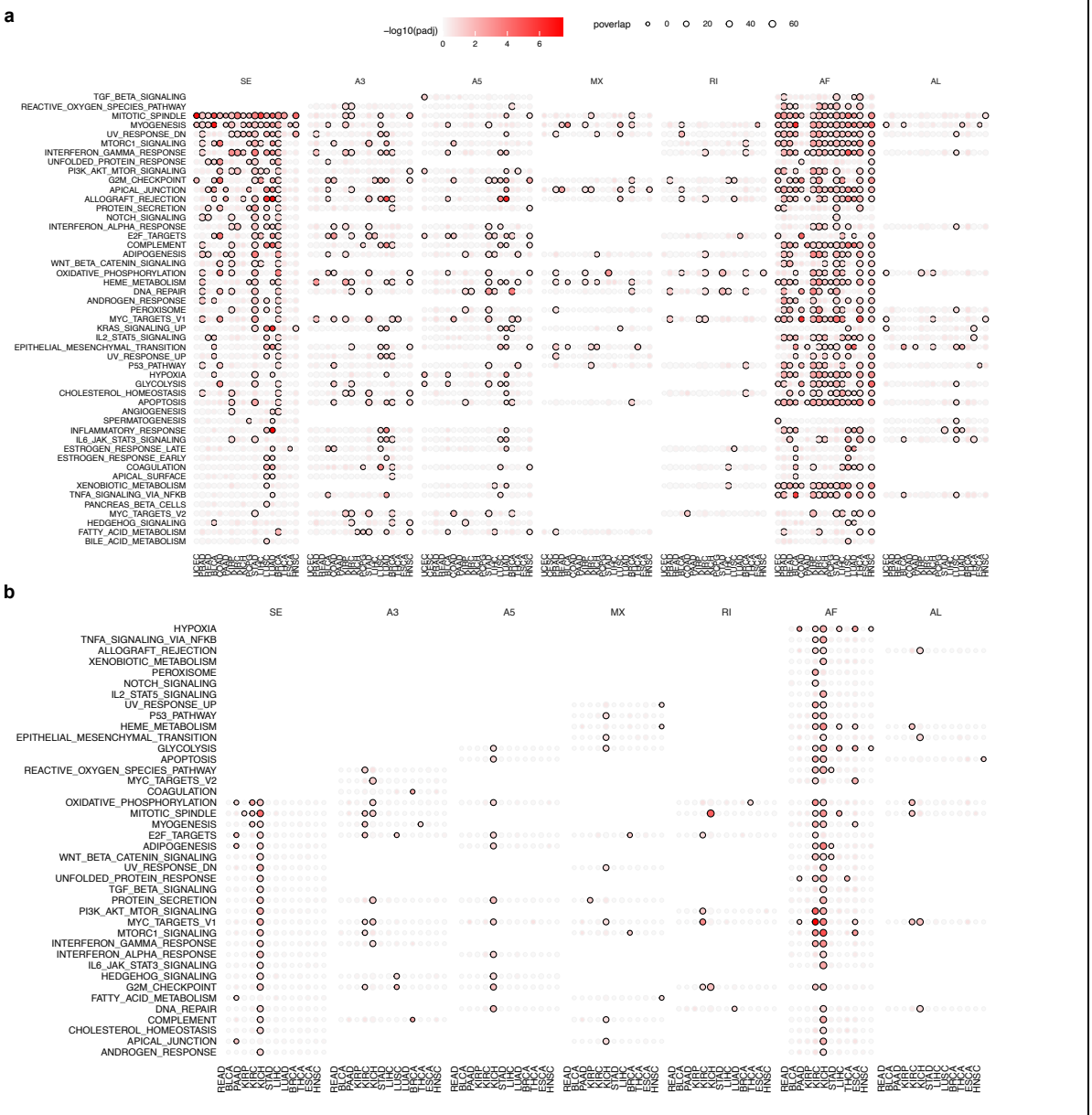


Figure S7 GORA of Hallmark gene sets across of DAS events

Enrichment of Hallmark pathways from MSigDB in genes with different types of **(a)** tumor-associated and **(b)** stage-associated DAS events. Solid black outlines indicate significant overrepresentation (FDR<0.2). Cancer names abbreviations detailed in **Figure S1**. Event type abbreviations described in the main text.

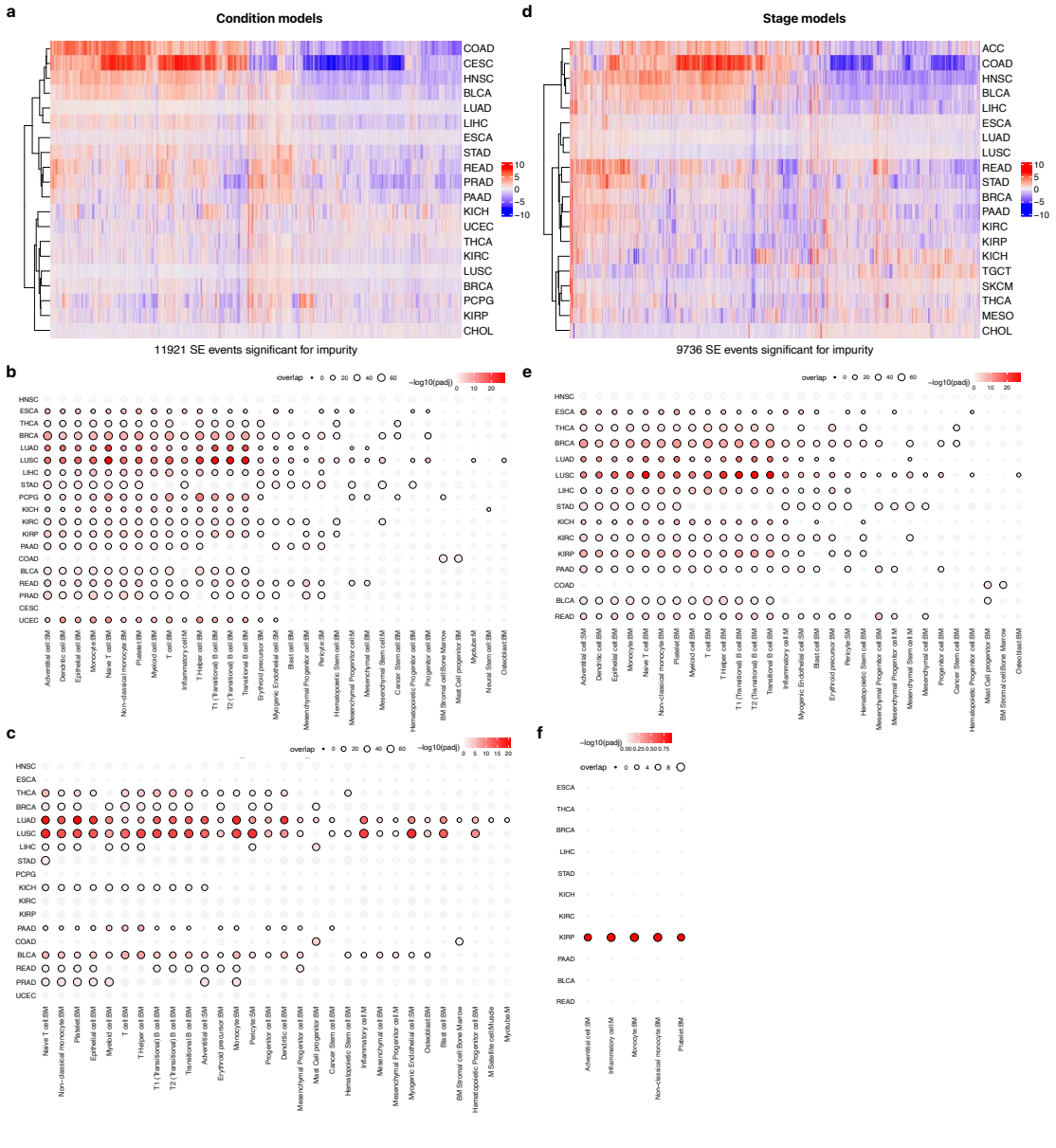


Figure S8 Impurity associated changes of SE events from both models

(a) Impurity-associated $\Delta logitPSI$ changes in SE events derived from the tumour model. GORA of immune cell markers in genes with at least one event significantly associated with **(b)** impurity in the tumour model (adjusted P-value of 0.05, and $|\Delta logitPSI| > 0.5$); and **(c)** the tumour coefficient from the same model. **(d)** Impurity-associated $\Delta logitPSI$ changes in SE events derived from the stage model. **(e)** GORA of immune cell markers in genes with events significantly associated with **(e)** impurity in the stage model; and **(f)** the stage coefficient from the same model. Cancer name abbreviations remain the same as in **Figure S1.**

