University of Torino

PhD program in Complex Systems for Quantitative Biomedicine

Computational investigation of alternative splicing in pediatric sarcomas

Francesca Priante

Internal advisor: Prof. Michele Caselle

External advisor: Prof. Matteo Cereda

Cycle XXXVI

Abstract

Sarcomas, primary malignant tumors affecting most children and young adults, present high metastasis recurrence and extremely poor clinical prognosis, posing an urgent need for new treatment strategies for these young patients. Given the limited mutational burden of this type of cancer, other types of dysregulation and potential therapeutic solutions need to be addressed. This thesis investigates the dysregulation of splicing factors and alternative splicing, considering their known implication in tumorigenesis in other cancers. Leveraging both computational and statistical methodologies, this research offers a comprehensive analysis of transcriptomic data from pediatric sarcomas, in particular, 8 Ewing sarcoma (EW) and 18 Osteosarcoma (OS) samples (enrolled within the national clinical trial SAR-GEN_ITA "Genomic Profiles Analysis in Children, Adolescents and Young Adult With Sarcomas", clinicalTrial.govid:NCT04621201), comparing them to publicly available osteoblast controls.

The investigation was conducted from three possible intervention points for personalized RNA-based therapies in pediatric sarcomas: (i) identification of alternative splicing events with negative effects on prognosis to be potentially targeted with splice-switching antisense oligonucleotides (SSOs); (ii) detection of RBPs with deleterious effect on splicing, to be targeted with small interfering RNA; and (iii) splicing derived neoepitopes discovery to guide targeted immunotherapy specifically to cancer cells.

For the first point, differential splicing analysis and cross-intersection with public available databases, identified alternative splicing events associated with patient prognosis, with 30 and 24 harmful events in OS and EW, respectively, and seven shared between the two diseases. These events represent potential targets for SSO-based therapies aimed at improving patient outcomes.

To address the second point, the RNAmars (RNA Motifs And cognate Regulators of alternative Splicing) algorithm was developed. RNAmars enables the discovery of the regulatory RBP in alternative splicing dysregulation and its motifs characterization. RNAmars revealed RBFOX2 as the main exon inclusion enhancer and U2AF2 the primarily exon silencers in both subtypes.

To tackle the last point, patient-specific splicing-derived neoepitopes were derived, with most being unique to individual patients. The most common neoepitope, originating from an alternative 3' splice site in the ATF6B gene, is shared among seven patients.

In conclusion, this research integrates computational and statistical methodologies to propose therapeutic solutions for pediatric sarcoma patients. These solutions encompass both personalized and shared strategies, demonstrating a comprehensive approach to advancing treatment for this challenging patient population.

Table of contents

List of abbreviations **6**

Introduction **8**

Data overview **15**

Sarcomas have low tumor mutational burden 15

Analysis on a clinical case of EW with Down syndrome 17

Chapter 1: Unveiling the transcriptomic alterations **22**

Development and cancer genes are deregulated in sarcomas 22

Landscape of alternative splicing in sarcoma 26

Identification of survival related alternative splicing events 31

Chapter 2: RNAmars - identification of multivalent RNA Motifs And cognate Regulators of alternative Splicing **39**

Identification of RBPs-specific binding regions 42

Characterization of RBPs binding and motifs preferences 46

Pipeline overview and performance evaluation 50

Testing RNAmars on HNRNPK silencing in PC3 cell line 53

RBFOX2 and U2AF2 regulate alternative splicing in sarcomas 57

Chapter 3: Splicing derived neoepitopes **79**

Alternative splice sites and exon skipping are major sources of neoepitopes 81

Neoepitopes are patient specific 83

Discussion **87**

Conclusions **90**

References **91**

List of abbreviations

OS: Osteosarcoma

EW: Ewing Sarcoma

RBP: RNA-Binding Protein

TF: Transcription Factor

EM: Epigenetic Modifier

CCD: Canonical Cancer Driver

PSI: Percentage Spliced In

SSO: Splice Switching Oligonucleotides

DEG: Differentially Expressed Genes

ORA: Over-Representation Analysis

PTC: Premature Termination Codon

NMD: Nonsense Mediated Decay

HR: Hazard Ratio

ASE: Alternative Splicing Event

CE: Cassette Exon

DSG: Differentially Spliced Genes

AS: Association Score

MHC: Major Histocompatibility Complex

Introduction

The genomic and transcriptomic landscape of pediatric sarcomas

Sarcomas are primary malignant tumors affecting children and young adults (Tirtei et al., 2020). They are a set of highly heterogeneous mesenchymal diseases with about 100 histological subtypes (Damerell, Pepper, and Prince 2021). They present a huge variety of genomic abnormalities and gene expression complexity. The traditional therapy for this type of tumors is a combination between radiation therapy, surgery and chemotherapy. However, sarcomas have a 50% chance of developing metastasis and 20% of recurrence after the treatment (Damerell, Pepper, and Prince 2021). The high metastasis recurrence and extremely poor clinical prognosis pose an urgent need for new treatment strategies for these young patients. Therefore the genomic investigation of these tumors is crucial for the achievement of efficient and personalized therapies. In particular, osteosarcoma (OS) and Ewing sarcoma (EW) present a high level of both inter- and intra-tumor heterogeneity leading to multiple clinical and pathologic consequences.

OS and EW are the most frequent bone sarcomas within the pediatric population (Tirtei et al. 2020). OS is developing mainly within the long bones of the arms and legs, while EW occurs in the pelvis, thigh, lower leg, upper arm, and rib. Furthermore, EW are characterized by a chromosomal translocation that fuses an RNA-binding protein from the FET family (encoded by FUS, EWSR1 and TAF15) with transcription factor from the ETS family (encoded mainly by FLI1 and ERG), creating an aberrant transcription factor. The most common fusion protein is EWSR1-FLI1 occurring 85% of the time (Grünewald et al. 2018).

These types of tumors are characterized by a limited number of somatic mutations. (Venkataramany et al. 2022; Gröbner et al. 2018) and are negatively affected by alternative splicing. Accumulating evidence highlights the relationship between alternative splicing and with patient survival, suggesting a role of RNA defects in tumorigenesis (Li et al. 2021; Yang et al. 2019).

For example, skipping exon 11 in the insulin receptor INSR gene in OS can lead to an increase in cell growth, migration, and viability (Venkataramany et al. 2022). Moreover, skipping exon 8 of the EWSR1 gene in EW is associated with the production of a functional EWSR1-FLI1 fusion protein (“Functional Genomic Screening Reveals Splicing of the EWS-FLI1 Fusion Transcript as a Vulnerability in Ewing Sarcoma” 2016).

A study of 236 patients from the The Cancer Genome Atlas Program (TCGA) detected 9674 alternative splicing events that were negatively associated with patient survival (Li et al. 2021).

Moreover, deregulated RNA Binding Proteins (RBPs) produce negative clinical outcomes of cancer treatment in sarcoma patients. For instance, the expression of HNRNPM in EW correlates with treatment resistance and poor patient survival (Passacantilli et al. 2017).

Furthermore, the Serine-rich splicing factor 3 (SRSF3) has been proven to be a proto-oncogene of osteosarcoma cell lines (U2OS). Specifically, SRSF3 controls the expression of 60 genes and the inclusion of 182 splicing events which are primarily associated with tumor cell proliferation (Ajiro et al. 2016).

For these reasons the comprehension of splicing regulation in this type of malignancies is fundamental for supporting the clinical research.

The splicing code

Pre-mRNA splicing is the process by which parts of the RNA molecule are excised (introns) and other parts are retained (exons). During constitutive splicing, introns are systematically excised, while exons are retained in a sequential manner.

This process is mediated by a large macromolecular complex called the spliceosome. This machinery recognizes specific sequences within introns and exons to perform splicing. Key consensus sequences involved in splicing include the 5' splice site (5'ss) with a GU dinucleotide at the intron's 5' end, the branch site typically containing an adenosine nucleotide, and the 3' splice site (3'ss) composed of a polypyrimidine tract and an AG dinucleotide at the intron's 3' end.

Alternative splicing refers to the process of selectively including or excluding different combinations of exons and introns in a gene's mRNA transcript.

This combinatorial process allows a single gene to produce multiple mRNA variants, leading to protein diversity within the cell. In particular, alternative splicing drives proteome diversity in 95% of human genes and it is finely controlled by both *cis*-acting regulatory elements and *trans*-acting splicing factors.

Specifically, RPBs bind to clusters of short sequences, referred to as multivalent RNA motifs, to promote or repress a given splicing event in a position-dependent manner (Cereda et al. 2014). Being multivalent, these binding sites are recognized by distinct RBPs. The same motif can facilitate the inclusion or the exclusion of an exon according to its location at the exon/intron junctions, even if bound by the same RBP. The multivalency of RNA motifs enlightened the presence of the 'splicing code' (Baralle and Baralle 2018; Barash et al.

2010). The complexity of this code is exacerbated by cooperative and competitive mechanisms exploited by RBPs for mRNA production (Corley et al, 2020).

For example, serine/arginine-rich proteins are usually exon enhancers while heterogeneous nuclear ribonucleoproteins (hnRNPs) are silencers (Bradley and Anczuków 2023; Van Nostrand, Freese, et al. 2020; Carazo, Romero, and Rubio 2019; Ule and Blencowe 2019; Stanley and Abdel-Wahab 2022) (Figure 1).

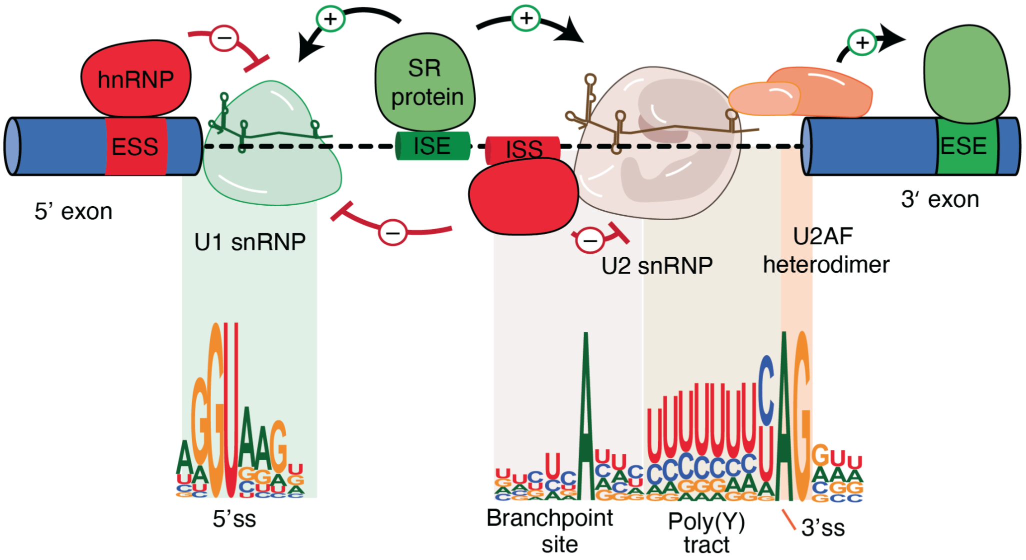

Figure 1: Adapted from (Stanley and Abdel-Wahab 2022): key *cis*-acting features and *trans*-acting splicing factors that govern splicing. Sequences that are required for spliceosome assembly include the GU at the 5′ss, the AG at the 3′ss, and the branch site residue A. In light green the spliceosome component U1 snRNP which initiates splicing by binding to the 5'ss consensus motifs. In brown another spliceosome component U2 snRNP is recruited at the branch site by the U2AF heterodimer, composed by U2AF1 and U2AF2 (orange), which in turn recognise the 3'ss motifs. Exonic splicing and intronic splicing enhancer motifs (ESE and ISE) are represented by red cylinders, while exonic splicing silencer and exonic splicing enhancers (ESS and ISS) are represented by green cylinders. These motifs are recognized by specific *trans*-acting RNA-binding proteins, including SR proteins (for enhancers) and hnRNPs (for silencers).

Therefore, cracking the splicing code requires a comprehensive characterization of all possible multivalent RNA motifs (*i.e. cis*-acting element) and cognate RBPs (*i.e. trans*-acting element).

Alterations of the RBPs-mRNA interactions can lead to cancer onset and progression (Julian P. Venables, Tazi, and Juge 2012; Lee and Abdel-Wahab 2016). Moreover, the altered expression of RBPs occurs in many different cancer types (Del Giudice et al. 2022; Kahles et al. 2018). For example, the dysregulation of transcription factors regulating RBPs such as

MYC and FOXA1 was shown to provoke a disruption of splicing landscape in prostate cancer (Phillips et al. 2020; Del Giudice et al. 2022).

However splicing alterations are not only due to RBPs altered expression, but are also caused by *cis*-acting mutation in the motifs and *trans*-acting alteration causing a loss or gain of functional domains of the RBPs. The loss of function in RBPs can occur due to somatic mutations in the RNA encoding the protein or as a result of alternative exon excision. This feedback-loop mechanism, wherein splicing events lead to abnormalities in the RBPs that govern splicing, is a well-known process used by proteins to regulate themselves (Del Giudice et al. 2022; Guo, Jia, and Jia 2015; Pervouchine et al. 2019).

Proposed strategies for RNA-based therapies

Considering the strong connection between splicing and sarcomas, RNA-based therapy can prove to be a possible solution for the cure of these tumors (Yuanjiao Zhang et al. 2021). Furthermore, the integration of data analysis and computational methodologies has paved the way for advancements in precision medicine (Del Giudice et al. 2021).

There are mainly three possible intervention points that could be targeted through drugs and that can be identified using computational approaches (Figure 2): (i) splice-switching antisense oligonucleotides (SSOs) binding to isoforms, (ii) small molecules targeting splicing factors, and (iii) Immunotherapy targeting splicing derived neoepitopes. These strategies tackle the tumor at three different levels: at the RNA level for the SSOs, the protein level for the small molecules and at the cellular level for the immunotherapy. In particular:

1. SSOs have the ability to form complementary pairs with specific RNA sequences and can induce the degradation of the targeted isoform. Consequently, the aberrant transcript harboring the alternative splicing event is prevented from being translated into protein, thereby neutralizing its potential harmful effects. This can be exploited to target alternative splicing events associated with cancer progression or patients' survival.

   Moreover, SSOs can also be exploited to modulate inclusion and exclusion of an exon by binding the *cis*-acting elements in the pre-mRNA sequence and preventing *trans*-acting elements from binding (Zhu et al. 2022).
2. Small molecules can bind to proteins and they can modulate their levels. Specifically, they can be exploited to inhibit splicing regulators that have deleterious effects on splicing and therefore limit the production of alternative isoforms (Zhu et al. 2022; Kim 2022).

3. The immunotherapy strategy is founded on the concept that tumors and normal cells possess distinct splicing junctions. When these splicing junctions are translated into proteins, they lead to the production of different antigens that are presented on the cell surface. Evidence suggests that splicing alterations have more potential to produce neoepitopes with respect to somatic mutation (Wang et al. 2021; Oka et al. 2021). This can be exploited to guide T-cells to target tumor cells which are specifically presenting tumor-specific neoepitopes.

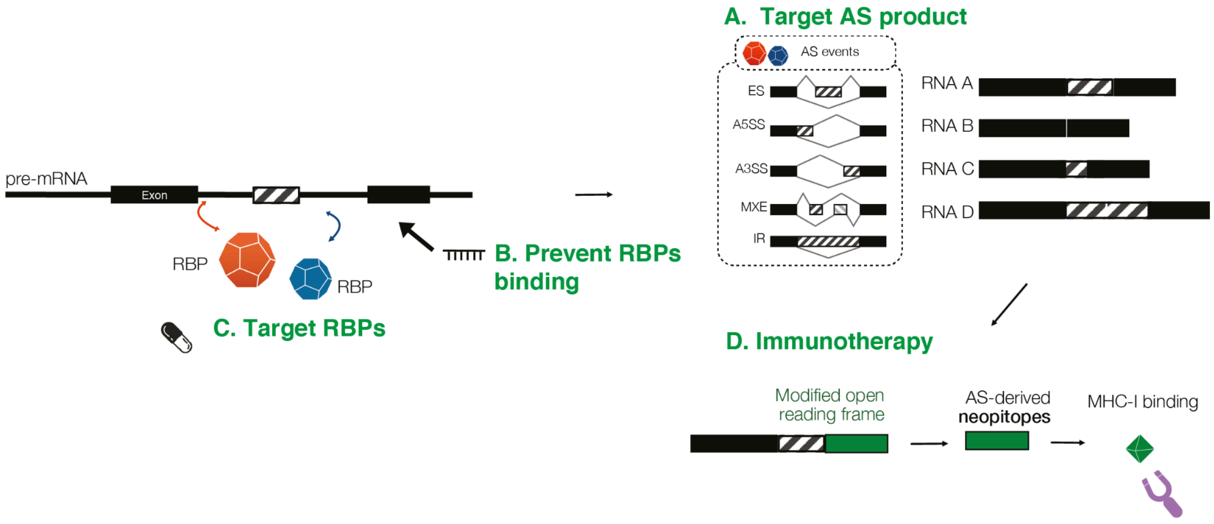

Figure 2: Schematic illustration of the possible RNA therapies intervention points: **A.** Exploit splice-switching antisense oligonucleotides (SSOs) to target poor prognosis-related isoforms. **B.** Employ SSOs to modulate inclusion of exons by preventing RBPs from binding. **C.** Utilize small interfering RNA to inhibit RBPs having deleterious effects on splicing. **D.** Leverage immunotherapy to guide T-cells to target tumor cells presenting tumor-specific and splicing derived neoepitopes

The thesis is divided into three major chapters that are aimed at the identification of candidates for the aforementioned intervention points via statistical and computational strategies. In the first chapter of the thesis, I will elucidate the methodology employed to identify alternative events that exhibit a negative correlation with patient survival, making them potential candidates for targeted interventions using SSOs.

To address the challenge of identifying the proteins accountable for alternative splicing, I developed a computational approach for the identification of multivalent RNA Motifs And cognate Regulators of alternative Splicing (RNAmars) algorithm, which will be explained in the second chapter. This tool is aimed at guiding the identification of the RBPs responsible for the splicing alteration.

In the third chapter, ISOTOPE pipeline is utilized to identify the splicing derived neoepitopes, which can be potentially targeted by immunotherapy (J. L. Trincado et al. 2021).

Data overview

Sarcomas have low tumor mutational burden

Within SAR_GEN-ITA (ClinicalTrial.gov id:NCT04621201), 46 young patients (up to 25 years old) affected by Osteosarcoma (OS) and Ewing Sarcoma (EW) were enrolled.

All patients underwent whole exome sequencing (WXS), and among them, 34 individuals also underwent total RNA sequencing (RNA-seq). WXS data revealed low tumor mutational burden (number of somatic mutation within 1M base pairs) and high copy number burden among patients (estimated by Sigminer, see methods) (Figure 3). In particular, 32 patients showed amplification or high-level amplification (more than four amplified copies) of different genes, indicating high genomic instability among these types of cancers. This observation corroborated the previous findings that sarcomas exhibit infrequent mutations but high genome instability (Gröbner et al. 2018).

Both OS and EW exhibited notable genetic heterogeneity, not only within individual tumors but also across different tumors of the same type. Furthermore, within the genetic landscape of these sarcomas, the pool of actionable targets, *i.e.* those genes that can be specifically targeted with therapeutic interventions (see Methods), is found to be relatively limited. The restricted mutational landscape poses the need to explore other types of dysregulations and potential therapeutic intervention points through alternative sources, such as RNA.

By analyzing RNA-seq data, the majority of EW (8 out of 11, 72%) were characterized by EWSR1–FLI1 fusion which occurs in 85% of the cases (Damerell, Pepper, and Prince 2021). A patient also manifested the EWSR1–ERG fusion, which arises at a 5-10% rate (Damerell, Pepper, and Prince 2021).

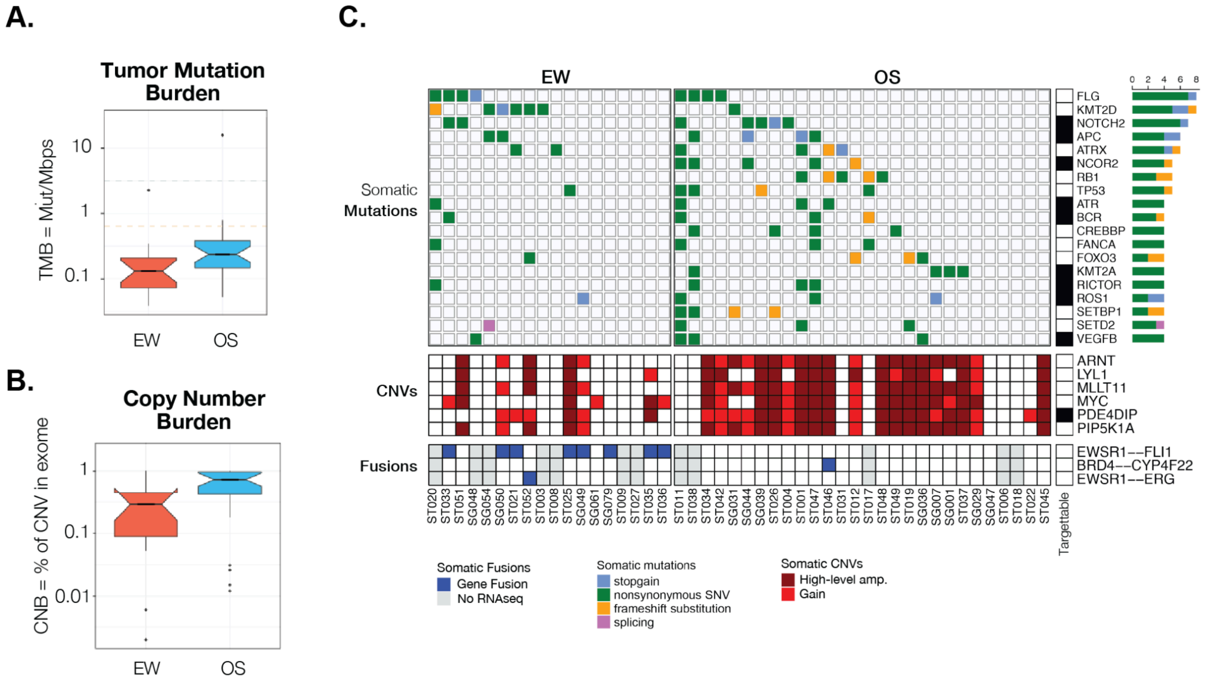

Figure 3: Tumor mutational burden and Copy number variation derived from WXS on 46 Ewing and Osteosarcoma patients. **A.** Tumor mutational burden expressed in the number of somatic mutations within 1M base pairs. **B.** Copy Number Burden expressed in percentage of copy number variation in the whole exome. Color code defines the disease subtype. **C.** The top heatmap collects patients in columns and genes in rows and the colored square defines the presence of the driver somatic mutation within the gene and the patient. Cell color is the mutation type. Lateral barplot measures the number of somatic mutations across patients within the same gene. Single column heatmap at the right is black for actionable genes. Middle heatmap represents the amplification level of each gene across patients. Bottom heatmap indicates gene fusions. When the sample did not undergo RNA-seq it was indicated by a gray cell.

Only 26 samples among all the RNA-sequenced had at least 5 million reads and were retained for all the downstream analysis (Figure 4). The selected patients were 14 years old on average at the moment of enrollment, with the youngest patient being two years old (st033).

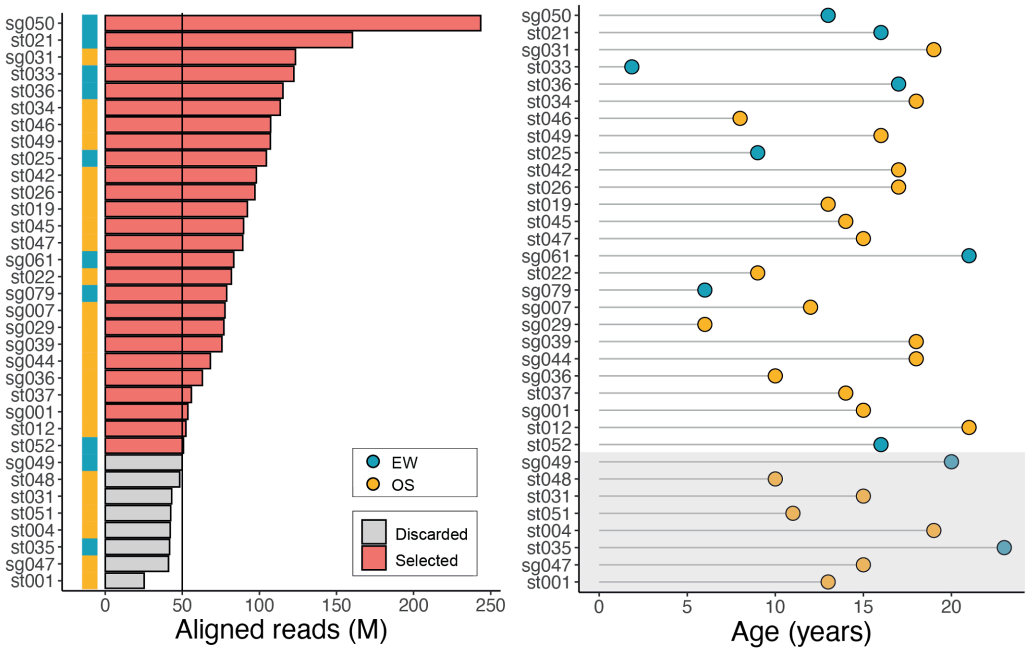

Figure 4: Landscape of the 34 patients that underwent RNA-seq. On the left hand panel the number of reads that were assigned to genes (units in million reads). Red bars are the samples with at least 50 million reads. On the right hand panel the age of the patients. Left side annotation and point color code indicate the disease subtype.

Analysis on a clinical case of EW with Down syndrome

An individual with a unique characteristic of having Down syndrome was examined separately among the patients in the study. This specific patient, referred to as "st033", is the youngest among the entire group (two years old). RNA-seq confirmed the presence of the EWSR1-FLI1 fusion (Figure 3). To understand the characteristics of this particular case study, the sample was compared with samples of similar age taken from the St. Jude database (McLeod et al. 2021). Over-representation analysis of pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) revealed an enrichment of immune- and infectious-disease related pathways in up-regulated genes. In particular, 20% of the differentially expressed genes belonged to the cytokine-cytokine receptor interaction pathway (Figure 5). Performing the same analysis using the Hallmark gene sets (Liberzon et al. 2015) showed an enrichment of immune related pathways in up-regulated genes (Figure 5). This work is available as a preprint in the MedRxiv portal (Peirone et al. 2023)

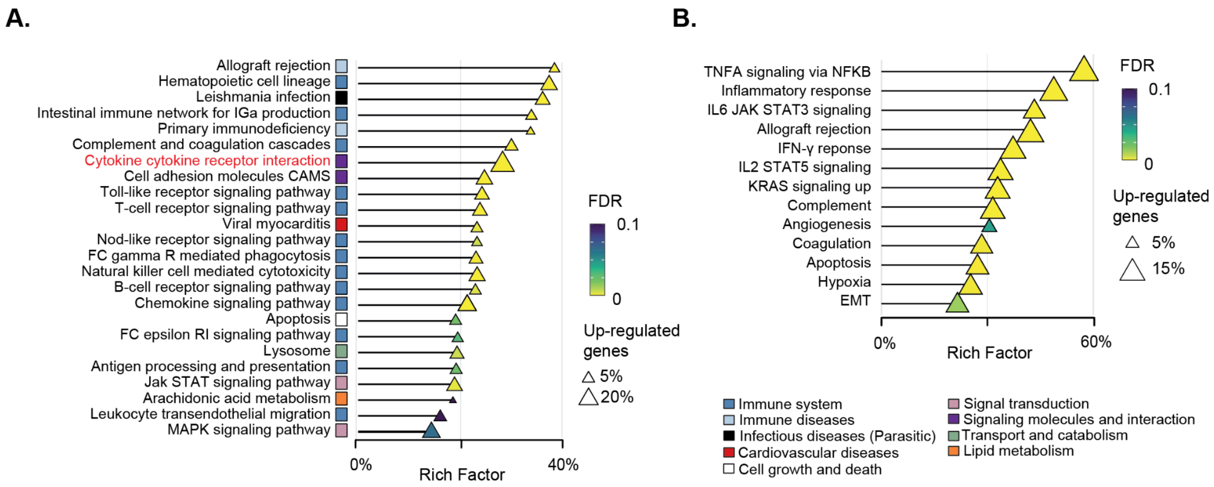

Figure 5: Over-representation analysis results on: **A.** KEGG genes sets **B.** Hallmark gene sets. Shape size indicates the fraction of DEG in each pathway. The Rich Factor represents the fraction of genes in a pathway that are differentially expressed divided by the genes annotated to that pathway. Color key represents the statistical significance (FDR) of the enrichment. Only enriched pathways (FDR<0.1), if any, are shown and sorted by statistical significance. No enrichment found for down-regulated genes.

Methods

Sequence alignment and variant calling

Somatic mutations were identified through the integration of the already published pipeline (Cereda et al. 2016) with the GATK Best Practice guidelines as implemented in the HaTSPiL framework (Morandi et al. 2019). In particular, reads from each sample were aligned to the human genome reference (GRCh37/hg19) using Novoalign (http://www.novocraft.com/) with default parameters.

At most three mismatches per read were allowed and PCR duplicates were eliminated using the Picard Markduplicates tool (Broad Institute 2022).

To enhance the accuracy of variant identification, local realignment around indels was performed using GATK RealignerTargetCreator and IndelRealigner tools.

Single nucleotide variants (SBSs) and minor insertion/deletions (IDs) were detected using independent analyses on tumor and normal samples via MuTect v.1.1.17 (Cibulskis et al. 2013), Strelka v.1.0.15 (Saunders et al. 2012), and Varscan2 v.2.3.6 (Koboldt et al. 2012).

Only variants designated as 'KEEP' in MuTect and 'PASS' in Strelka were taken into consideration. SBSs and IDs were retained if they met two criteria: (i) possessing an allele frequency of at least 5%, and (ii) occurring at a genomic position with a minimum coverage of 10 reads.

Copy number detection and purity and ploidy estimation

Somatic CNV regions were detected employing Sequenza v.3.0.0 (Favero et al. 2015) with parameters window=5mb and min.reads.baf=4. Only positions covered by a minimum of 10 reads were retained in the analysis. To identify amplified and deleted genes, the genomic coordinates of the aberrant regions were intersected with those of 20,297 human protein coding genes of the GENCODE GRCh37 version 28 (Frankish et al. 2019). A gene was deemed as altered if a minimum of 80% of its length fell within an aberrant region. Copy Number burden was estimated using the “Sigminer” package starting from Sequenza results.

Identification of cancer driver mutations

In the tumor sample, SBSs and IDs from the three different tools were identified as somatic if absent in the normal counterpart. Nonsilent mutations (*i.e.* nonsynonymous, stopgain, stoploss, frameshift, nonframeshift and splicing modifications) were identified using

ANNOVAR (K. Wang, Li, and Hakonarson 2010) with RefSeq v.64 (http://www.ncbi.nlm.nih.gov/RefSeq/) as a reference protein dataset.

SBSs and ID falling within 2 bp from the splice sites of a gene in one of the three datasets were considered as splicing mutations.

The Network of Cancer Genes v.5 (An et al. 2016) (http://ncg.kcl.ac.uk/) was used to collect a list of cancer genes. This list was exploited to select 183 and 518 pediatric and adult cancer driver genes, respectively. Of these, 23 and 63 were pediatric and adult sarcoma driver genes, respectively. Moreover, a compilation of 164 genes exhibiting actionable alterations was assembled using the 'PrecisionTrialDrawer' R package (Melloni et al. 2018) and subsequently regarded as actionable genes. Genes harboring nonsilent mutations were annotated using these two gene lists.