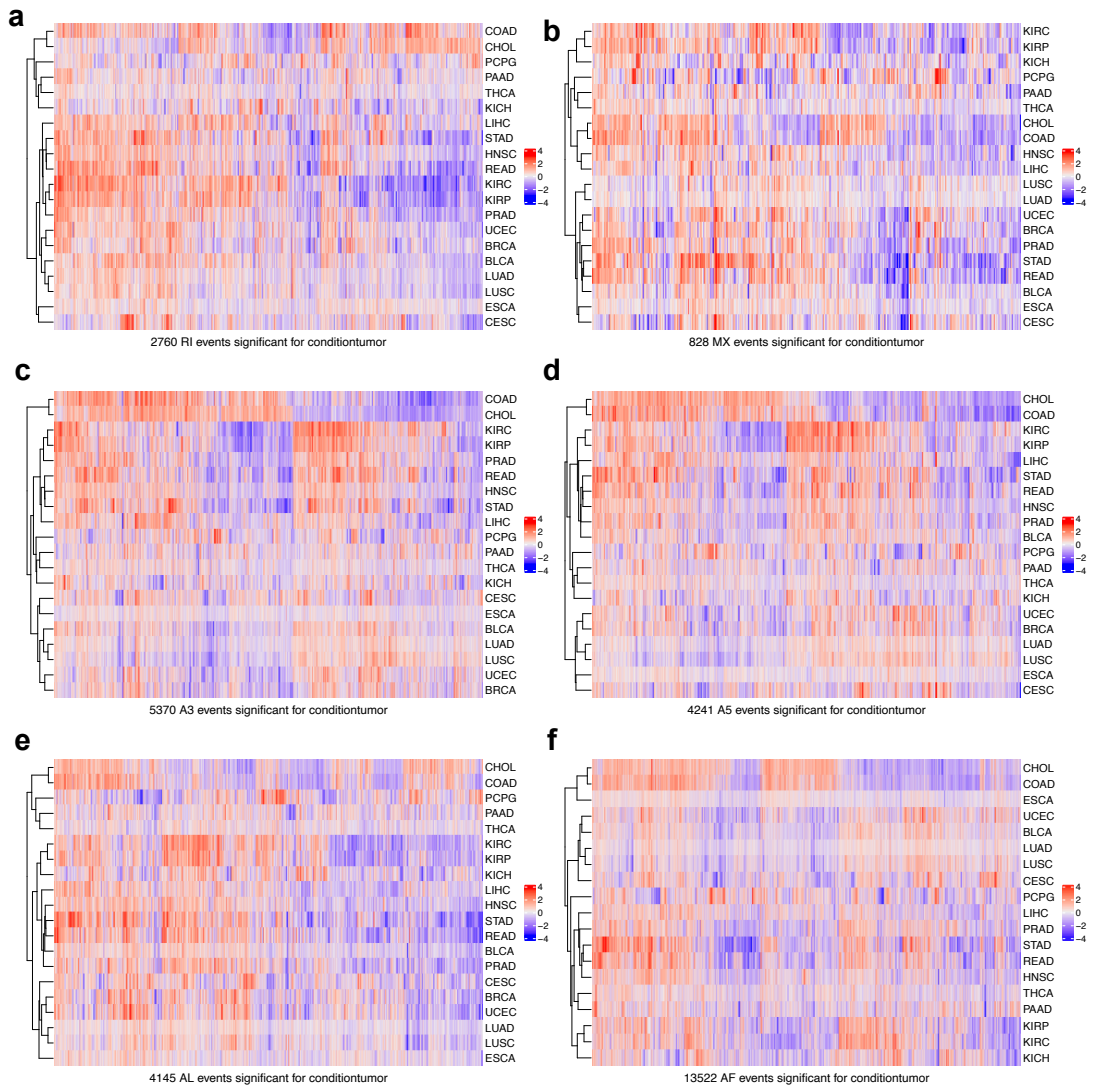


Figure S9 Clustering of tumour-associated changes per AS mechanism

Heatmaps showing estimated tumour-associated $\Delta logit\text{PSI}$ for **(a)** RI, **(b)** MX, **(c)** A3, **(d)** A5, **(e)** AL and **(f)** AF events. All heatmaps were clustered using the Ward method with a Pearson correlation metric. Cancer name abbreviations remain the same as in **Figure S1.**

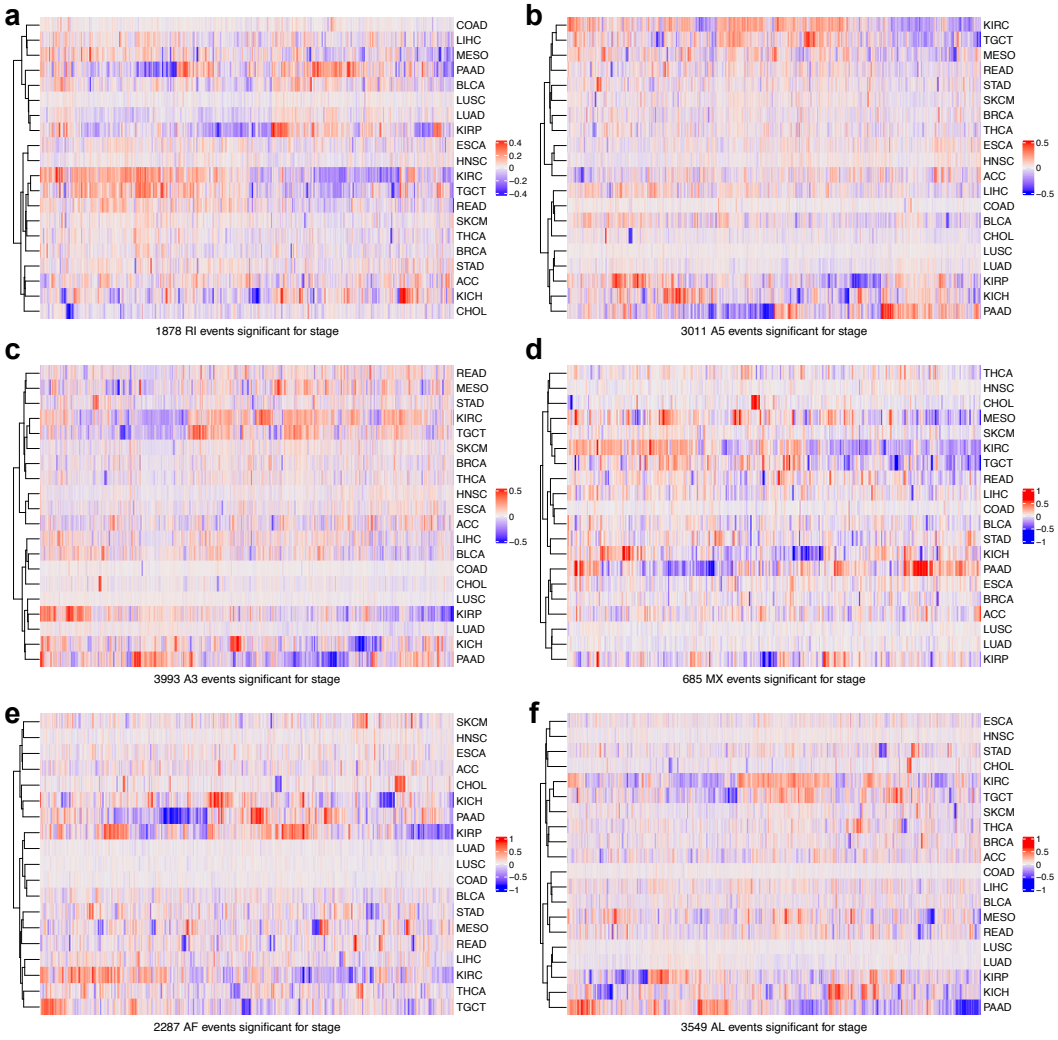


Figure S10 Clustering of stage-associated changes

Heatmaps showing stage $\Delta logit$PSI for **(a)** RI, **(b)** MX, **(c)** A3, **(d)** A5, **(e)** AL and **(f)** AF events. All heatmaps were clustered using the Ward method with a Pearson correlation metric. Cancer name abbreviations remain the same as in **Figure S1.**

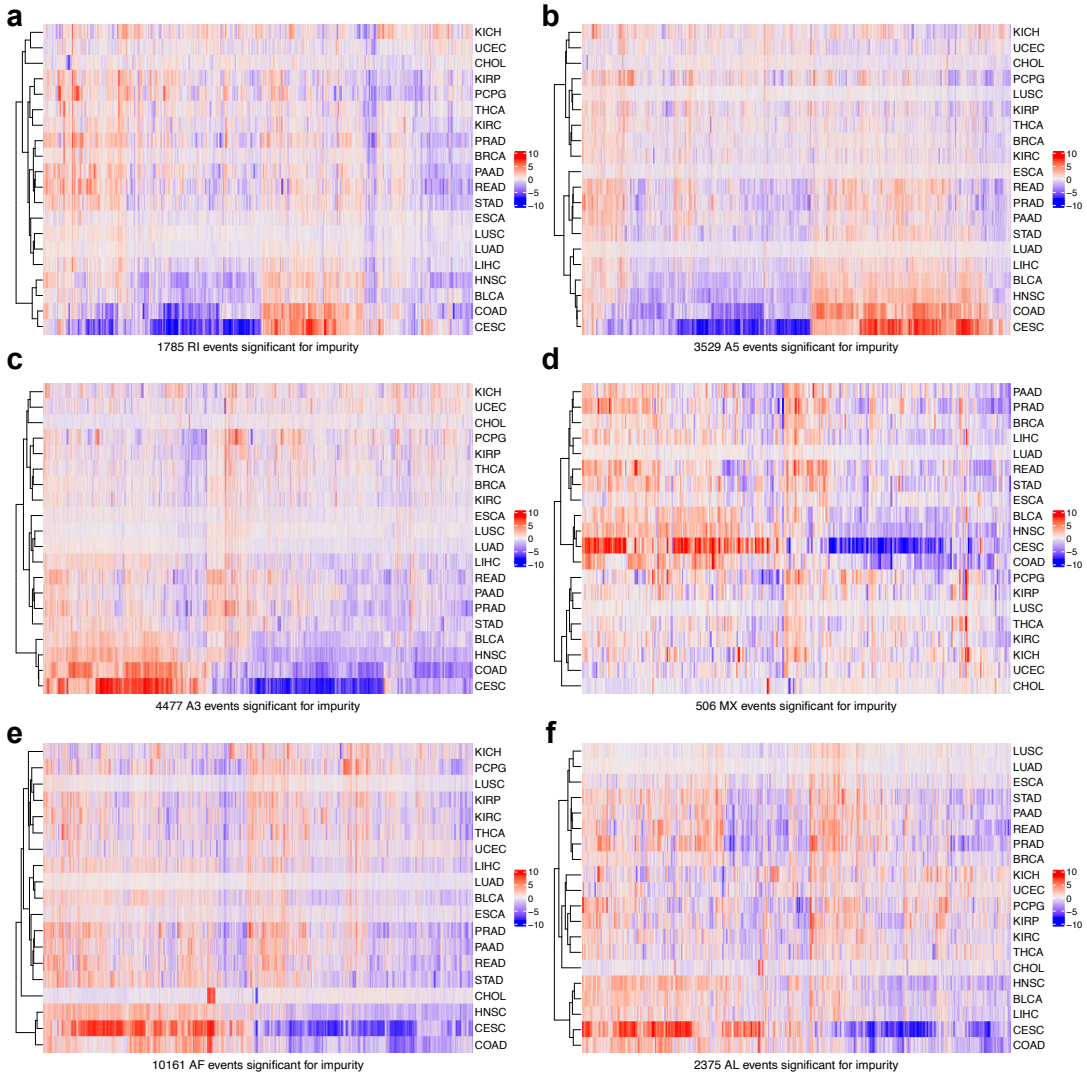


Figure S11 Clustering of impurity-associated changes in the tumour models

Heatmaps showing impurity $\Delta logit$PSI for **(a)** RI, **(b)** MX, **(c)** A3, **(d)** A5, **(e)** AL and **(f)** AF events in the tumour models. All heatmaps were clustered using the Ward method with a Pearson correlation metric. Cancer name abbreviations remain the same as in **Figure S1.**

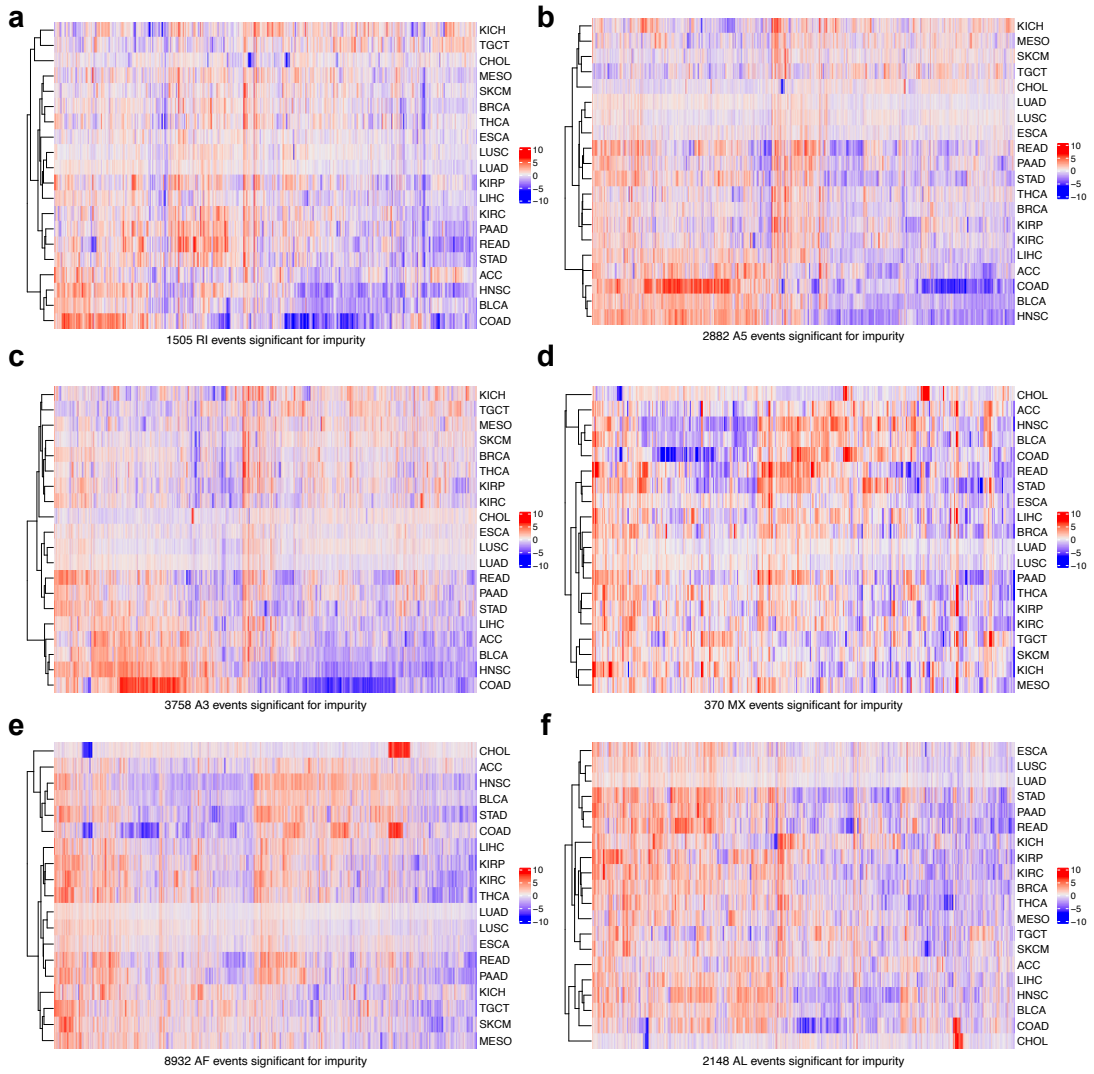


Figure S12 Clustering of impurity-associated changes from the stage models

Heatmaps showing impurity $\Delta logit$PSI for **(a)** RI, **(b)** MX, **(c)** A3, **(d)** A5, **(e)** AL and **(f)** AF events in the stage models. All heatmaps were clustered using the Ward method with a Pearson correlation metric. Cancer name abbreviations remain the same as in **Figure S1.**

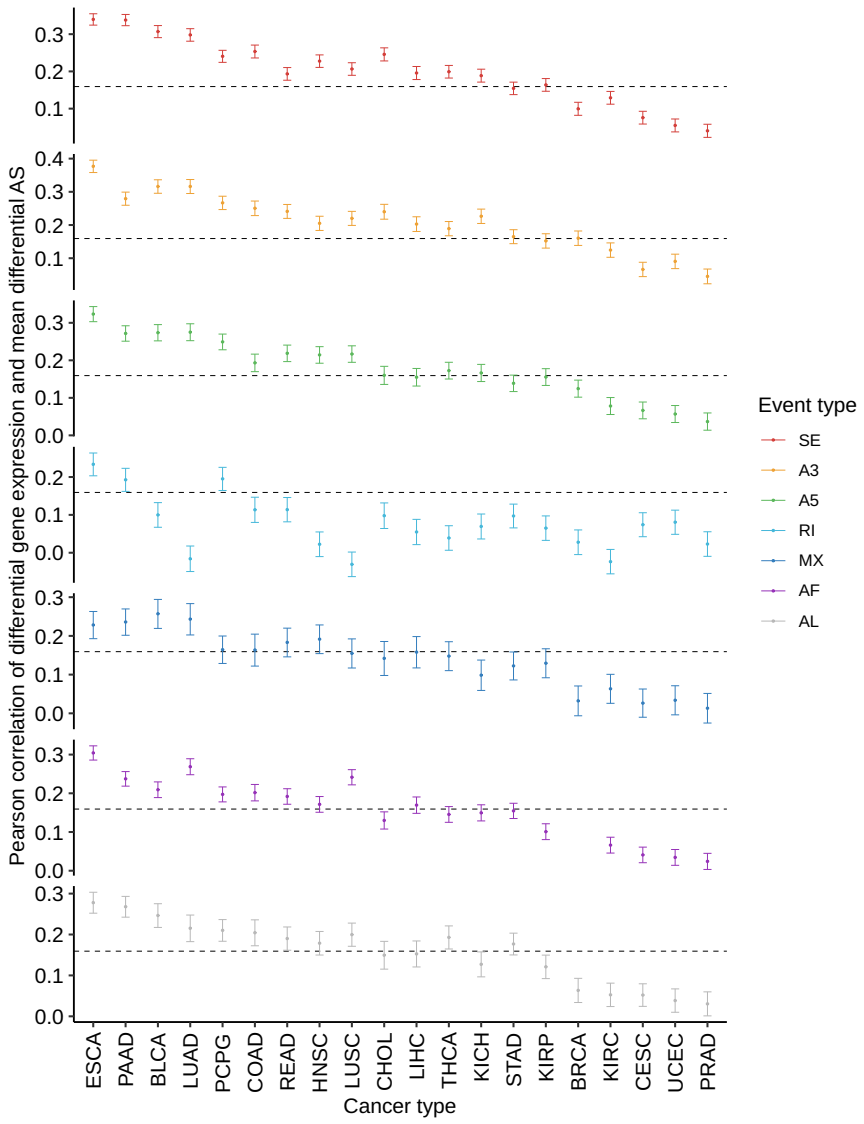


Figure S13 Comparison of differential alternative splicing effects and differential gene expression changes

Correlations between effect sizes of tumour-associated DGE and mean DAS effects. Event-wise DAS estimates are collapsed into a gene-level metric by taking the average $\Delta logit\text{PSI}$ of all the events of the same type within a given gene. DGE effects correspond to a gene-level DESeq2 model contrasting tumour vs normal samples while accounting for impurity, age, and sex. Comparisons performed for every cancer type separately.