All non-silent mutations, excluding frameshift substitutions, were preserved under two conditions: (i) they were identified by a minimum of two variant callers, or (ii) they were located within genes annotated as cancer drivers and/or actionable.

Gene fusion and expression analyses of RNA-seq data

Raw sequencing reads underwent trimming to eliminate nucleotide overlaps between read pairs at both ends using the bbduck tool from bbmap (Bushnell 2014) v.38.18 with parameters forcetrimright=50 and minlength=30. Trimmed reads were aligned to the human genome reference GENCODE GRCh38 version 33 (Frankish et al. 2019) using STAR v.2.7.3a (Dobin et al. 2013) in basic two-pass mode removing duplicates and preventing multimappings (*i.e.* --bamRemoveDuplicatesType UniqueIdentical and --outFilterMultimapNmax 1). Moreover, the following parameters were used: --alignInsertionFlush Right --outSAMstrandField intronMotif --outSAMattributes NH HI NM MD AS XS --peOverlapNbasesMin 20 --peOverlapMMp 0.25 --chimSegmentMin 12 --chimJunctionOverhangMin 8 --chimOutJunctionFormat 1 --chimMultimapScoreRange 3 --chimScoreJunctionNonGTAG -4 --chimMultimapNmax 20 and --chimNonchimScoreDropMin 10. Gene fusions were identified using STAR-Fusion v. 1.9.0 with options --min_FFPM 0 --FusionInspector validate --examine_coding_effect. Only fusions (FFPM≥0.1, LargeAnchorSupport=”YES”, LeftBreakEntropy≥1 and RightBreakEntropy≥1) were retained for further analysis. Read counts at gene level were estimated using featureCounts from Subread v. 2.0.0 (Liao, Smyth, and Shi 2014) with parameters -O --primary -Q 1 -J -s 2 -p -B. The number of transcripts per million reads (TPM) was calculated using the expression values of 19,923 protein coding genes.

Chapter 1: Unveiling the transcriptomic alterations

In this chapter, I will thoroughly investigate the transcriptomic alterations landscape, encompassing gene expression patterns and alternative splicing events. Considering the lack of matched normal tissue, I have used publicly available osteoblast data that have previously been used as controls within a study of osteosarcomas samples (Moriarity et al. 2015) (Accession code GSE57925).

Through gene set enrichment analysis, I will derive key characteristics specific to OS and EW. Additionally, I will identify a subset of alternative splicing events that significantly impact patient survival, designating them as potential targets for therapeutic intervention using SSOs.

Results

Development and cancer genes are deregulated in sarcomas

Sarcomas and Osteoblast gene expression data were visually inspected through Principal Component Analysis (PCA) which revealed a good separation between disease types and controls (Figure 6). Differentially Expressed Genes (DEGs) were then defined with both a parametric approach (DESeq2) and a non parametric approach (Wilcoxon test, see Methods). These approaches were followed by bootstrapping simulations to mitigate the impact of sample size disparities and estimate empirical significance levels similarly to a previously described approach (Del Giudice et al. 2022). In both subtypes the number of upregulated genes in sarcomas was higher than the downregulated genes. Interestingly, there was a substantial overlap of commonly dysregulated genes between the two subtypes (1376 genes with Jaccard index = 38,3%). The gene dysregulations were consistent between the two sarcomas subtypes ($R=0.95$, $p<2.2e-16$) and there were no genes with opposite log2 Fold Change. This finding suggests that even though OS and EW are distinct diseases with unique clinical characteristics, they exhibit a coherent pattern of gene expression alterations.

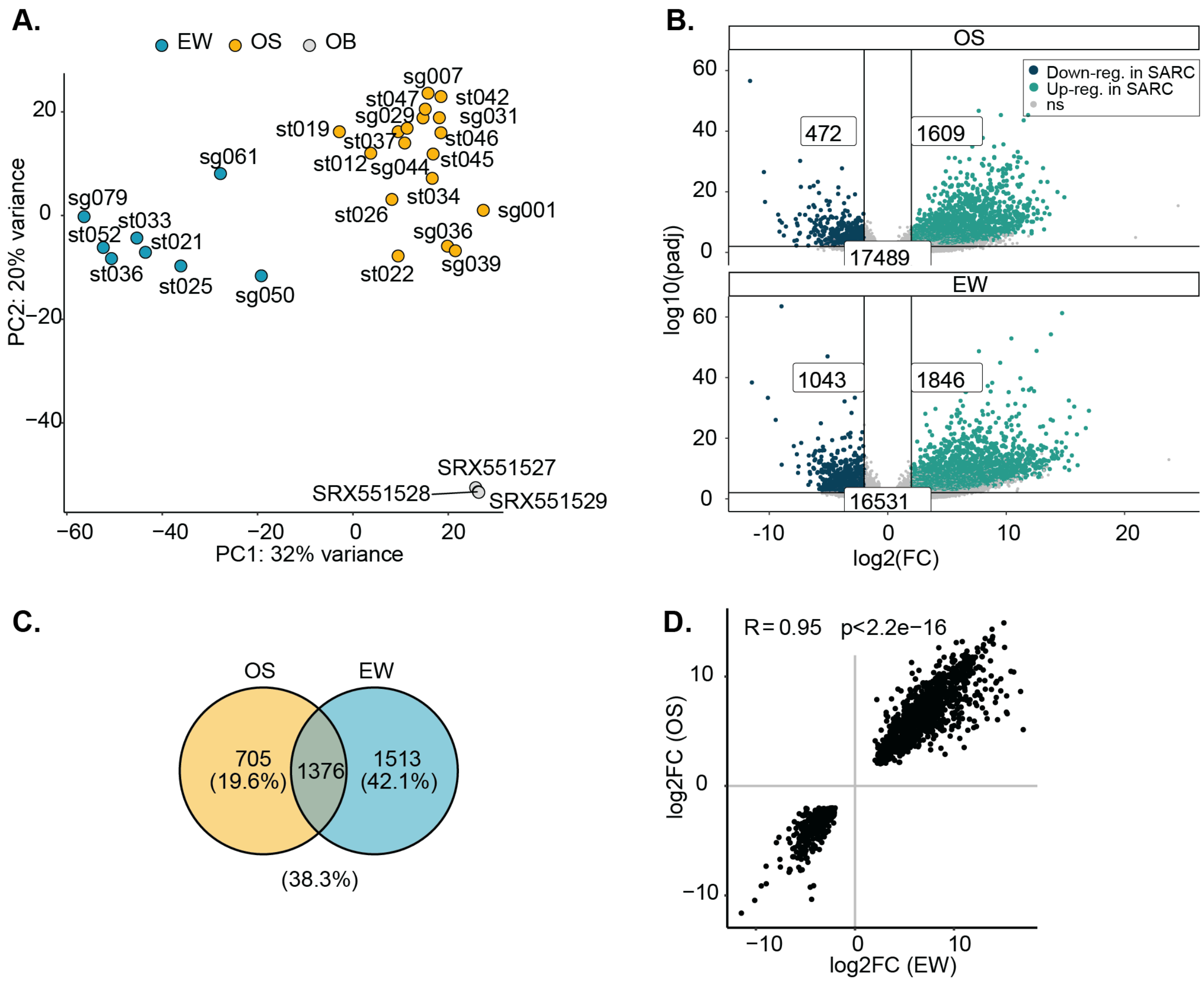

Figure 6: A. Principal Component Analysis of gene expression of 18 OS patients (gold), eight EW patients (light blue) and 3 OB cell lines (gray). **B.** Volcano plots of DESeq2 p-value adjusted and log2 Fold Change. Horizontal line represents p-value adjusted=0.01 and vertical lines represent absolute log2 Fold Change = 2. Gray dots represent genes that are not defined as differentially expressed (see Methods). **C.** Number of differentially expressed genes for OS and EW and their overlap. Percentages represent the Jaccard index. **D.** Scatter plot and Pearson correlation between OS and EW log2 Fold Change

To assess the biological processes affected by DEGs I performed gene ontology analysis. Cell-adhesion related terms (leukocyte_cell_cell_adhesion and cell_substrate_adhesion) were among the top five enriched terms in up-regulated genes in sarcomas compared to osteoblast controls, suggesting a possible disruption of cells communication (Figure 7). Indeed, cell-to-cell adhesion has been extensively associated with cancer progression (Janiszewska, Primi, and Izard 2020; Farahani et al. 2014). Moreover, I also found immunity related terms within up-regulated genes, such as leukocyte migration and cell activation involved in immune response. Conversely, genes that experienced a reduction in expression were associated with terms related to bone development, such as "response to BMP" (Bone

Morphogenetic Protein), "ossification," "collagen fibril organization," and collagen metabolic process. This observation reveals a deficiency in genes related to bone and tissues development within Sarcomas.

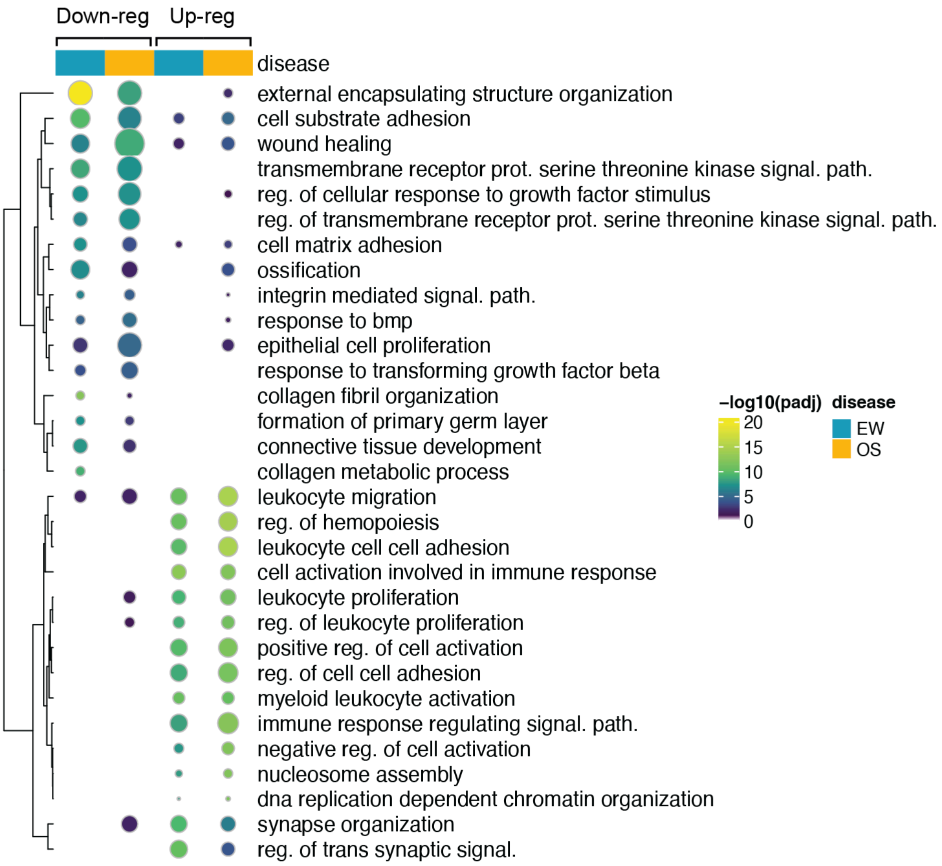

Figure 7: Biological processes Gene Ontology. Columns are stratified by disease subtype (OS, EW) and by direction of regulation (Up or Down). Rows represent biological pathways. Size and ball color are proportional to -log10(p-value adjusted).

To take the pathways analysis further, I wondered whether other specific classes of genes were enriched in DEGs.

In particular, I was interested in transcription related pathways such as Epigenetic Modifiers (EMs), Transcription Factors (TFs) and RBPs. Moreover, due to the well-documented disruption of these pathways in cancers, I took into account both the list of cancer genes and development-related genes (Nwabo Kamdje et al. 2017; Dressler et al. 2022). Finally, in light of the cell-adhesion pathway disruption identified in the previous analysis, I expanded the investigation to incorporate the ligands and receptors derived from the FANTOM database (Lizio et al. 2019). This additional step aimed to assess the extent of deregulation within the cell signaling network. I tested the enrichment of DEGs within these six groups of genes using the Fisher's enrichment test.

No enrichment was found in EM, TFs and RBPs (Figure 8). On the contrary, there was an enrichment of DEGs in development and cancer genes (pv<1e-34 and pv<3.3e-9) for both the cancer subtypes (Figure 8). Children's growth is characterized by an intense cell proliferation and differentiation starting from the early fetal development (Moore 2009) therefore, the disruption of developmental process has the potential to trigger oncogenesis. In previous studies it has been suggested a link between cancer and developmental genes (An et al. 2015). The significant enrichment of dysregulated developmental genes in this analysis endorse this hypothesis.

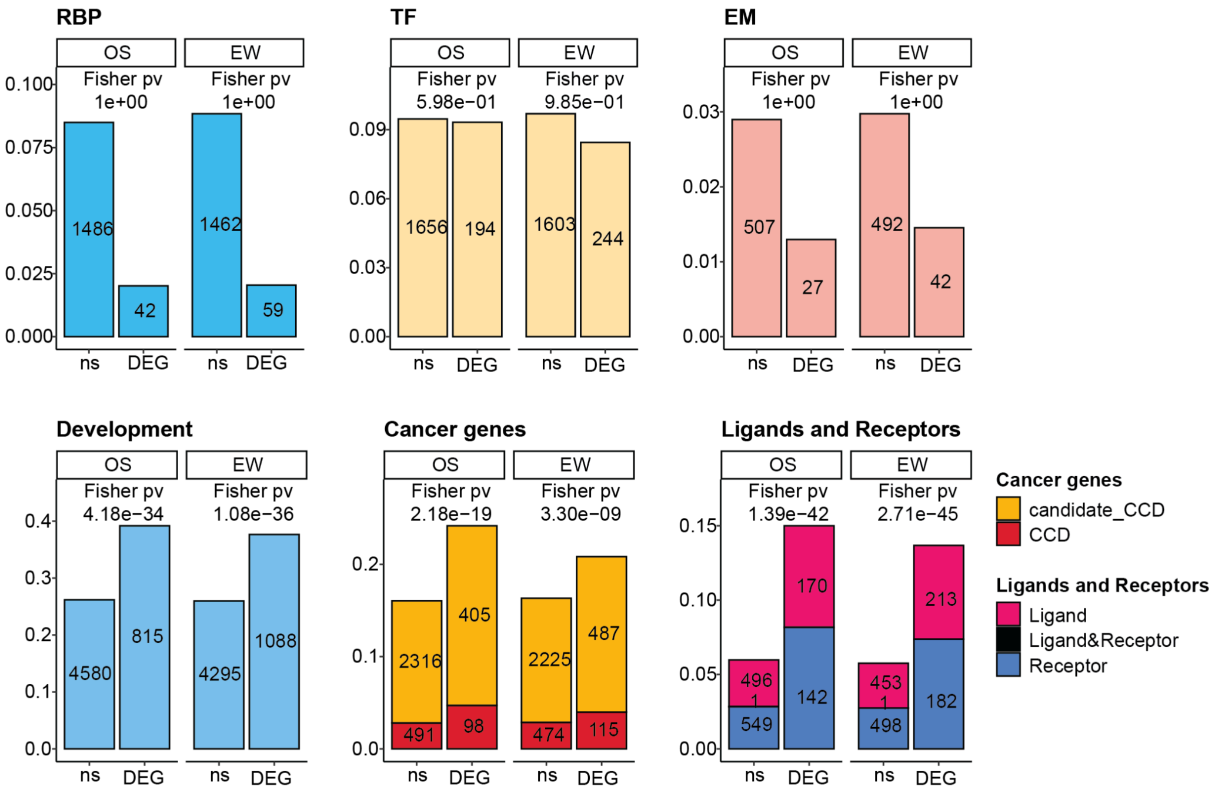

Figure 8: Fisher's enrichment test within different class of genes. Test if performed between DEGs and non significant genes (ns). Y-axis represents the proportion of genes within the class and the absolute number is noted inside the bar.

The results of this preliminary analysis on gene expression expose that the transcriptional dysregulation is not affecting RNA nor DNA binding proteins. Instead, a significant portion of the differentially expressed genes appears to be intricately linked to developmental processes and intercellular communication pathways. However, to understand the alterations more deeply, a finer and more comprehensive analysis at the transcripts level is necessary.

Landscape of alternative splicing in sarcoma

To retrieve alternative splicing events I exploited rMATS (Shen et al. 2014) software, which enables the identification of exons skipping (ES), mutually exclusive exons (MXE), intron retention (IR), alternative 3’ splice site (A3SS) and alternative 5’ splice site (A5SS). For each event type rMATS produces three files: one for annotated events, one for novel junctions (novel JN) and one for novel splice sites (novel SS). Therefore, each event can be labeled as novel or annotated. However taken as it is, rMATS result does not clarify whether the novelty relies on the inclusion junction or exclusion junction (or both).

To overcome this issue I developed a computational strategy to assign the novelty type to each event (Figure 9). This method firstly creates a list of exons and introns coordinates and a set of subsequential intron chains from each transcripts of the reference comprehensive annotation (GRCh38 v33).

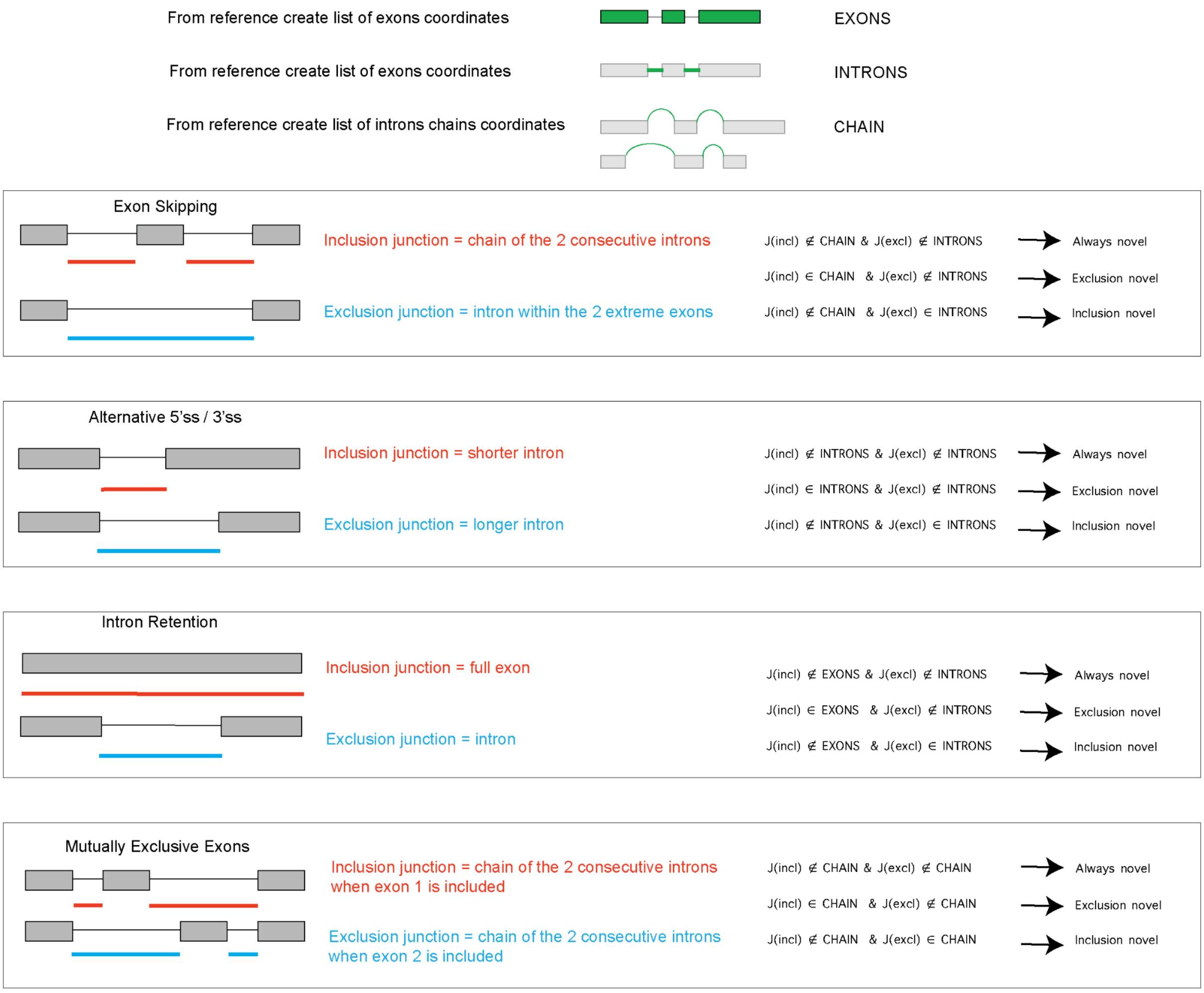
From reference create list of exons coordinates
EXONS
From reference create list of exons coordinates
INTRONS
From reference create list of introns chains coordinates
CHAIN
Exon Skipping
Inclusion junction = chain of the 2 consecutive introns
Exclusion junction = intron within the 2 extreme exons
J(incl) ∉ CHAIN & J(excl) ∉ INTRONS
J(incl) ∈ CHAIN & J(excl) ∉ INTRONS
J(incl) ∉ CHAIN & J(excl) ∈ INTRONS
Always novel
Exclusion novel
Inclusion novel
Alternative 5’ss / 3’ss
Inclusion junction = shorter intron
Exclusion junction = longer intron
J(incl) ∉ INTRONS & J(excl) ∉ INTRONS
J(incl) ∈ INTRONS & J(excl) ∉ INTRONS
J(incl) ∉ INTRONS & J(excl) ∈ INTRONS
Always novel
Exclusion novel
Inclusion novel
Intron Retention
Inclusion junction = full exon
Exclusion junction = intron
J(incl) ∉ EXONS & J(excl) ∉ INTRONS
J(incl) ∈ EXONS & J(excl) ∉ INTRONS
J(incl) ∉ EXONS & J(excl) ∈ INTRONS
Always novel
Exclusion novel
Inclusion novel
Mutually Exclusive Exons
Inclusion junction = chain of the 2 consecutive introns when exon 1 is included
Exclusion junction = chain of the 2 consecutive introns when exon 2 is included
J(incl) ∉ CHAIN & J(excl) ∉ CHAIN
J(incl) ∈ CHAIN & J(excl) ∉ CHAIN
J(incl) ∉ CHAIN & J(excl) ∈ CHAIN
Always novel
Exclusion novel
Inclusion novel

Figure 9: Strategy for right classification of novel junctions.

Then it compares each inclusion and exclusion junction from all rMATS events to check whether they are found in the reference lists. If none of them are matching the reference, the event will be labeled as “always novel”. If only inclusion or exclusion JN are found then the event is labeled as “exclusion novel” or “inclusion novel” respectively.

The aforementioned strategy allowed me to notice some false novel splice sites and false novel junctions identified by rMATS. Indeed, I found 849 and 681 misclassified junctions out of 348,325 and 214,028 total junctions in OS and EW respectively (Figure 10A). From hereafter I used the corrected junction type assignment.

Alternative Splicing Events (ASEs) were detected through a process that involved filtering events based on rMATS statistics, in conjunction with a non-parametric method, as detailed in the Methods section. A total of 2158 and 1367 ASEs were found, prevalently composed by MXE and ES (Figure 10B). Only 219 exons were common between the two diseases (Jaccard index of 12%), but they showed high DPSI correlation (R=0.96, p<2.2e-16) and no opposite behavior (*i.e.* opposite sign of DPSI) (Figure 11A-B).

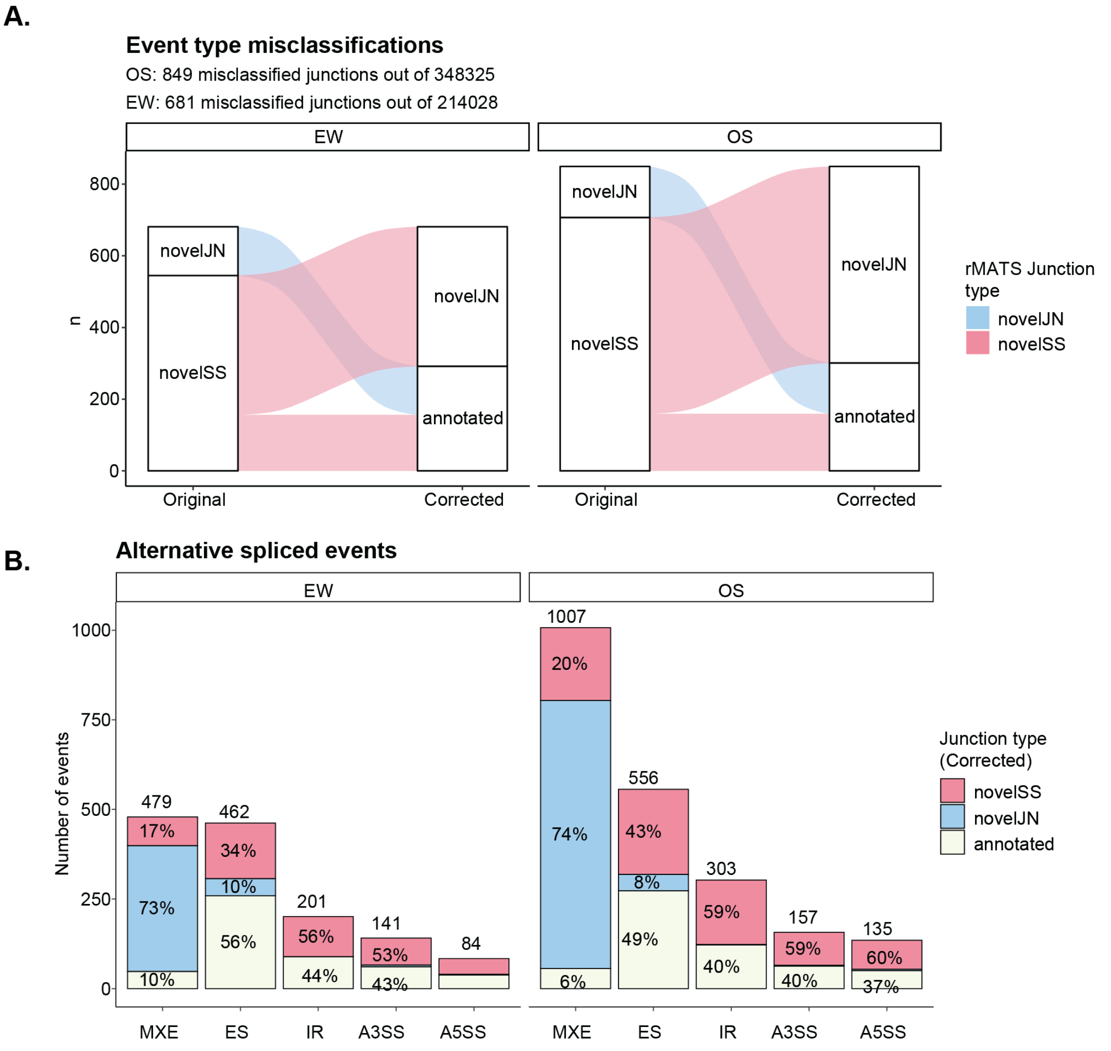

Figure 10: rMATS events. **A.** Junctions rMATS misclassification, on the left side of the Sankey diagram there are the original rMATS labels (also defined by color code), on the right side the

corrected labels retrieve from annotation comparison. **B.** Number of alternative splicing events stratified by junction type and event.

I wondered whether ASEs would enrich the same classes as DEGs. To address this question I performed a one-tailed Fisher's enrichment test (with alternative='greater') on alternative versus constitutive exons in the same classes used in the DEG analysis in the previous section. The results showed an opposite pattern with respect to DEG enrichments. For example, both TFs and EMs displayed significant enrichment in both diseases, indicating a more refined transcriptomic regulation by these classes through splicing rather than at the gene expression level (Figure 11C). Conversely, RBPs were found enriched by ASEs only in OS. Furthermore, developmental genes, cancer drivers and ligands and receptors genes which exhibited pronounced enrichment in DEGs, were not enriched in alternative spliced genes (Figure 11C). As a matter of fact, the intersection between differentially spliced genes (DSGs) and DGEs was minimal in both subtypes, yielding a Jaccard index of 0.011 and 0.015 for OS and EW respectively (Figure 11D).

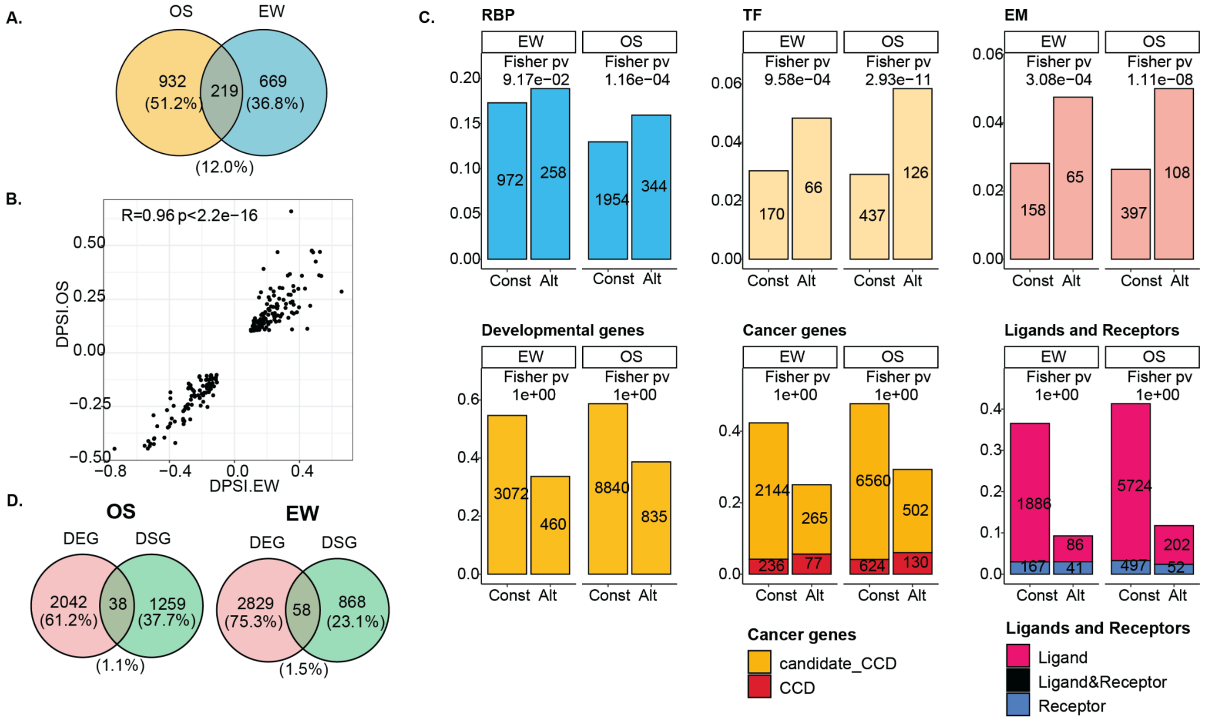

Figure 11: landscape of ASEs. **A.** Number of total ASEs in the two diseases and the intersection and Jaccard index in percentage between the two. **B.** Scatter plot and Pearson's correlation coefficient of inclusion levels between OS and EW. **C.** Proportion of ASEs and Fisher's enrichments within classes of genes. **D.** Number of differentially expressed genes (DEGs) and differentially spliced genes (DSGs) and their overlap in the two subtypes.

To identify the biological processes underlying the genes affected by splicing, I assessed the over-representation analysis (ORA) of genes with at least one ASE in the list of KEGG canonical pathways (Kanehisa and Goto 2000) and Biological Processes gene ontology term up to 8th level. “Spliceosome" appeared among the top 10 most significant terms (p-value adjusted ≤ 0.1) in the KEGG pathways, and “mRNA processing” was the most enriched term in biological processes in both disease subtypes (Figure 12). Also “RNA splicing” pathway was heavily enriched in both OS and EW, suggesting a widespread alternative splicing dysregulation in splicing regulation genes.