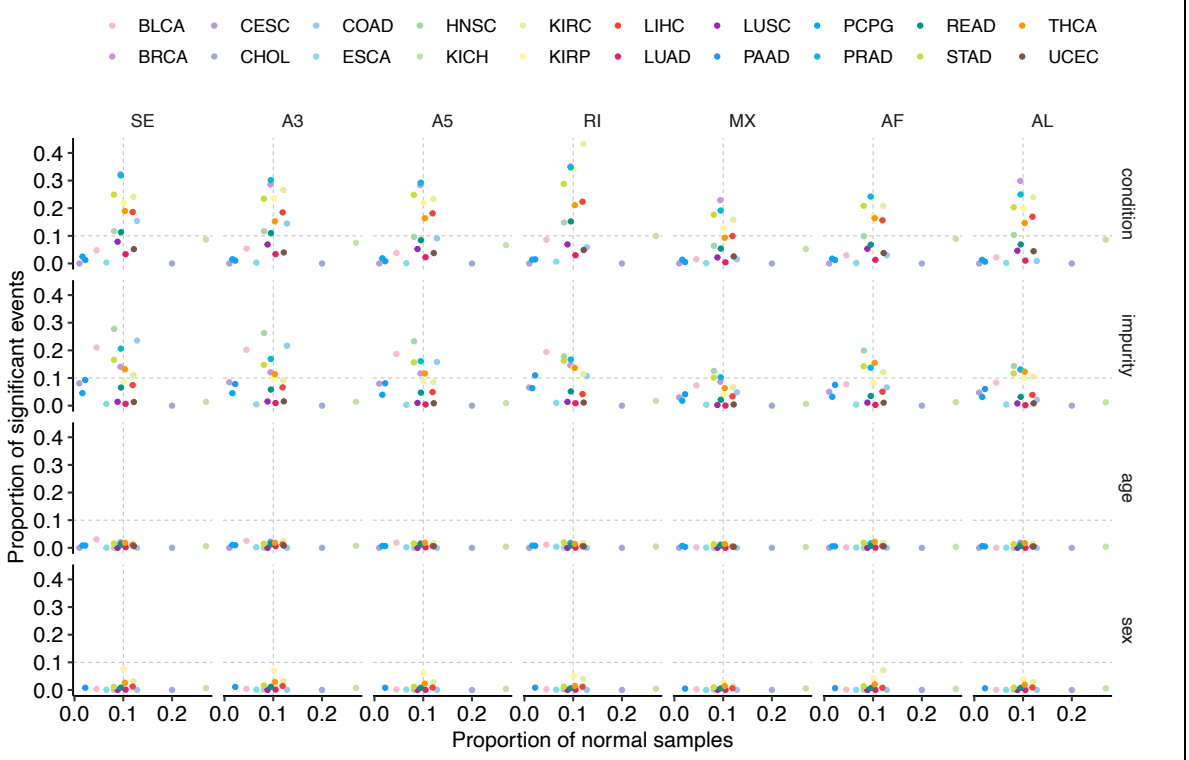


Figure S14 Relationship between fraction of normal samples and significant events

Proportion of normal samples vs the fraction of events, per type (columns), significantly associated (FDR<0.05) with each variable (rows) in the tumour models. Each dot represents one cancer type; name abbreviations shown in **Figure S1.**

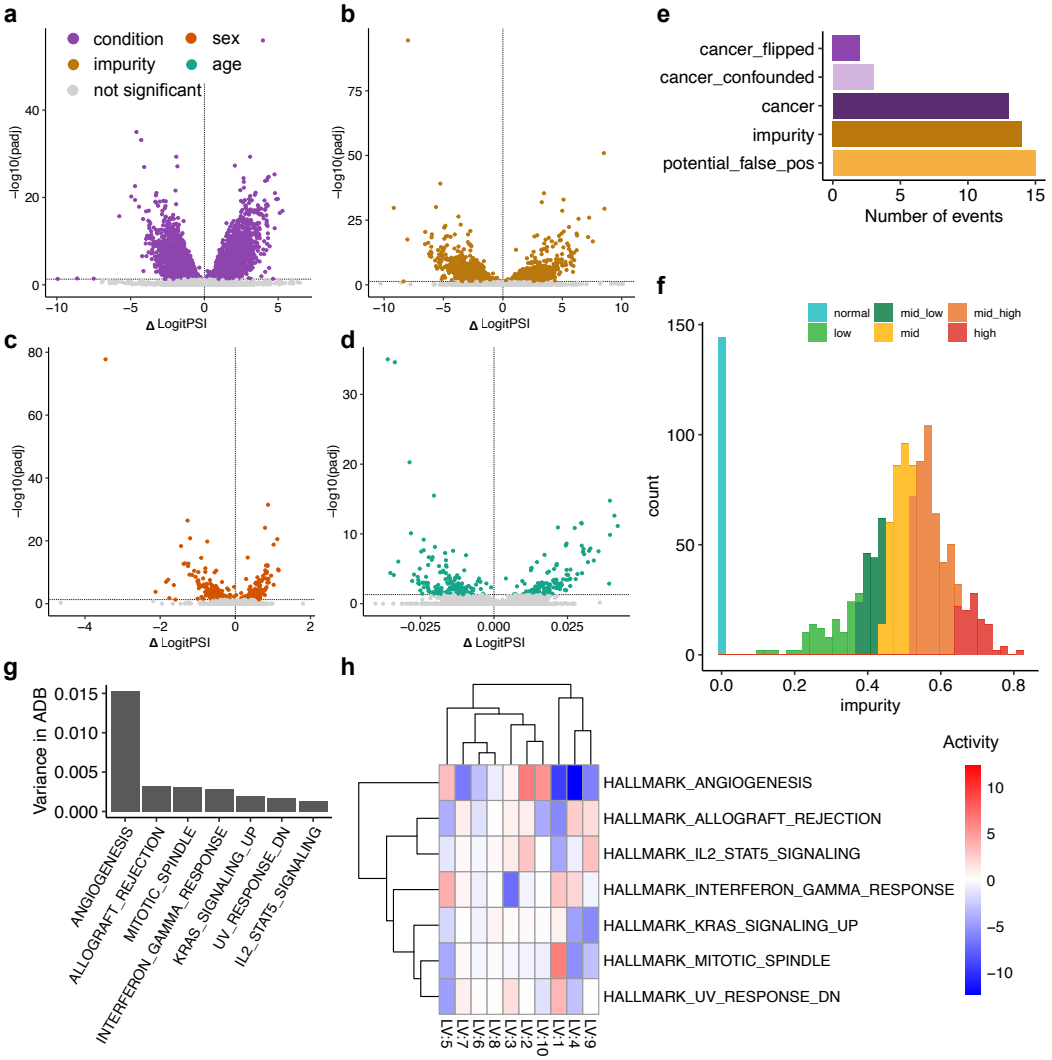


Figure S15 KIRC case study

Volcano plots of **(a)** tumour, **(b)** impurity, **(c)** sex, and **(d)** age associated SE events. **(e)** Number of events reported in the ASCancer database per event category. **(f)** Histogram of number of samples per impurity bin. **(g)** Variance explained by each Hallmark pathway and **(h)** projection of pathway activities onto the latent variables learned by GEDI.

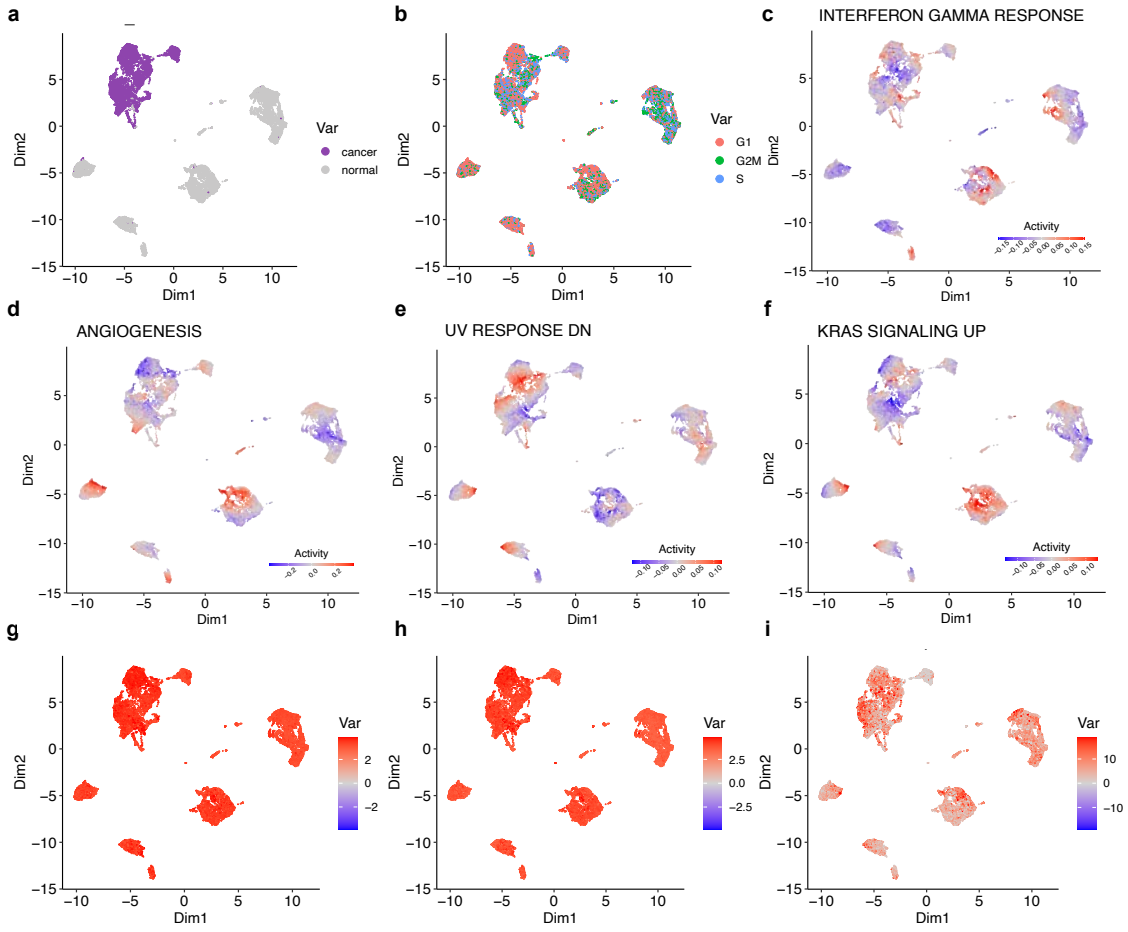


Figure S16 Additional variables measured in the single cell ccRCC dataset

(a) InferCNV status calls. **(b)** Assigned cell cycle phases. GEDI-inferred activities of Hallmark pathways: **(c)** *Interferon Gamma Response*, (d) *Angiogenesis*, **(e)** *UV Response (down)*, and **(f)** *KRAS signaling (up)*. **(c)** Number of detected genes in logarithmic scale; **(d)** total UMIs per cell in logarithmic scale; and **(e)** percentage of mitochondrial genes.

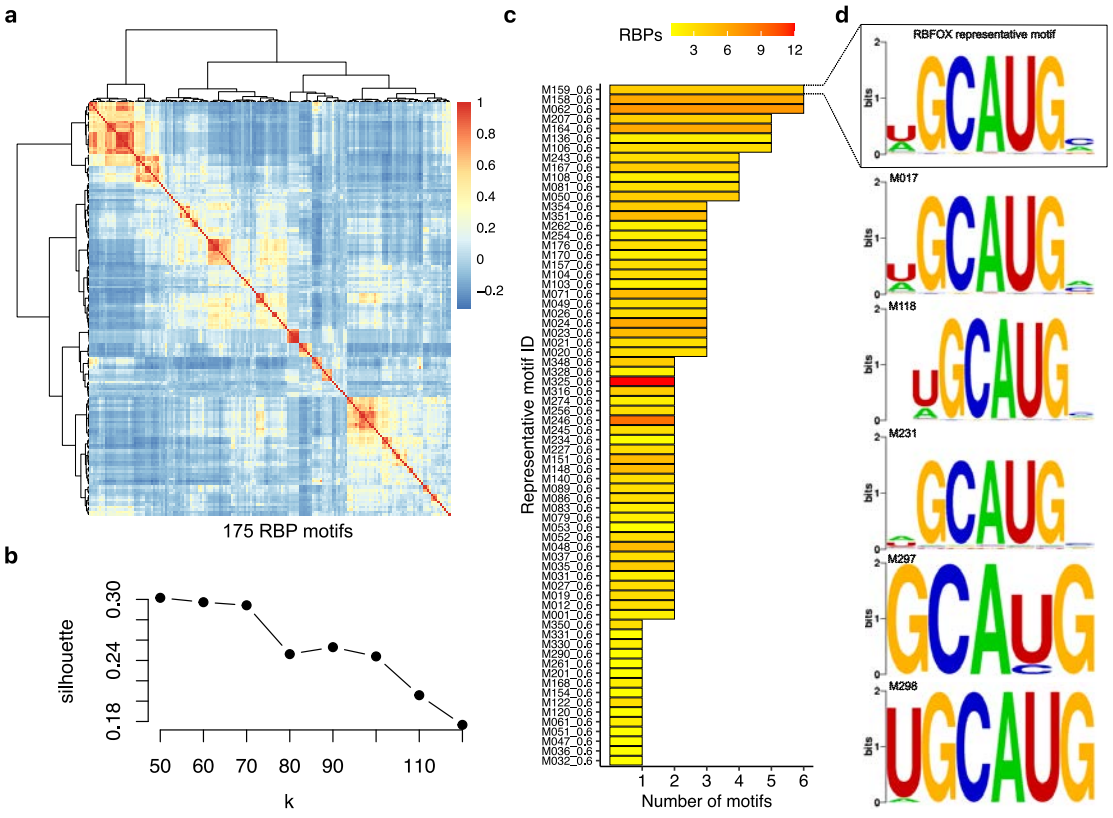


Figure S17 RBP motif clustering

(a) Heatmap of clustered pairwise Pearson correlations of the logarithm of binding affinities of each motif over a set of 50,000 random sequences. **(c)** Histogram of cluster sizes, color scale represents the number of RBPs associated with each representative motif. **(d)** motif logos downloaded from CISBP (http://cisbp-rna.ccbr.utoronto.ca) of the motifs in the RBFOX cluster.

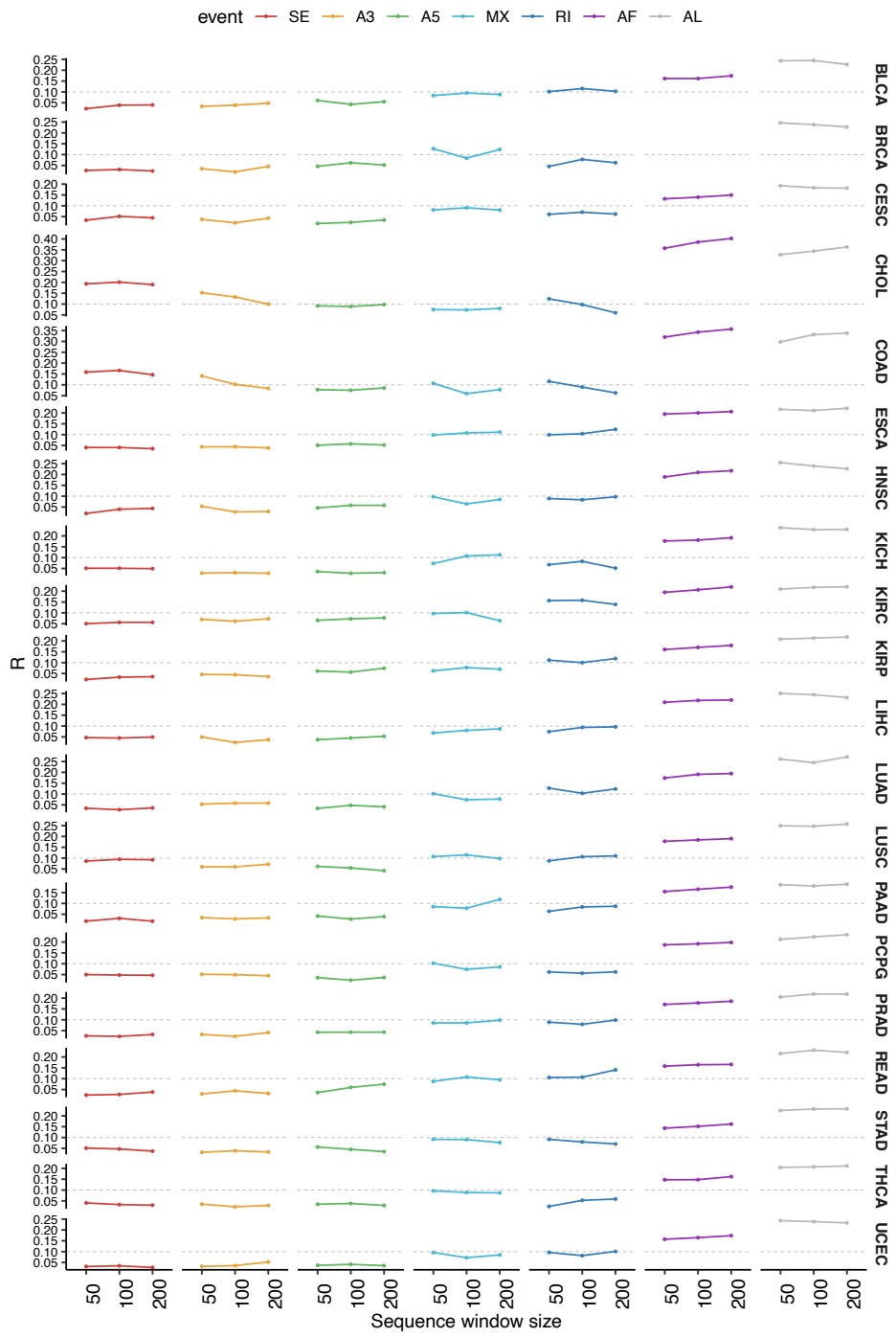


Figure S18 Performance of RBP DSR models

Average Pearson correlation coefficients of 10-fold cross validation tests during model fitting. Cancer name abbreviations shown in **Figure S1**.