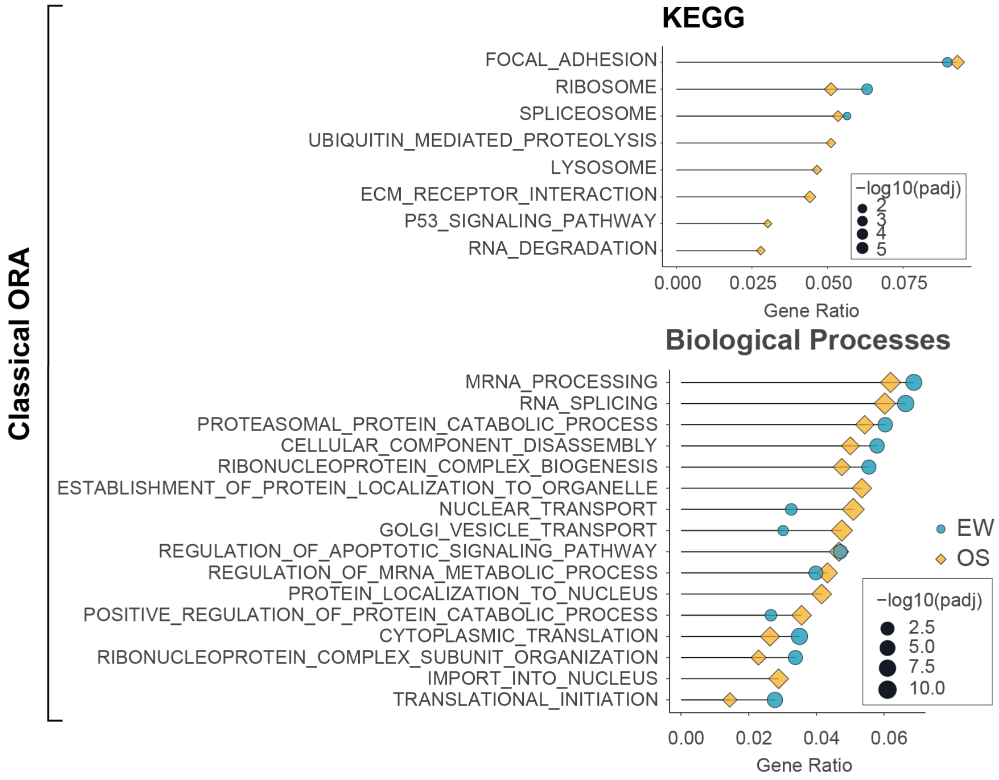

Figure 12: Over representation analysis of genes containing at least one alternative splicing event. Top panel collects pathways from KEGG and bottom panel collects pathways from Biological Process gene ontology terms. Shape and color define the disease. Point size defines -log10(p-value adjusted)

However, it is important to note that the classical ORA does not take into account the number of alternative exons within a gene, potentially introducing bias toward genes with numerous exons. Given that Spliceosome genes have significantly more exons (Figure 13) (considering the longest transcript among genes) the classical ORA could lead to false results.

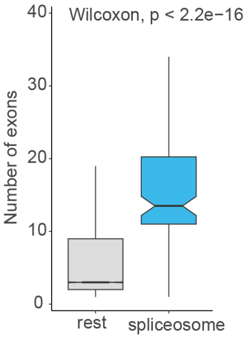

Figure 13: Number of exons within the longest transcript in genes. Wilcoxon test is performed between spliceosome genes and the rest of protein coding genes

To overcome this problem I developed a binomial ORA inspired by previously published analysis (Wang et al. 2021), (see Methods). Exploiting the binomial test enables to keep into consideration multiple alternative exons within the same gene, thereby enhancing the accuracy of the analysis. With the binomial ORA, “RNA Splicing” term is still highly significant (FDR = 1.98e-08 for OS and FDR=9.98e-03 for EW). However, when ranking the pathways by FDR this term is no more among the top 10 ranked terms being the 23-th place for OS and 218-th place for EW (Figure 14). Same held true for the “mRNA processing” pathway which was still significant but not among the top 10. In contrast, within the top enriched terms I found cell cycle related terms (Figure 14). Also the “Spliceosome” term in the KEGG pathway is still significant (FDR = 2.06e-4 for OS and 0.078 for EW) but no more at the top 10 terms (16th and 34th for OS and EW). Interestingly, the top term in the KEGG pathway is “MAPK signaling pathway” whose activation has been already related to cancer proliferation (Yuan et al. 2020).

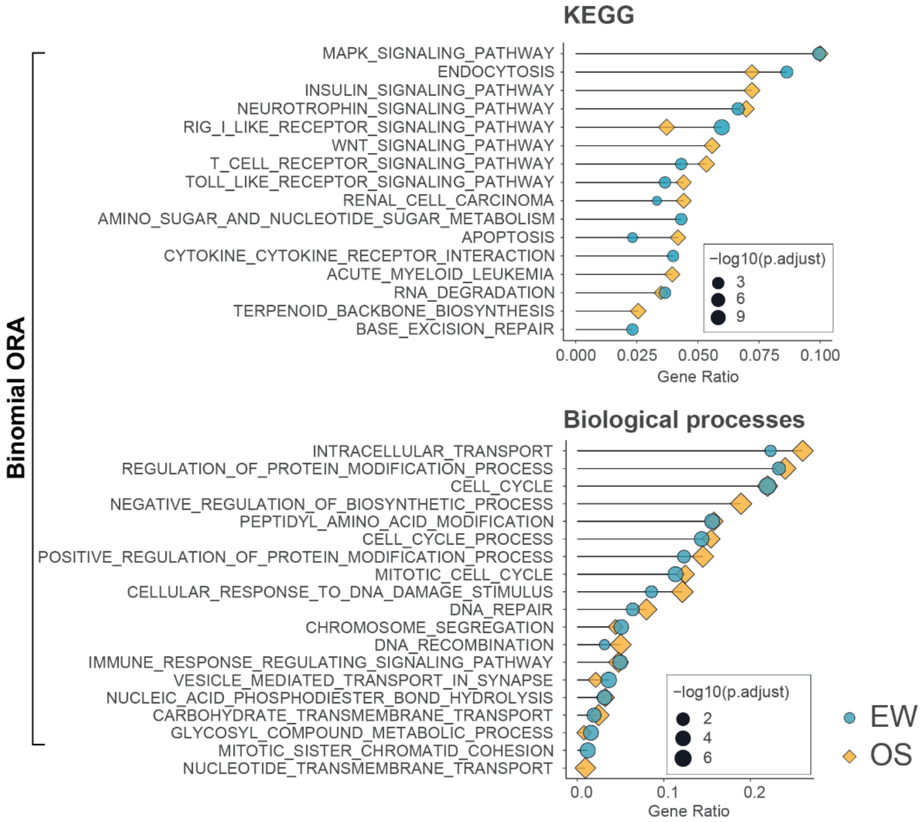

Figure 14: Binomial tests on the number of alternative exons within the total number of exons in the pathway. Shape and color define the disease. Point size defines -log10(p-value adjusted)

In summary, in pediatric sarcomas there exists a limited intersection of target genes between transcriptional and splicing alterations. Notably, alternative splicing primarily influences genes linked to transcription factors (TFs) and epigenetic modifiers (EMs), in addition to the MAPK signaling pathways, while transcriptional alterations target developmental processes, cancer genes and cell membrane proteins.

Identification of survival related alternative splicing events

Given previous findings that demonstrated a connection between ASEs and patients' prognosis (Li et al. 2021), my goal was to identify any splicing abnormalities in our patients that could potentially affect their prognosis. To accomplish this, I utilized differential splicing data from a previously published analysis that was conducted on the sarcomas cohort using TCGA data (Kahles et al. 2018). I stratified 181 sarcomas patients with available clinical data according to low and high cumulative event inclusion of each group (see Methods). I considered four types of events: exon skipping, alternative 3’ and 5’ and intron retention and then a univariate Cox proportional hazard model was used to estimate Hazard Ratio (HR) and log-rank p-value for each event. Events with log-rank p-value≤0.1 were considered as significant related to patient survival. TCGA events were intersected with alternative events as defined before by evaluating the junction similarity (see Methods). I found a highly significant enrichment of survival related events in the alternative events for both diseases (54 and 57 for OS and EW, respectively, Fisher's p-value <1e-12) (Figure 15).

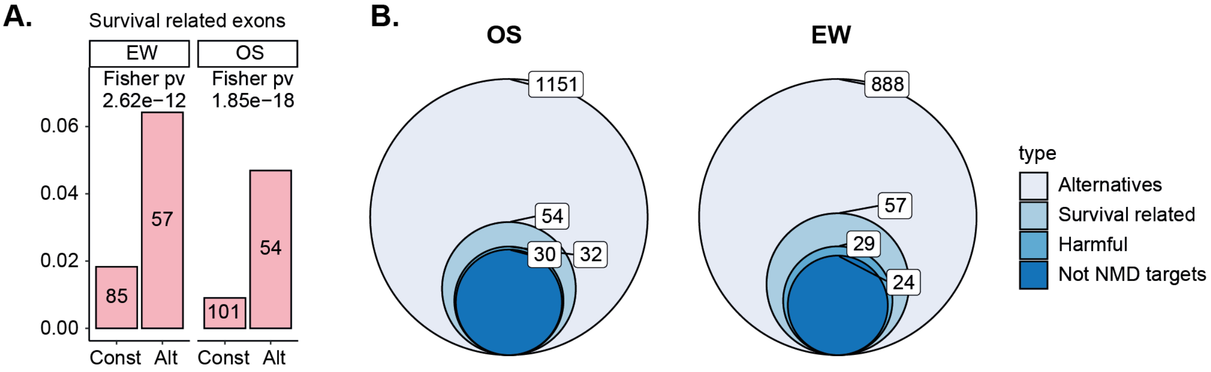

Figure 15: **A.** Proportion of survival related exons in alternative and constitutive exons. Number in the bar is the absolute number of exons. **B.** Venn Diagrams of the subsequent filters to retrieve damaging events candidates in the two diseases.

To obtain a list of potential events significantly correlated with patient survival, called hereafter as "damaging events", I employed a systematic filtering approach in order to ensure a high level of confidence in the results. First, I narrowed down the survival-related events to a specific set of exons characterized by detrimental levels of inclusion (Figure 15). In particular, I adopted a previously established approach (Del Giudice et al. 2022), wherein I defined as "harmful" those ASEs with DPSI>0 and HR>1 (bad prognosis with inclusion) or DPSI<0 and HR<1 (bad prognosis with exclusion) remaining with 32 and 29 events for OS and EW respectively.

To be more stringent with the prognostic candidates I wanted to further filter out transcripts that could potentially undergo degradation. In some cases, the inclusion or the exclusion of an exon alters the coding part of a transcript introducing a PTC. This phenomenon leads to the activation of a surveillance system that degrades transcripts with Premature Termination Codon (PTC) called Nonsense Mediated Decay (NMD) (Pervouchine et al. 2019).The exon that leads to PTC formation through inclusion is referred to as a "poison" exon, while its exclusion is termed an "essential" exon. Therefore, to delve deeper into this mechanism I intersected the cassette exons with the previously annotated poison and essential exons (Pervouchine et al. 2019). I defined poison exons with DPSI>0 or essential exons with DPSI<0 as "NMD targets" in sarcomas. The transcripts containing these events are presumably going to be degraded and therefore they are not going to produce an aberrant protein that could impact prognosis. Only a limited subset of harmful exons were targeted by NMD and filtered out.

Following this filtering process, a total of 30 events in OS and 24 events in EW were identified as having a negative association with patient prognosis. Notably, seven of these events were found to be shared between the two diseases, indicating potential commonalities in their impact on patient outcomes (Figure 16 and Table 1). These identified events represent potential candidates for SSOs targeting. Such targeting has the potential to induce degradation, ultimately favoring an improvement in patient prognosis.

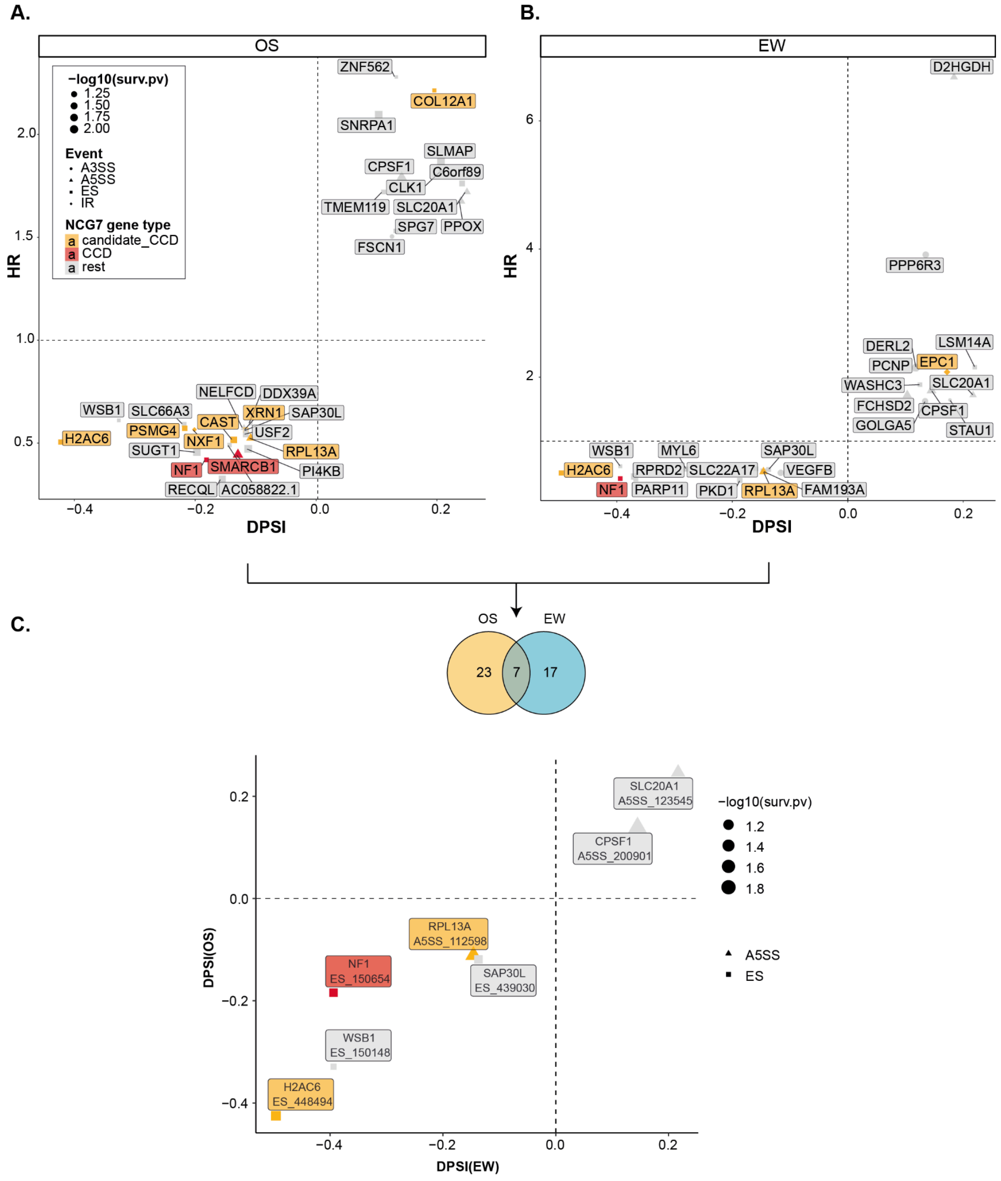

Figure 16: Selection of candidates of damaging events. **A-B**: Scatter plot between Hazard Ratio (HR) and DPSI in OS (A) and EW. **C.** Common damaging events between the two diseases and their HR-DPSI scatterplot.