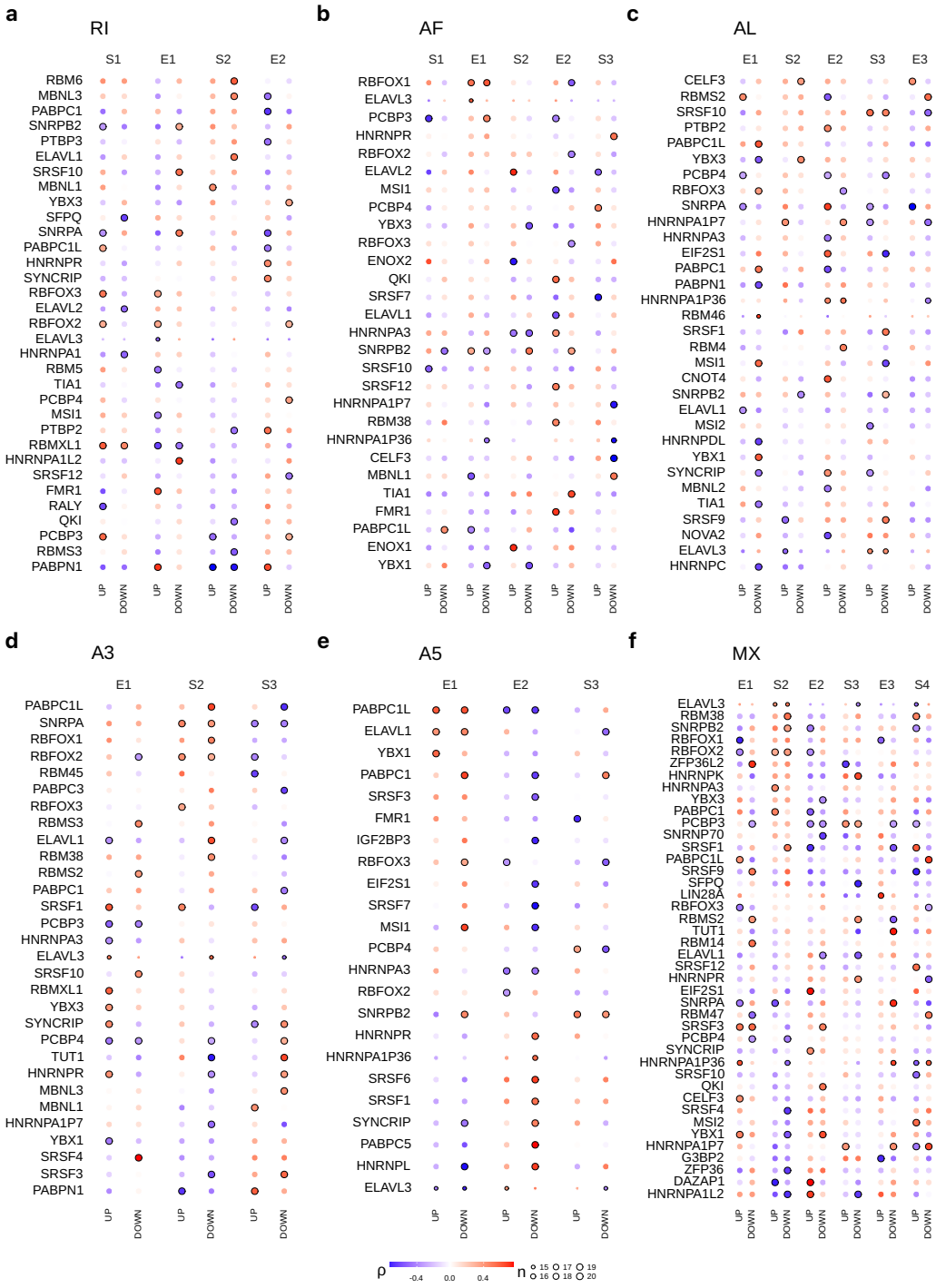


Figure S19 Correlations between RBPs DSR and expression

Pearson correlation coefficients between each RBP's differential expression in tumour vs normal, and its DSR coefficient in **(a)** RI, **(b)** AF, **(c)** AL, **(d)** A3, **(e)** A5 and **(f)** MX events per position. E = end, S = start

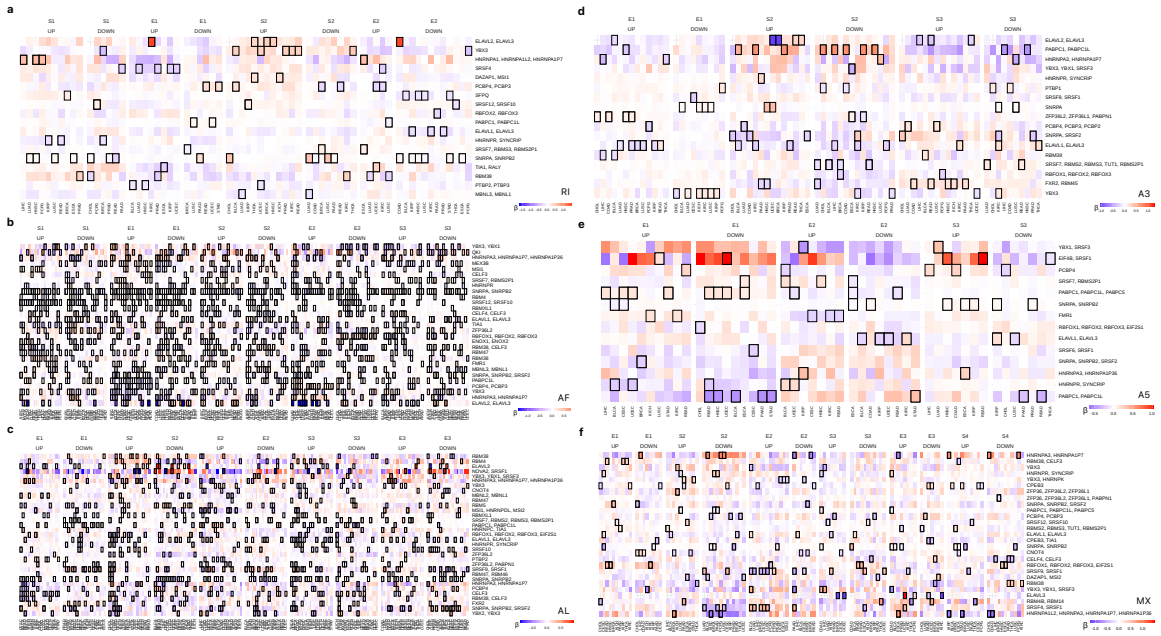


Figure S20 DSR of RBPs per event type

Coefficients of differential RBP binding per splice region across cancer types for **(a)** RI, **(b)** AF, **(c)** AL, **(d)** A3, **(e)** A5 and **(f)** MX events. Solid outlines indicate significant coefficients (nominal P value < 0.01).

7.2 Supplementary Notes

7.2.1 TRex design matrix implementation in DESeq2

Assuming an experiment with 2 experimental conditions T and N with two replicates each, event counts of type A and C are calculated separately in each sample and then merged back in an augmented expression matrix. This augmented matrix now includes two entries per sample, one corresponding to each count type (A or C). The count type is then specified as an additional variable in the design matrix. If we express Eqn. 3 and Eqn. 4 from Section 2.1 in terms of this design matrix, the model for any event k in the augmented matrix becomes

$$\log(\bar{\mu}_k) = D \times \beta$$

Where $\bar{\mu}_k$ is now the vector of both count types corresponding to event k in all samples (**Figure S21**).

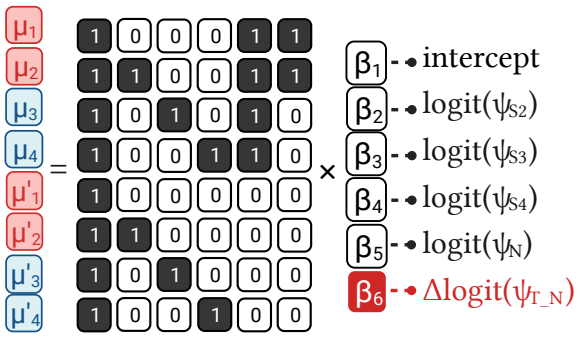


Figure S21 TRex design matrix and coefficients

In an example experiments with two tumour samples (red) and two normal samples (blue) the counts for each outcome of a given event in every sample are separated and treated as if they were independent measurements (vector of μ and μ'). The count type is then added to the formula used for DESEq2 so that the design matrix contains the following columns in the order shown: intercept, sample 2, sample 3, sample 4, count type, interaction between condition and count type.

Where β is the vector of model coefficients:

$$\beta_1 = \log(\gamma'_{k1}) + \log(q'_k)$$
$$\beta_2 = \log(\gamma'_{k2}) + \log(\gamma'_{k1})$$
$$\beta_3 = \log(\gamma'_{k3}) + \log(\gamma'_{k1})$$
$$\beta_4 = \log(\gamma'_{k4}) + \log(\gamma'_{k1})$$
$$\beta_5 = \log(r_{k,4}) + \log(q_k) - \log(q'_k)$$
$$\beta_6 = \log(r_{k,1}) - \log(\mathrm{r}_{k,4})$$

Given that $r_{k,1} = {}^{\gamma}/_{\gamma'}$ in the tumour condition and $r_{k,4} = {}^{\gamma}/_{\gamma'}$ in the normal group, we can express b_6 in terms of PSI (shown in **Figure 4**) as

$$b_6 = \mathrm{logit}(\Psi_{k,\rho(T)}) - \mathrm{logit}(\Psi_{k,\rho(N)})$$

Therefore b_6 directly becomes the effect size estimator of differential splicing of event k between conditions tumour and normal.

7.3 Supplementary Tables

Table S1 R package versions			
R Package	***Version***	***R Package***	***Version***
AnnotationHub	3.0.2	HDF5Array	1.20.0
aplot	0.1.9	hms	1.1.3
binr	1.1	lme4	1.1-29
BiocParallel	1.26.2	Matrix	1.5-1
biomaRt	2.48.3	Matrix	1.3-4
caret	6.0-93	matrixStats	0.63.0
ComplexHeatmap	2.8.0	msigdbr	7.5.1
cowplot	1.1.1	optparse	1.7.1
data.table	1.14.6	org.Hs.eg.db	3.13.0
DESeq2	1.32.0	parallel	4.1.2
doParallel	1.0.17	patchwork	1.1.2
dplyr	1.0.9	pheatmap	1.0.12
DT	0.26	plotly	4.10.0
EnhancedVolcano	1.13.2	plyr	1.8.8
estimate	1.0.13	precrec	0.12.9
fgsea	1.18.0	purrr	0.3.5
fishpond	1.8.0	readr	2.1.4
furrr	0.3.1	rstatix	0.7.0
GEDI	0.0.0.9000	rtracklayer	1.52.1
GenomicFeatures	1.44.2	Rtsne	0.16
getopt	1.20.3	scater	1.20.1
ggfortify	0.4.15	scran	1.20.1
ggnewscale	0.4.9	scuttle	1.2.1
ggplot2	3.4.0	SingleCellExperiment	1.14.1
ggpubr	0.4.0	stringr	1.5.0
ggrastr	1.0.2	SummarizedExperiment	1.22.0
ggsci	2.9	TCGAbiolinks	2.22.4
ggtext	0.1.2	tidyr	1.2.0
ggVennDiagram	1.2.2	tximport	1.20.0
glmnet	4.1-8	uwot	0.1.14
grid	4.1.2	viridis	0.6.2
gridExtra	2.3		

Table S2 RBP motif clustering			
C	**Representative motif**	**Motifs in cluster**	**RBPs in cluster**
1	M136_0.6	M069_0.6, M136_0.6, M317_0.6, M318_0.6, M319_0.6	SNRPA, SNRPB2
2	M350_0.6	M350_0.6	ZFP36, ZFP36L2, ZFP36L1
3	M089_0.6	M089_0.6, M169_0.6	HNRNPL, ENSG00000215042, hnRNPLL
4	M026_0.6	M026_0.6, M161_0.6, M260_0.6	hnRNPK, RBM6, CSDA
5	M227_0.6	M227_0.6, M228_0.6	PTBP1, PTBP2, ROD1
6	M261_0.6	M261_0.6	SF3B4
7	M120_0.6	M120_0.6	CPEB3
8	M122_0.6	M122_0.6	MEX3C, MEX3D, MEX3B
9	M037_0.6	M037_0.6, M320_0.6	MBNL3, MBNL2, MBNL1
10	M140_0.6	M140_0.6, M147_0.6	ENOX1, ENOX2, CNOT4
11	M103_0.6	M065_0.6, M102_0.6, M103_0.6	SRSF9, SRSF1
12	M035_0.6	M035_0.6, M070_0.6	LIN28A, LIN28B, SRSF2, ENSG00000180771
13	M023_0.6	M022_0.6, M023_0.6, M271_0.6	HNRNPA1, HNRNPA3, ENSG00000215492, ENSG00000231942, HNRNPA1L2
14	M159_0.6	M017_0.6, M118_0.6, M159_0.6, M231_0.6, M297_0.6, M298_0.6	RBFOX2, RBFOX3, A2BP1, EIF2S1
15	M053_0.6	M053_0.6, M145_0.6	RBM5
16	M050_0.6	M044_0.6, M050_0.6, M054_0.6, M109_0.6	PPRC1, RBM4B, RBM4, RBM8A
17	M256_0.6	M256_0.6, M347_0.6	ACO1, SNRPA, SNRPB2
18	M290_0.6	M290_0.6	EIF4B
19	M024_0.6	M024_0.6, M141_0.6, M307_0.6	HNRNPA2B1, HNRNPA3, ENSG00000215492, ESRP2, ESRP1, NONO
20	M036_0.6	M036_0.6	MATR3
21	M254_0.6	M082_0.6, M250_0.6, M254_0.6	YBX2, CSDA
22	M167_0.6	M013_0.6, M040_0.6, M167_0.6, M210_0.6	DAZAP1, MSI1, MSI2, HNRPDL
23	M168_0.6	M168_0.6	SFPQ, PSPC1
24	M083_0.6	M083_0.6, M273_0.6	ZC3H10, SRSF1
25	M246_0.6	M246_0.6, M247_0.6	RBMY1F, RBMXL2, RBMXL3, RBMXL1, RBMY1J, RBMY1A1, RBMY1E, RBMY1B, RBMY1D
26	M001_0.6	M001_0.6, M085_0.6	A1CF, ZCRB1
27	M176_0.6	M033_0.6, M160_0.6, M176_0.6	KHDRBS3, KHDRBS1, KHDRBS2
28	M234_0.6	M234_0.6, M235_0.6	RBM47
29	M164_0.6	M055_0.6, M073_0.6, M143_0.6, M164_0.6, M353_0.6	RBMS2, RBMS3, ENSG00000213250, STAR-PAP, RBMS1, SRSF7

30	M157_0.6	M004_0.6, M157_0.6, M229_0.6	BRUNOL4, CELF3, BRUNOL5
31	M207_0.6	M043_0.6, M177_0.6, M188_0.6, M207_0.6, M211_0.6	PCBP4, PCBP3, PCBP2, PCBP1
32	M071_0.6	M002_0.6, M071_0.6, M152_0.6	ANKHD1, ANKRD17, ENSG00000249536, SRSF7, FXR1
33	M081_0.6	M056_0.6, M081_0.6, M111_0.6, M121_0.6	SRSF3, CSDA, YB-1
34	M316_0.6	M316_0.6, M346_0.6	FUS, TAF15, SNRPA, SNRPB2
35	M019_0.6	M019_0.6, M088_0.6	SRSF12, Fusip1, BX511012.1
36	M148_0.6	M148_0.6, M269_0.6	PABPN1, PABPN1L, ZFP36, ZFP36L2, ZFP36L1
37	M330_0.6	M330_0.6	ELAVL2, ELAVL3
38	M154_0.6	M154_0.6	SRSF1
39	M051_0.6	M051_0.6	RBM41
40	M031_0.6	M031_0.6, M232_0.6	ELAVL1, ELAVL3
41	M328_0.6	M328_0.6, M329_0.6	ELAVL2, ELAVL3
42	M108_0.6	M108_0.6, M112_0.6, M124_0.6, M127_0.6	ELAVL1, ELAVL3
43	M052_0.6	M052_0.6, M068_0.6	RBM46, RBM47, SNRNP70
44	M325_0.6	M325_0.6, M344_0.6	NOVA2, ENSG00000248163, ENSG00000249644, RBMX, RBMY1F, RBMXL2, RBMXL3, RBMXL1, RBMY1J, RBMY1E, RBMY1B, RBMY1D
45	M020_0.6	M016_0.6, M020_0.6, M209_0.6	FMR1, FXR2, RBM45
46	M151_0.6	M151_0.6, M153_0.6	HNRNPH2, HNRNPH1, HNRNPF, LIN28A, LIN28B
47	M262_0.6	M046_0.6, M142_0.6, M262_0.6	QKI, RBM42
48	M061_0.6	M061_0.6	SAMD4A, SAMD4B
49	M348_0.6	M348_0.6, M352_0.6	SNRPA, SNRPB2, SRSF2, ENSG00000180771
50	M012_0.6	M012_0.6, M077_0.6	CPEB3, CPEB2, U2AF2
51	M170_0.6	M170_0.6, M175_0.6, M240_0.6	RBM38, RBM24
52	M048_0.6	M048_0.6, M349_0.6	CIRBP, RBM3, PABPC1, PABPC1L, ENSG00000250177
53	M104_0.6	M104_0.6, M105_0.6, M126_0.6	SRSF1, SRSF4, SRSF6
54	M106_0.6	M072_0.6, M106_0.6, M272_0.6, M292_0.6, M333_0.6	SRSF9, SRSF1, EIF4B
55	M354_0.6	M296_0.6, M345_0.6, M354_0.6	IGF2BP1, SNRPA, SNRPB2, YTHDC1
56	M047_0.6	M047_0.6	RBM28
57	M027_0.6	M027_0.6, M163_0.6	HNRNPL, ENSG00000215042, IGF2BP3
58	M079_0.6	M079_0.6, M155_0.6	CELF3, ZNF638
59	M032_0.6	M032_0.6	IGF2BP2

60	M245_0.6	M245_0.6, M334_0.6	NCL, SRSF4, SRSF6
61	M049_0.6	M049_0.6, M178_0.6, M238_0.6	RBM38, BRUNOL6, CELF3
62	M243_0.6	M195_0.6, M236_0.6, M242_0.6, M243_0.6	SF3B4, HNRNPR, SYNCRIP
63	M021_0.6	M021_0.6, M074_0.6, M205_0.6	G3BP2, TARDBP, SF3B4
64	M351_0.6	M291_0.6, M323_0.6, M351_0.6	EIF4B, NOVA2, ENSG00000248163, ENSG00000249644, SRSF1
65	M201_0.6	M201_0.6	SF3B4
66	M158_0.6	M025_0.6, M075_0.6, M149_0.6, M150_0.6, M156_0.6, M158_0.6	HNRNPC, TIA1, CPEB3, CPEB4, RALY, HNRNPCL1
67	M062_0.6	M042_0.6, M062_0.6, M144_0.6, M146_0.6, M162_0.6, M275_0.6	PABPC4, PABPC1L, ENSG00000250177, SART3, PABPC3, PABPC1, PABPC5
68	M086_0.6	M086_0.6, M087_0.6	SRSF12, Fusip1, BX511012.1
69	M274_0.6	M274_0.6, M332_0.6	SRSF2, ENSG00000180771
70	M331_0.6	M331_0.6	SRSF7

Table S3 Tumour-associated events								
Cancer	**A3**	**A5**	**AF**	**AL**	**MX**	**RI**	**SE**	**Total**
BLCA	869	547	1895	374	69	626	1842	6222
BRCA	4779	4239	0	5888	1158	2599	12943	31606
CESC	0	1	0	0	1	0	0	2
COAD	2156	1196	1576	120	59	399	5458	10964
ESCA	54	17	145	47	8	54	145	470
HNSC	1951	1430	7047	1971	327	1094	4658	18478
KICH	1186	943	5708	1501	244	704	3319	13605
KIRC	4563	3606	16195	5275	868	3261	9998	43766
KIRP	4031	3332	12695	4137	668	2557	8949	36369
LIHC	2841	2434	8722	2535	414	1557	6839	25342
LUAD	493	289	661	128	15	201	1180	2967
LUSC	1155	783	3737	890	110	512	3162	10349
PAAD	191	133	1110	170	31	116	565	2316
PCPG	273	303	1532	337	83	110	1077	3715
PRAD	5138	4437	18165	5274	1026	2600	12992	49632
READ	1850	1261	4934	1367	278	1133	4535	15358
STAD	4192	4027	17841	4926	1027	2244	10699	44956
THCA	2582	2482	12127	3093	489	1580	7684	30037
UCEC	697	588	2890	985	138	375	2176	7849
Total	39001	32048	116980	39018	7013	21722	98221	354003

Table S4 Stage-associated events								
Cancer	**A3**	**A5**	**AF**	**AL**	**MX**	**RI**	**SE**	**Total**
BLCA	130	72	103	14	3	36	279	637
BRCA	316	184	0	246	35	103	507	1391
ESCA	36	4	32	3	2	4	35	116
HNSC	73	32	120	13	1	33	124	396
KICH	138	99	334	86	21	58	321	1057
KIRC	2267	1817	7304	2056	445	1192	5250	20331
KIRP	756	581	2724	651	75	436	1453	6676
LIHC	147	85	232	48	8	43	232	795
LUAD	5	2	0	0	0	2	7	16
LUSC	1	0	0	2	0	0	0	3
PAAD	213	140	677	199	48	50	565	1892
READ	123	64	392	97	19	29	240	964
STAD	311	182	1384	292	54	90	547	2860
THCA	290	222	1257	405	60	94	623	2951
Total	4806	3484	14559	4112	771	2170	10183	40085

7.4 Supplementary Data Tables

The following list indicates the contents of extensive data tables that were not included in the body of this thesis. These tables are available the online GitHub repository.

Supplementary Data Table 1: Pan-cancer models
Includes detailed information regarding the number of samples in different experimental groups, as well as the number of events and models fitted across cancer types.

Supplementary Data Table 2: Correlation of differential AS and differential expression
Shows the Pearson correlation coefficients and associated nominal P values for the comparisons between differential gene expression and mean $\Delta logitPSI$ of all the events of the same type in each gene. Correlations were calculated for every cancer and event type separately.