Gene name	TCGA event id	surv.logrank_pval	surv.hazard_ratio	DPSI.OS	DPSI.EW	Alternative coordinates
CPSF1	alt_5prime_200901	1.3E-02	1.79	0.139	0.145	chr8:144398778-144398793
WSB1	exon_skip_150148	9.7E-02	0.61	-0.329	-0.394	chr17:27307743:27307919
RPL13A	alt_5prime_112598	3E-02	0.52	-0.111	-0.146	chr19:49490090-49489912
SLC20A1	alt_5prime_123545	5.2E-02	1.72	0.247	0.217	chr2:112652947-112652798
SAP30L	exon_skip_439030	8E-02	0.56	-0.119	-0.137	chr5:154452455:154452499
H2AC6	exon_skip_448494	6.2E-02	0.51	-0.425	-0.496	chr6:26127697:26127761
NF1	exon_skip_150654	8E-02	0.42	-0.184	-0.394	chr17:31252937:31253000

Table 1**:** Common damaging events candidates between OS and EW.

One of the common candidates, NF1 exon 23 (exon_skip_150654) that when skipped is associated with worse prognosis (Figure 16 and Figure 17), is an interesting candidate as it is a canonical cancer driver given NCG7 annotation, in particular a tumor suppressor gene. NF1 is a negative regulator of RAS activity pathway arising from the RasGAP domain where exon 23 falls in. RasGAP domain inactivates RAS making it shift from its active GTP-bound form to its inactive GDP-bound (Biayna et al. 2021; Tomazini and Shifman 2023). Exon 23 of NF1 gene is 63 nucleotides long, meaning that its skipping induces a loss of a part of the RasGAP domain. This could lead to a malfunctioning of RAS repression, yielding a RAS proteins activation. Hence, controlling the inclusion of NF1 exon 23 serves as a mechanism to ensure Ras signaling remains at the right levels (Hinman et al. 2014). RAS proteins have extensively associated with cancer since its mutations are found in 28% of human tumors (Liu, Xie, and Chen 2023; “Comprehensive Pancancer Genomic Analysis Reveals (RTK)-RAS-RAF-MEK as a Key Dysregulated Pathway in Cancer: Its Clinical Implications” 2019).

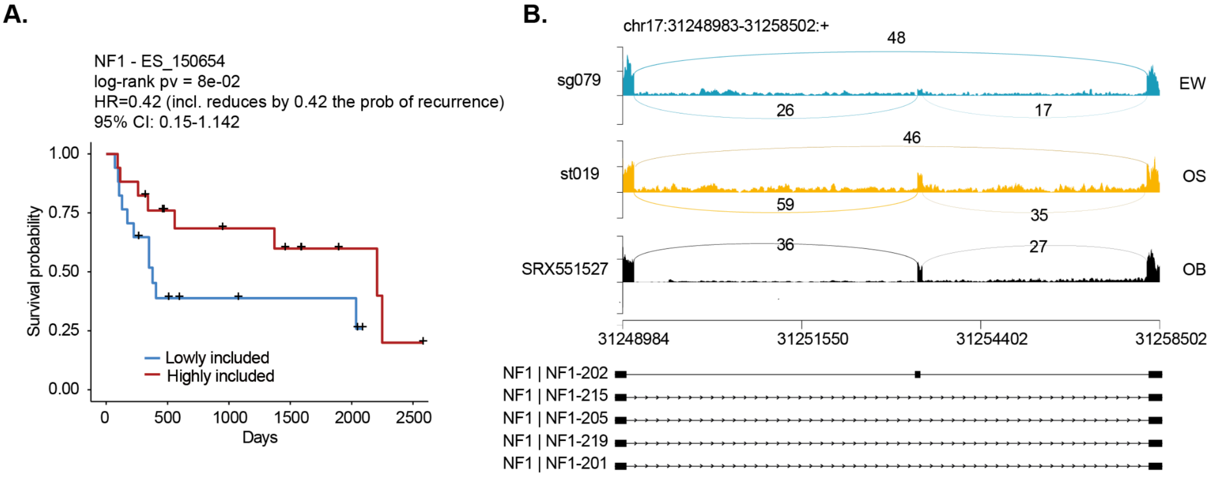

Figure 17: NF1 exon 23 candidate: **A.** Kaplan Meier curve stratified by inclusion level. Crosses represented censors. **B.** Sashimi plot of three samples, one EW, one OS, one OB respectively. Number of reads are written over the splicing junction. Below the sashimi there are all the isoforms of NF1 gene of the GENCODE comprehensive annotation.

Nonetheless, it is worth noting that the incorporation of exon 23 of NF1 has been associated with the activation of RAS/MAPK oncogenic pathways, as observed in High-grade diffuse glioma (Siddaway et al. 2022). Hence, the consequences of exon 23 inclusion may vary across different cancer subtypes.

To assess the behavior of NF1 exon 23 across different tumor types I retrieved data from Oncosplicing database (Yangjun Zhang et al. 2021) and checked the survival HR regarding its inclusion. Only tumors with matched normal were available in the database. Interestingly, NF1 exon 23 was always significantly associated with patient prognosis (log-rank p-value ≤ 0.1). However, the impact of its inclusion on survival outcomes varied depending on the specific tumor subtype. For instance, in Lung Squamous Cell Carcinoma (LUSC) and Esophageal Carcinoma (ESCA), its exclusion correlated with lower survival (Figure 18), as observed for sarcomas. Whereas, in Bladder Urothelial Carcinoma (BLCA), Brain Lower Grade Glioma (LGG) and Skin Cutaneous Melanoma (SKCM) its inclusion led to a low survival probability (Figure 18). This underscores the significance of exon 23 modulation across a diverse spectrum of cancers, with distinct effects attributed to its inclusion or exclusion.

The NF1 gene is also highly mutated in osteosarcomas and its loss promotes colony formation (Moriarity et al. 2015). Its amplification is mainly related to the tumor predisposition syndrome "Neurofibromatosis type 1" (Yap et al. 2014). This disease is thought to be the precursor of many cancer types, also some types of sarcomas, such as Ewing sarcoma (Fernandez et al. 2019; Chowdhry et al. 2009), osteosarcomas (Afşar et al. 2013) and rhabdomyosarcoma (Brems et al. 2009). Considering its correlation with prognosis and consistency between the two diseases, it is reasonable to think that a relationship might also exist with OS and EW as well.

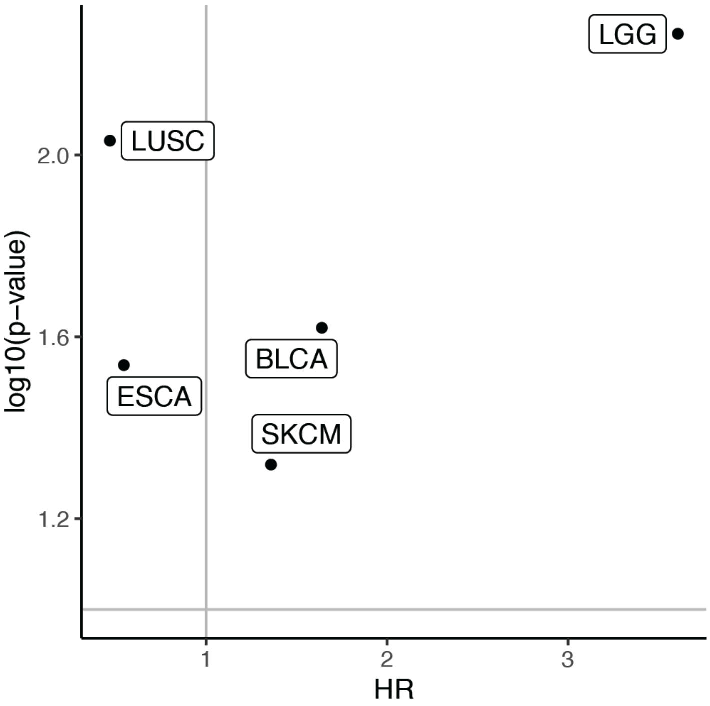

Figure 18: Log-rank p-value and Hazard Ratio from Oncosplicing database of five cancer types.

Methods

Differential expression analysis

Differential expression analysis was taken forward with two different approaches, similarly as previously done in unbalanced datasets (Del Giudice et al. 2022): exploiting a parametric method such as DESeq2 (Love, Huber, and Anders 2014) and a non parametric method using Wilcoxon test to account for the imbalance between groups. First, DESeq2 was used to compare OS vs OB and EW vs OB. Concomitantly, a Wilcoxon test was performed between the same group and a permutation test was used to retrieve the null distribution. Empirical p-value was defined as follows:

$$p.emp \ = \ \frac{1+\#(p_{permutation} < p_{wilcoxon})}{1+\#permutations}$$

where $p_{wilcoxon}$ is the p-value is Wilcoxon p-value between the non permuted groups, $p_{permutation}$ is the Wilcoxon p-value of a single permutation. Differentially Expressed Genes (DEGs) were defined as those with DESeq2 p-value adjusted ≤ 0.01, absolute log2 Fold Change > 2 and Wilcoxon p-value adjusted ≤ 0.1 and p.emp ≤0.1.

Gene annotations and enrichment analysis

Transcription factors (TFs), RNA-binding proteins (RBPs) and Epigenetic Modifiers (EMs) lists were manually curated exploiting different studies (Huether et al. 2014; Van Nostrand, Freese, et al. 2020; Gerstberger, Hafner, and Tuschl 2014; Attig et al. 2018; Lambert et al. 2014). Canonical Cancer Drivers (CCD) genes and putative CCD were retrieved from the Network of Cancer Genes (NCG7.1) (Dressler et al. 2022). Development related genes were retrieved from parsing the term “development” among all the ten branches of Biological Processes of Gene Ontology (GO:0008150) using ‘GO.db’ v3.16.0 R package. Ligands and receptors gene names were downloaded from FANTOM website (Lizio et al. 2019). Enrichment of DEGs versus the non significant genes in each annotation class was performed through a one-tailed Fisher’s enrichment test using *fisher.test* function in ‘stats’ R package v4.2.3.

Binomial Over-representation analysis

The binomial over-representation analysis is evaluating the probability for a list of genes to have at least X number of alternative exons, among all the possible exons Y in the same list (or pathway), considering as background probability $p_0 = \frac{K}{N}$ where K is the total number of alternative exons and N is the total number of exons across all pathways. In particular, the probability P of having exactly $X = \sum_{i=1}^{m} k_i$ number of alternative exons in a pathways with a total number of exons $Y = \sum_{i=1}^{m} n_i$:

$$P(X\ among\ Y\ in\ the\ pathway) = \binom{Y}{X} {p_0}^{X}(1 - p_0)^{X-Y}$$

where n_i total number of exons (sum of constitutive and alternative) of gene i and k_i total number of alternative exons in gene i, while *m* is the total number of genes in the considered pathway. The p-value is the probability of having at least X alternative exons among Y exons, considering as distribution under the null hypothesis the binomial distribution Bin(Y, p) with $p = p_0$. P-values are then corrected for multiple tests using the false discovery rate (FDR) by the Benjamini–Hochberg method.

“Gene Ratio” term is defined as the ratio between alternative exons in the pathway and all the alternative exons.

Alternative spliced events

The software rMATS v4.1.1 (Shen et al. 2014) was used to retrieve splicing events with the --novelSS flag to allow for the discovery of novel junctions. rMATS was run for OS and EW samples using the OB as control. A non-parametric test, Wilcoxon test, was also used to account for the unbalanced groups as previously proposed(Del Giudice et al. 2022). A permutation test was used as done for differential gene expression to retrieve the null distribution. I defined as Alternative Spliced Events ASEs those with rMATS FDR<0.1 and |DPSI|>0.1 and with Wilcoxon p-value ≤ 0.05 and p-emp ≤ 0.05 (however all of the events with Wilcoxon p-value ≤ 0.05 had the p-emp below as well). The constitutive events were those with rMATS FDR ≥ 0.1 and |DPSI| ≤ 0.1 which were used as controls.

Sashimi plots were derived using ‘trackplot’ python script (Yiming Zhang, Zhou, and Wang 2022) setting the minimum number of reads supporting the junctions to 10 (-t 10).

Survival analysis

Differential splicing of the TCGA database elaborated in a previous pan-cancer analysis (Kahles et al. 2018) were downloaded from Genomic Data Common. Only sarcoma patients were selected, *i.e.* barcoded as ‘SARC’. Only patients belonging to the previously annotated whitelist were considered (Kahles et al. 2018). Since TCGA data were mapped to hg19 coordinates, all the rMATS events were transposed to hg19 using LiftOver R package v1.22.0. All the exons coordinates involved in the event were transformed, in the case at least one coordinate was not found in the target genome, the whole event was discarded. A TCGA event was considered equivalent to an rMATS event if they shared the alternative regions and the internal splice sites of the flanking exons. For example, for the exon skipping case, the coordinates considered for the intersection were the start and the end of the alternative exon and the end of the upstream and the start of the downstream exon.

Disease-free survival (DFS) was defined as the time between primary treatment and the diagnosis of disease progression, as defined by biochemical or clinical recurrence, or the end of follow-up. Patients were stratified into high and low expressors based on the 25th and 75th percentile of the exon PSI distribution. Exploiting this stratification, survival analysis was performed by fitting a univariate Cox proportional hazards model with log-rank test (Therneau and Grambsch 2013) using the *coxph* function in the R ‘survival’ package. To ensure an adequate number of observations, only events that had a minimum of five patients in both of the two inclusion groups were taken into account.

Chapter 2: RNAmars - identification of multivalent RNA Motifs And cognate Regulators of alternative Splicing

Experimental methods for identifying RBPs-RNA interactions are traditionally based on RNA immuno-preciptation (RIP) (Keene, Komisarow, and Friedersdorf 2006) and ultraviolet cross-linking and immunoprecipitation (CLIP) (Ule et al. 2003; Underwood et al. 2005), enabling RBPs binding site identification in their cellular contexts. Recently, the Encyclopedia of DNA Elements (ENCODE) developed an enhanced CLIP protocol (eCLIP-seq) to reveal RNA-RBP interactions of more than 150 different RBPs (Van Nostrand, Freese, et al. 2020). Both methods have a low resolution (50-500nt) compared to the short motifs on which RBPs bind (Carazo, Romero, and Rubio 2019). Integrating sequence information can lead to a finer identification of binding. *In silico*, this can be performed with RNAcompete (Ray et al. 2009) which uses a pool of randomly generated k-mers to determine the high affinity RNA sequence of an RBP. *In vitro*, instead, binding specificities can be assessed with RNA Bind-N-Seq (RBNS) where recombinant purified RBPs reacting with pools of random RNA oligonucleotide are high-throughput sequenced. RBNS facilitates the identification of RBPs binding motifs but lacks regulatory activity information (Lambert et al. 2014).

However, these experimental methods are highly expensive and time consuming. The development of computational methods is therefore essential to conveniently decipher RBPs-RNA interactions. Various algorithms have been developed to discover *cis*-acting motifs of RBPs. For instance, MEME software enables the motifs discovery starting from unaligned RNA sequence (Timothy L. Bailey 1994). This software has then been improved by GLAM2 which identifies gapped motifs (T. L. Bailey et al. 2009). Another recent tool which enables gapped motifs discovery is RBPmap (Paz et al. 2022). RNAmotifs (Cereda et al. 2014) assesses clusters of repeated tetramers enrichment (multivalent motifs) allowing for higher-order dependencies.

While these softwares are exclusively based on the raw sequences, mCROSS (Feng et al. 2019) model RBP binding specificity exploiting CLIP data and resulting in a list of position weight matrices (PWMs).

Several recent approaches exploit neural networks and deep-learning to overcome the complexity of the task. Some methods such as DeepPN are based on the sequence information alone (J. Zhang et al. 2022). Other methods such as iDeepS, DeepRKE and DeepRiPe integrate multiple convolutional neural networks to gather different RNA features together, e.g. RNA sequence, secondary structure or transcript region (Pan et al. 2018; Deng et al. 2020; Ghanbari and Ohler 2020).

These deep learning algorithms improve accuracy of predictions at the expense of interpretability. Moreover, these methods do not take into consideration splicing alterations. There are many softwares, such as rMATS(Shen et al. 2014), Whippet (Sterne-Weiler et al., n.d.), DiffSplice(Hu et al. 2013) that successfully identify alternative splicing events between two conditions starting from RNA-seq data. Nevertheless, these tools do not give insights on the actual regulatory mechanisms for which these alterations happen. A study from Sebastién et al. extensively evaluated the associations of alternative splicing events and RBP in TCGA data through motifs discovery and network analysis between their expression and the splicing pattern (Sebestyén et al. 2016).

CoSpliceNet is another method to derive associations between splicing factors and their putative mature transcript product via a co-expression and de novo motif prediction at the splice junctions (Aghamirzaie et al. 2016).

However, these algorithms explore RBP-RNA interactions only through motif discovery, which can be limiting considering that some RBPs can have overlapping consensus (Fu and Ares 2014). Carazo *et al.* showed that using CLIP data narrows down the list of putative regulatory RBPs of alternative splicing (Carazo et al. 2019). A more precise strategy could be to integrate both the sequence information and eCLIP experiments. However, to my knowledge, no technique exists for predicting RBPs utilizing both eCLIP experiment data and motifs data.

RNAmotifs (Cereda et al. 2014), a software developed for the identification of clustered and repetitive tetramers near and within alternative cassette exons, identifies *cis*-acting elements related to input exons. However, it does not link the *cis*-acting elements (*i.e.* motifs) to the putative *trans*-acting factors (*i.e.* RBPs). The user has to manually inspect the resulting motifs and compare them to known PWMs to have insights into the putative regulatory RBP. A robust and systematic method to overcome this problem was still to be developed. In this section, I propose RNAmars an highly interpretable algorithm for the identification of RNA Motifs And cognate Regulators of Alternative Splicing (RNAmars). RNAmars is an extension of RNAmotifs, that enables the discovery of the RBPs underlying the alternative splicing regulation. RNAmars exploits both the *cis*-acting element enrichments and the *trans*-acting element preferences, taking advantage of eCLIP data from ENCODE (Van Nostrand, Freese, et al. 2020). Learning from binding sites and sequence preferences, RNAmars provides a deeper understanding of cooperation mechanisms that regulate alternative splicing. These findings are consistent with known motifs and binding preferences and contribute to reveal RBP-RBP interactions.

Results

Identification of RBPs-specific binding regions

To associate motifs to RBPs, I retrieved alternative splicing events upon RBP silencing (short-harpin) and iCount peaks of eCLIP for HepG2 and K562 cell lines from the ENCODE database (König et al. 2010; Van Nostrand, Freese, et al. 2020). Since both exons coordinates and binding sites were needed, I kept only RBPs for which both the experiments were performed, for a total of 70 and 80 RBPs in HepG2 and in K562, respectively (Figure 19). Moreover, to avoid RBPs lacking binding specificity, I retained only the experiments for which the PWMs were available in mCROSS database (Feng et al. 2019), remaining with 64 and 72 RBPs for HepG2 and K562 (Figure 19). Finally, only RBPs containing at least 50 ASEs were retained, for a total of 15 RBPs and 13 RBPs in the two cell lines (Figure 19). Seven RBPs were common between the cell lines, namely U2AF1, U2AF2, PTBP1, SRSF1, PRPF8, HNRNPU and SF3B4.

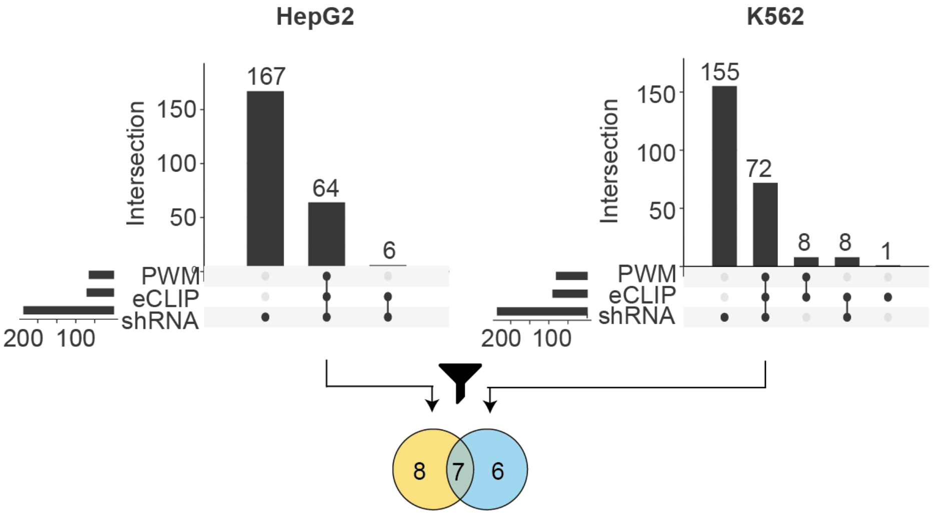

Figure 19: Number of proteins in the databases for HepG2 cell line (left) and K562 cell line (right). PWMs are retrieved from the mCROSS database. eCLIP and shRNA are collected from ENCODE database

The majority of the RBPs are splicing regulators or spliceosome components (Figure 20), according to a previously curated annotation (Van Nostrand, Freese, et al. 2020). I noticed that, in both cell lines, U2AF2 was among the RBPs with the highest number of regulated events (Figure 20), 603 and 550 ASEs for HepG2 and K562 respectively, which was expected since U2AF2 is required for the binding of U2 snRNP to the pre-mRNA branch site

(Fu and Ares 2014). For four proteins no annotation was available. In particular, two of them (UCHL5 and AGGF1) are known as “Novel RBP”, for which the function is yet unknown.

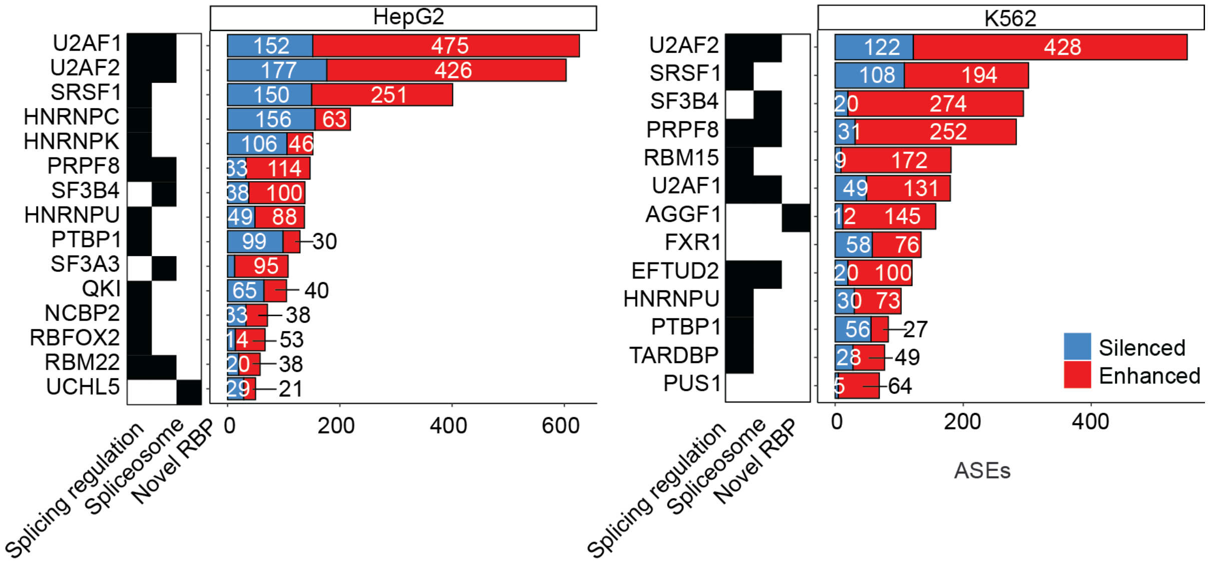

Figure 20: Number of alternative exons in the two cell lines. Color code defines the direction of inclusion promoted by the protein. Left heatmaps are annotations of the RBP type, black if protein belongs to the class, white if not.

To properly decipher motif enrichments around the RBP regulated exons, I had to take into account the binding clustering properties of RBPs. In particular, RBP domains bind to repeated and degenerated motifs that define their specificity (Stitzinger, Sohrabi-Jahromi, and Söding 2023; Jolma et al. 2020; Dominguez et al. 2018). These motifs are about four nucleotides (nts) and spaced a few bases apart (Cereda et al. 2014). RNAmotifs software (Cereda et al. 2014) suited very well in discovering these types of motifs. Specifically, RNAmotifs is a software that accounts for the ability to bind to multiple proximal and degenerate motifs (*multivalent* motifs). To do so, the algorithm identifies clusters of tetramers within scrolling windows centered on each considered position at exon-intron boundaries of alternative exons. In particular, two internal parameters define the regulatory region features: the scrolling window *n* and the enrichment window *e*. The scrolling window defines the region in which the search for the repeated multivalent motifs is done, controlling in this way the tetramer sparsity at each position; the enrichment window *e* is the region around the alternative exons where to look for the tetramer. In other words, the algorithm evaluates the enrichment of the tetramers’ instances across three regions (R_1, R_2 and R_3) surrounding the splice sites of the alternative exons, compared to control ones and the enrichment window *e* controls the width of the three regions, defining the alternative splicing regulatory principles of RBPs.

Since different RBPs can have different regulatory regions, I decided to fine-tune RNAmotifs parameters to retrieve the optimal *n* and *e* for each RBP (see Methods). Briefly, I evaluated the strength of the association between each tetramer and RBP using a weighted cosine similarity (see Methods). Such similarity was defined as Association Score (AS), ranging from 0 to 1, where 1 means that the protein is probably binding to that multivalent motif to promote the alternative splicing event. It is the result of the multiplication between two terms: (i) a profile similarity between the tetramer splicing maps and the binding profiles and (ii) a signal recovery rate, that measures the ability for a dataset to recover the final binding profile given a subset of randomly picked exons (see Methods). Binding profiles are derived from a statistical procedure in which an enrichment test of binding peaks in alternative exons versus constitutive ones is performed (see Methods). The derived profile similarity is specific to each tetramer-RBP pair and evaluates the consistency of the sequence signal with the expected binding signal (see Methods). On the other hand, the signal recovery rate does not rely on RNAmotifs results, as it depends only on the regulated exons and the eCLIP peaks. Two signal recovery rates for each protein were evaluated, for silenced and enhanced exons respectively. The signal recovery rate goes from 0 to 1 and penalizes weak binding profiles (*i.e.* those for which subsets of exons are not able to accurately represent the final signal). In particular, the mean value of the signal recovery rate distribution was 0.78 with a maximum of 0.92 for HNRNPC silenced exons in HepG2 which shows high robustness in binding profile (Figure 21). Interestingly, U2AF1 and U2AF2 had higher signal recovery rate in silenced exons, even though the majority of their regulated exons were enhanced. This happened probably because the maximum Binding Strength (BS) (see Method) occurred in silenced exons, promoting a stronger binding signal in these sets of exons.

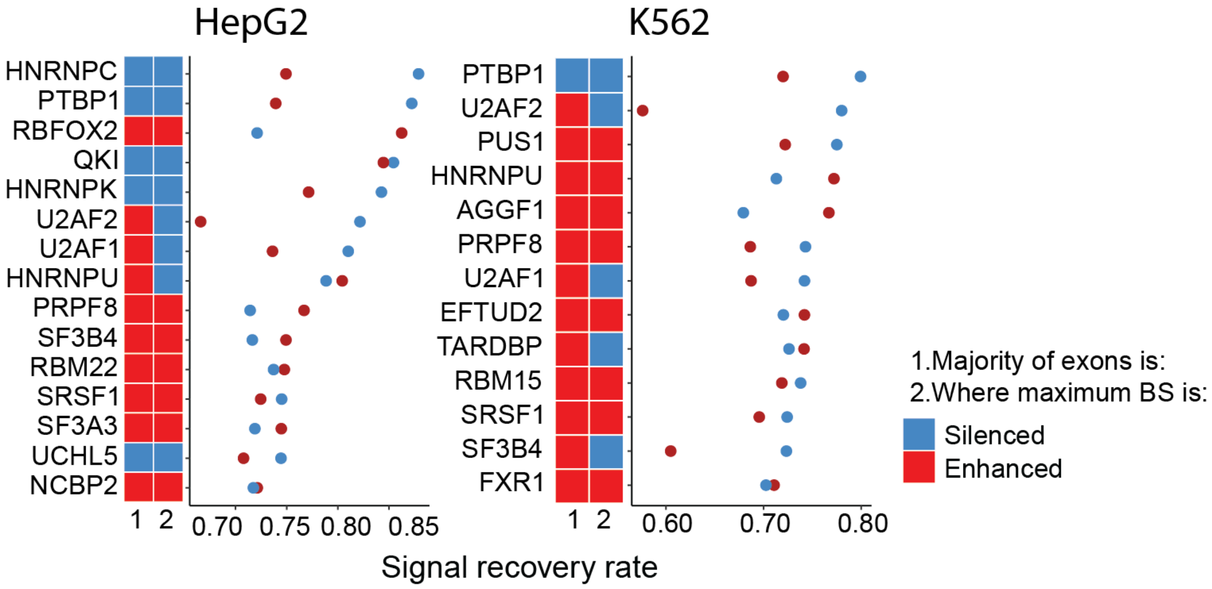

Figure 21: Signal recovery rate for RBPs in the two cell lines and separated by enhanced and silenced set of exons (point color code). Lateral annotation indicates in the first column if the exons are prevalently silenced (blue) or enhanced (red). The second column whether the maximum binding strength appear in silenced (blue) or enhanced (red) exons

Given T tetramers deriving from a RNAmotif run and P evaluated proteins, ASs were computed for each pair, yielding a PxT matrix, defined as AS matrix. In the set of exons regulated by a specific RBP, the ASs of enriched tetramers found by RNAmotifs and the same RBP are expected to be the highest, while the ASs of the same tetramer and the other protein are expected to be lower. This concept can be thought of as a classification problem: each pair of parameters (e, n), *i.e.* each RNAmotifs run, is related to the AS matrix specific to tetramers found in that run.

For each AS matrix I evaluated whether the AS values exceed a threshold for the expected RBP and whether it remained below for the rest of RBPs (see Methods). The optimal set of parameters were given by the run that yielded the highest AUROC for each RBP (see Methods). The analysis produced a total of 14 unique sets of parameters across all datasets (Figure 22). I noticed that some RBPs, such as PTBP1, HNRNPC, HNRNPK, SRSF1 and U2AF2, presented almost invariant AUROC across parameters, meaning that the AS were not dependent on clustering and enrichment windows for these proteins (Figure 22). In fact, the corresponding AUROC curves changed by just a few percentage points among all the parameter pairs (Figure 22). However, the majority of the proteins were highly affected by the choice of parameters. For example, QKI had an optimal enrichment window of 100 nts, same as RBFOX2 (Figure 22). This was expected since both of them are known to have a large binding in the downstream region and to cooperate during the splicing process (D. Zhou et al. 2021; Van Nostrand, Freese, et al. 2020). HNRNPK and TARDBP are known to have a wide intron binding (Van Nostrand, Pratt, et al. 2020) and this was reflected in the enrichment window of e=300 and e=200 nts, respectively (Figure 22).