7.5 Copyright clearance

Figures 1 to 5 were created either fully or partially using BioRender. Publication Licences from BioRender follow.

49 Spadina Ave. Suite 200
Toronto ON M5V 2J1 Canada
www.biorender.com

Confirmation of Publication and Licensing Rights

December 3rd, 2023
Science Suite Inc.

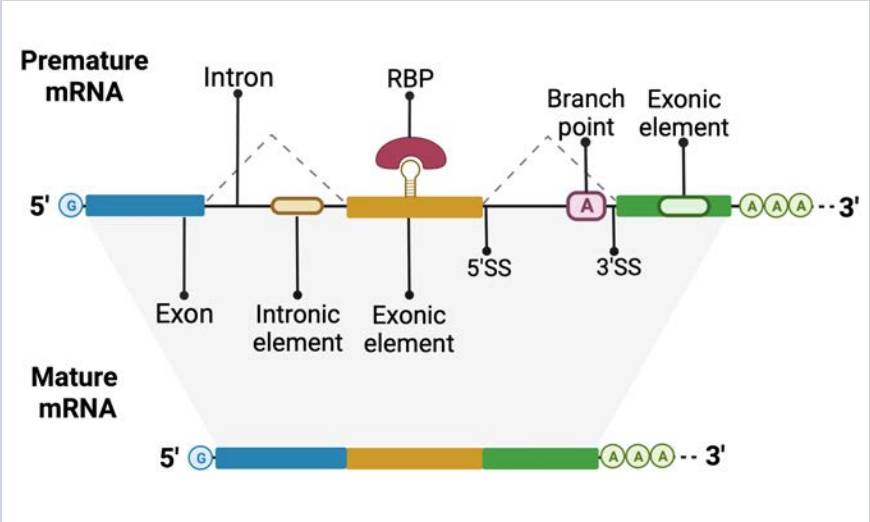
Subscription:	*Institution*
Agreement number:	*KZ2668GZ04*
Journal name:	*Published Thesis*

To whom this may concern,

This document is to confirm that Larisa Soto has been granted a license to use the BioRender content, including icons, templates and other original artwork, appearing in the attached completed graphic pursuant to BioRender's Academic License Terms. This license permits BioRender content to be sublicensed for use in journal publications.

All rights and ownership of BioRender content are reserved by BioRender. All completed graphics must be accompanied by the following citation: "Created with BioRender.com".

BioRender content included in the completed graphic is not licensed for any commercial uses beyond publication in a journal. For any commercial use of this figure, users may, if allowed, recreate it in BioRender under an Industry BioRender Plan.



For any questions regarding this document, or other questions about publishing with BioRender refer to our BioRender Publication Guide, or contact BioRender Support at support@biorender.com.

49 Spadina Ave. Suite 200
Toronto ON M5V 2J1 Canada
www.biorender.com

Confirmation of Publication and Licensing Rights

December 3rd, 2023
Science Suite Inc.

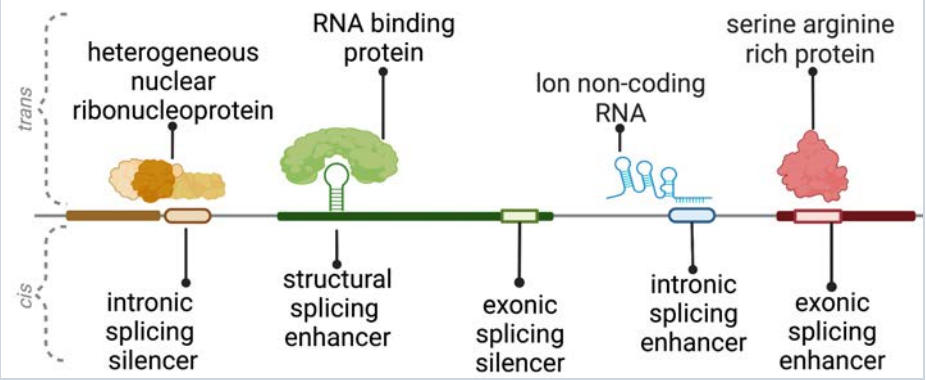
Subscription:	*Institution*
Agreement number:	*TO2668GWNI*
Journal name:	*Published Thesis*

To whom this may concern,

This document is to confirm that Larisa Soto has been granted a license to use the BioRender content, including icons, templates and other original artwork, appearing in the attached completed graphic pursuant to BioRender's Academic License Terms. This license permits BioRender content to be sublicensed for use in journal publications.

All rights and ownership of BioRender content are reserved by BioRender. All completed graphics must be accompanied by the following citation: "Created with BioRender.com".

BioRender content included in the completed graphic is not licensed for any commercial uses beyond publication in a journal. For any commercial use of this figure, users may, if allowed, recreate it in BioRender under an Industry BioRender Plan.



For any questions regarding this document, or other questions about publishing with BioRender refer to our BioRender Publication Guide, or contact BioRender Support at support@biorender.com.

49 Spadina Ave. Suite 200
Toronto ON M5V 2J1 Canada
www.biorender.com

Confirmation of Publication and Licensing Rights

December 3rd, 2023
Science Suite Inc.

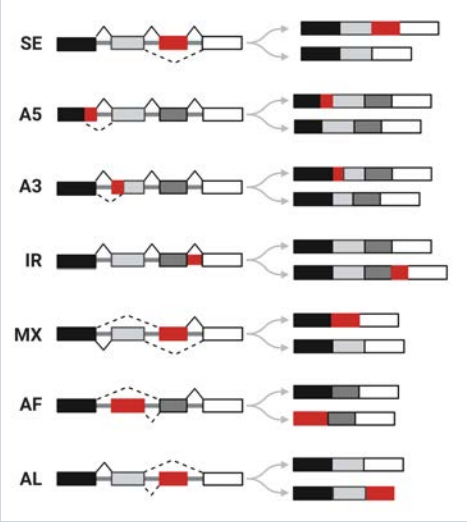
Subscription:	*Institution*
Agreement number:	*XC2668GRG8*
Journal name:	*Published Thesis*

To whom this may concern,

This document is to confirm that Larisa Soto has been granted a license to use the BioRender content, including icons, templates and other original artwork, appearing in the attached completed graphic pursuant to BioRender's Academic License Terms. This license permits BioRender content to be sublicensed for use in journal publications.

All rights and ownership of BioRender content are reserved by BioRender. All completed graphics must be accompanied by the following citation: "Created with BioRender.com".

BioRender content included in the completed graphic is not licensed for any commercial uses beyond publication in a journal. For any commercial use of this figure, users may, if allowed, recreate it in BioRender under an Industry BioRender Plan.



For any questions regarding this document, or other questions about publishing with BioRender refer to our BioRender Publication Guide, or contact BioRender Support at support@biorender.com.

49 Spadina Ave. Suite 200
Toronto ON M5V 2J1 Canada
www.biorender.com

Confirmation of Publication and Licensing Rights

December 3rd, 2023
Science Suite Inc.

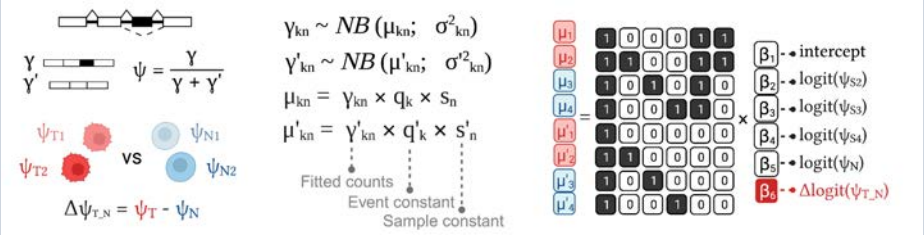
Subscription:	*Institution*
Agreement number:	*VL2668GPAT*
Journal name:	*Published thesis*

To whom this may concern,

This document is to confirm that Larisa Soto has been granted a license to use the BioRender content, including icons, templates and other original artwork, appearing in the attached completed graphic pursuant to BioRender's Academic License Terms. This license permits BioRender content to be sublicensed for use in journal publications.

All rights and ownership of BioRender content are reserved by BioRender. All completed graphics must be accompanied by the following citation: "Created with BioRender.com".

BioRender content included in the completed graphic is not licensed for any commercial uses beyond publication in a journal. For any commercial use of this figure, users may, if allowed, recreate it in BioRender under an Industry BioRender Plan.



For any questions regarding this document, or other questions about publishing with BioRender refer to our BioRender Publication Guide, or contact BioRender Support at support@biorender.com.

49 Spadina Ave. Suite 200
Toronto ON M5V 2J1 Canada
www.biorender.com

Confirmation of Publication and Licensing Rights

December 3rd, 2023
Science Suite Inc.

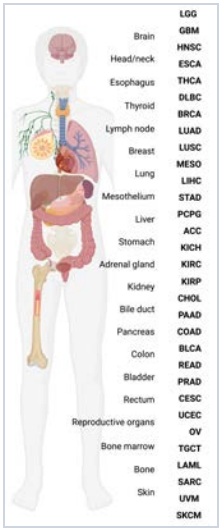
Subscription:	*Institution*
Agreement number:	*UX2668HQPF*
Journal name:	*Published Thesis*

To whom this may concern,

This document is to confirm that Larisa Soto has been granted a license to use the BioRender content, including icons, templates and other original artwork, appearing in the attached completed graphic pursuant to BioRender's Academic License Terms. This license permits BioRender content to be sublicensed for use in journal publications.

All rights and ownership of BioRender content are reserved by BioRender. All completed graphics must be accompanied by the following citation: "Created with BioRender.com".

BioRender content included in the completed graphic is not licensed for any commercial uses beyond publication in a journal. For any commercial use of this figure, users may, if allowed, recreate it in BioRender under an Industry BioRender Plan.



For any questions regarding this document, or other questions about publishing with BioRender refer to our BioRender Publication Guide, or contact BioRender Support at support@biorender.com.