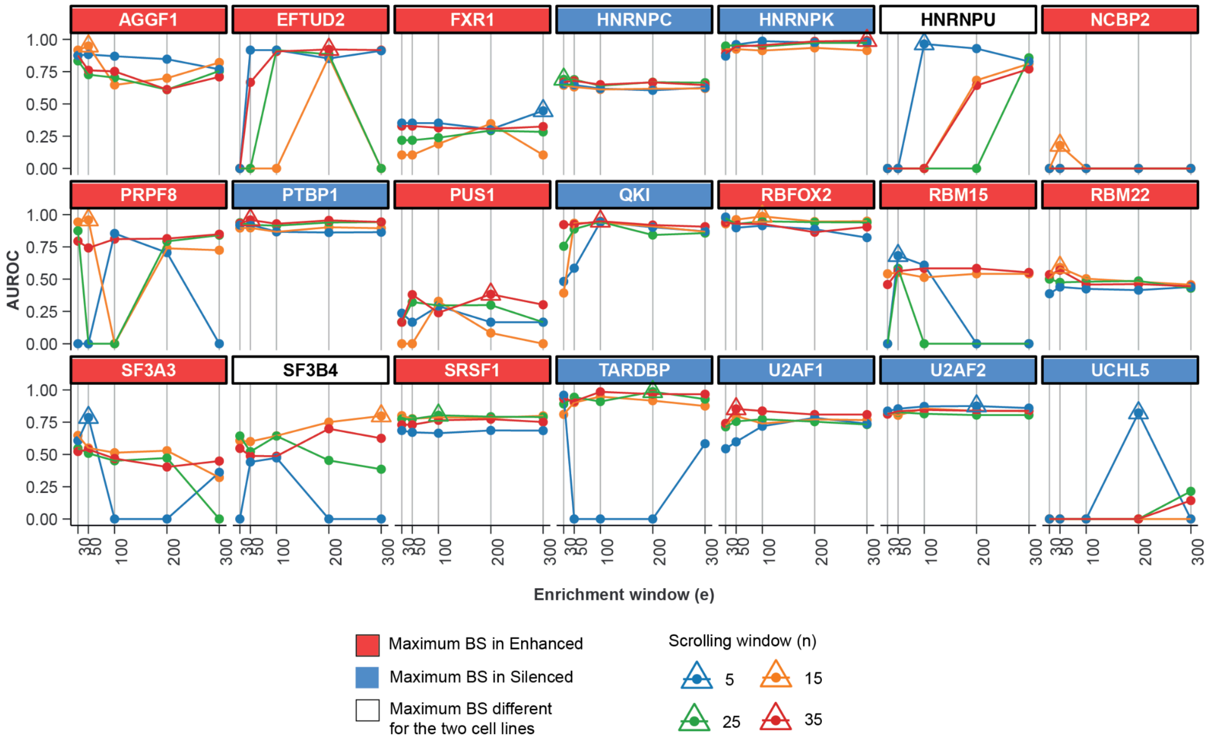

Figure 22: AUROC scores (y-axis) for different values of the scrolling window n (line color code) and enrichment window e (x-axis). Triangles define the highest score within the RBP. Facet color indicates whether the maximum BS is found in enhanced exons in both cell lines (red), silenced exons in both cell lines (blue), or in different regulations type across the two cell lines.

Characterization of RBPs binding and motifs preferences

Given RNAmotifs optimal parameters specific for every RBP, I proceeded to validate whether the tetramer enrichments in these optimal runs aligned with the expected binding profile. In other words, my objectives were (i) confirming that the tetramer enrichments were located in the same regions as the binding enrichments of the respective RBP, and (ii) to assess the motifs peculiarity of each dataset to check their agreement with expected sequence specificity of RBPs.

To evaluate the compatibility of tetramer enrichments with binding profiles I compared the region-wise enrichments (tetramer strength) with the Binding Strength (BS) of the same protein within the same cell line (see Methods). In detail, tetramer strength was defined as the cumulative tetramer scores in R_1, R_2, R_3 in the silenced and enhanced exons, while BS was proportional to the maximum binding profile enrichment in the same regions and regulation type (see Methods). The two six-dimensional vectors, representing tetramer strength and BS, were compared using the cosine similarity (Figure 23). In figure, each vector was normalized to one for visualization purposes. Fourteen datasets presented a high cosine similarity with values above 0.75. For HNRNPU and PUS1, no tetramer resulted as enriched under the optimal parameters, probably due to the low number of input exons (101 for HNRNPU and 69 for PUS1). Interestingly, among the highest similarity datasets (similarity ≥ 0.75), the most frequent region of interest in which both the maximum binding strength and maximum tetramer strength were found was R_1 for silenced exons in 9 over 14 cases (Figure 23). I next wondered whether datasets with enrichments in the same regions also shared a sequence preference, therefore I created PWMs starting from tetramers scores (see Methods). Surprisingly, I observed that in cases in which the maximum BS and tetramer strength was R_1 in silenced exons, the tetramer PWMs were mainly pyrimidine-rich or T-rich (Figure 23).

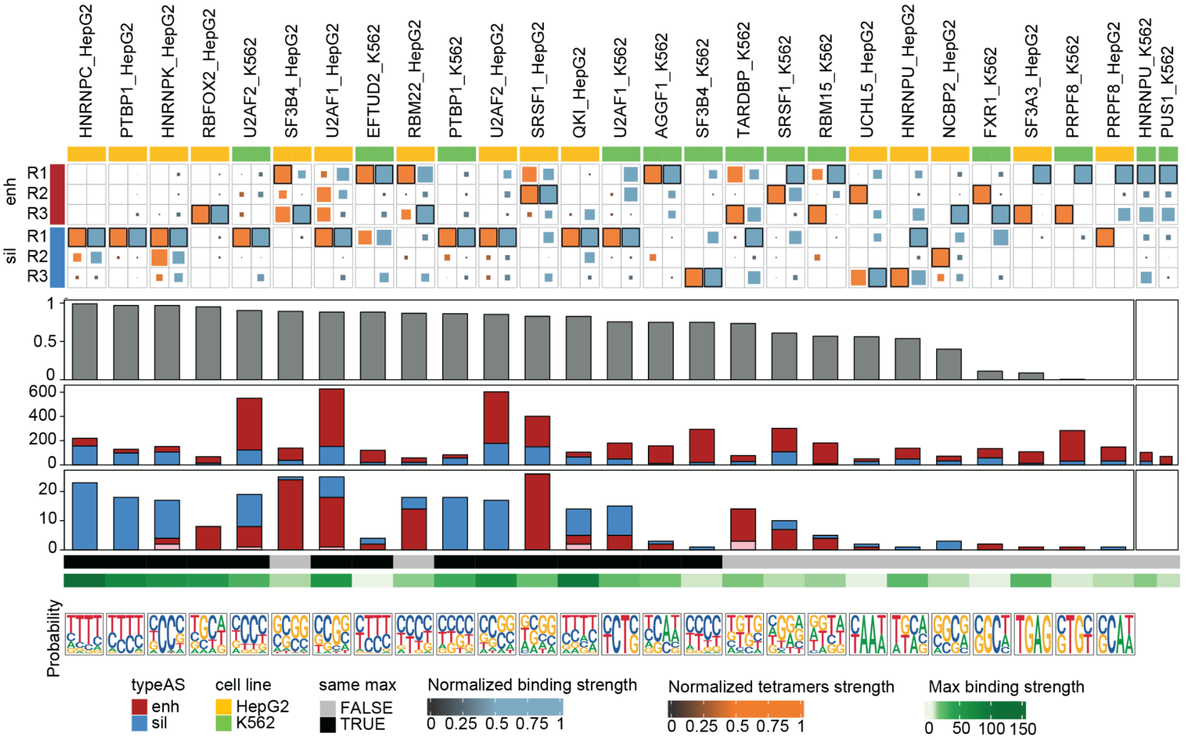

Figure 23: Comparison between tetramer strength and binding strength: Heatmap collects binding strength (in blue) and tetramer strength (in orange), the maximum value is highlighted with black border. Along the column there are the 28 input datasets *i.e.* 21 proteins among HepG2 and K562 cell lines, along the rows there are the regulatory regions of the alternative splicing event. Gray barplot indicates the cosine similarity between tetramer and binding strength. The two bottom barplots are the number of input events and number of retrieved tetramers in the dataset (pink in tetramers barplot indicates that the tetramer is enriched in both enhanced and silenced exons). Black cells indicate whether the maximum binding strength is in the same region of the maximum tetramer strength. Green cells are the non normalized values of the maximum binding strength. Bottom annotation is the list of logos reconstructed from the enriched tetramers (see Methods).

This finding reflects some already known interaction mechanisms of RBPs. For example, PTBP1 and U2AF1 have similar binding affinity to pyrimidine-rich tract that can result in actual competition for binding at the 3’ splice site (Izquierdo et al. 2005; Saulière et al. 2006). HNRNPC, instead, binds to the upstream region to uridine rich sequences to prevent exon inclusion (König et al. 2010) and it is also known to compete with U2AF1 for binding (Zarnack et al. 2013). Overall, by collapsing all the enriched tetramer in the 6 regions (see Methods), an enrichment of pyrimidine-rich PWMs in silenced exons and a minor enrichment of serine-rich PWMs for enhanced exons emerged (Figure 24). In both silenced and enhanced exons the Information Content (IC) of tetramers PWMs was higher in the R_1 region, in other words there was a stronger sequence signal in the upstream intron region (Figure 24).

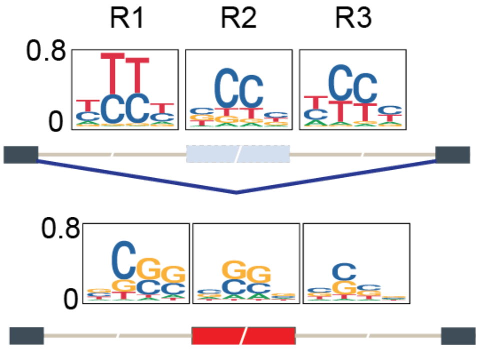

Figure 24: tetramers derived PWMs in the three regulatory regions for silenced and enhanced separately. Y-axis is the Information Content.

The similarity between the binding strength and the tetramer strength became weaker as fewer enriched tetramers were found. Under this light, I wondered whether the worsening of the compatibility between motifs and binding enrichments were due to intrinsic aspects of the dataset. I found a high correlation between the cosine similarity and both the number of enriched tetramers (r=0.71, pv=4.0e-4) and the overall tetramer score (r=0.56, pv=2.5e-3) within the dataset (Figure 25). Cosine similarity also correlated with the maximum binding strength of the same protein (r=0.52, pv=6.4e-3) (Figure 25), which in turn showed a relationship with overall tetramers score (r=0.69, pv=1.1e-4). These results suggested that in the cases where a weak sequence signal (*i.e.* overall tetramer scores) is observed, the binding signal (*i.e.* binding strength) is also weak.

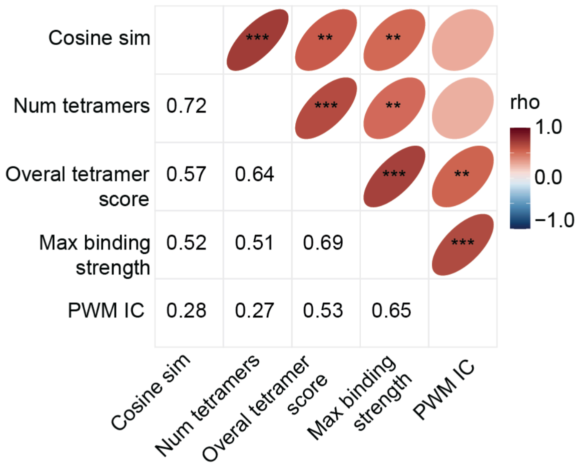

Figure 25: Pearson's correlations of cosine similarities between tetramer strength and binding strength and other features of the input dataset: number of tetramers, overall tetramer score, maximum binding strength, PWM information content (IC). Lower triangular matrix is the value of Pearson's correlation coefficient. Upper triangular matrix is the p-value from the correlation test (meaning: *** < 10e-3, ** < 10e-2)

To better understand the sequence specificity of an RBP, I made a list of tetramers for each RBP across the two cell lines and performed an intersection between pairs of proteins. I observed that some proteins, such as U2AF2, U2AF1 and SRSF1, had many tetramers in common with others (80%, 79%, 64% tetramers are shared, respectively). On the contrary, other RBPs showed a highly specific tetramer list, such as TARDBP, that presented 14 enriched tetramers and only 4 of them were shared with others, or HNRNPC (5 shared tetramers among 23) (Figure 26).

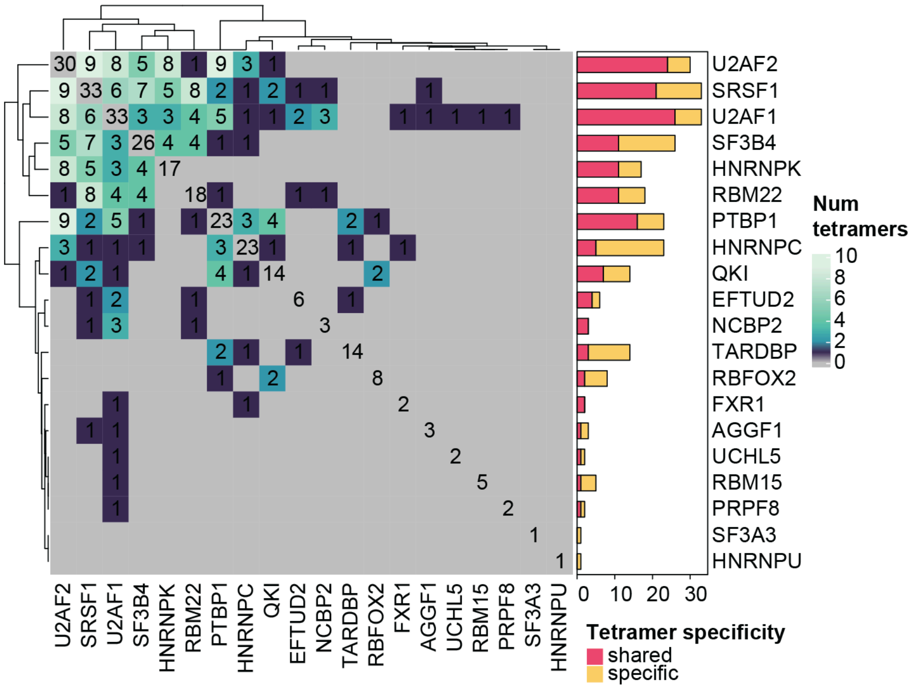

Figure 26: Number of common significant tetramers between each pair of RBPs. Lateral barplot identifies the number of shared or protein specific tetramers

Pipeline overview and performance evaluation

Considering the reliability of RNAmotifs results in terms of sequence and region enrichments, I developed the RNAmars algorithm to link tetramers with the cognate RBPs. Given a set of alternative and constitutive exons, RNAmars aims at retrieving putative RBP regulating the differential inclusion (see Methods). RNAmars is able to identify the splicing regulators starting exclusively from exons coordinates. This allows users to perform solely RNA-seq experiments upon conditions, dismissing the necessity of conducting expensive eCLIP experiments.

The idea behind RNAmars is to connect *cis*-acting elements' enrichments with *trans*-acting factors' enrichments. The first is given by the position-wise enrichment scores of RNAmotifs tetramers (tetramer specific splicing map), while the second is retrieved from ENCODE RBPs binding profiles (see Methods). The strength of the association between tetramers and RBPs is given by the AS (see Methods). The final result of RNAmars is a matrix in which each column corresponds to an enriched tetramer and each row is a protein. The proteins are ranked from the top-associated to the lowest (see Methods).

Given a set of exons regulated by a specific RBP, the top associated protein is expected to be the RBP itself. Therefore, to evaluate RNAmars performance I used input exons regulated by the RBPs discussed in the previous section, for which I found the set of optimal parameters. Since tetramers are actually clusters of degenerated sequences of four nucleotides, and they are not specific to only one protein, it can be possible to have strong associations with multiple RBPs, nevertheless I expected that the strongest associated protein was the actual regulatory one.

As an example of the resulting visualization of RNAmars, I showed PTBP1 silenced exons heatmap in Figure 27. Considering that the PTBP1 splicing map is well established, characterized by Pyrimidine motif enrichment at 3′ splice sites of silenced exons, I thought it could fit well as a benchmark for explanatory purposes (Cereda et al. 2014). RNAmars accurately predicted PTBP1 as the major regulator, since it had the highest association score mean across tetramers (Figure 27). PTBP1 was mainly associated with pyrimidine-rich tetramers found in the upstream region, consistently with the already known PTBP1 exon skipping mechanisms (Llorian et al. 2010; Saulière et al. 2006; Xue et al. 2009).

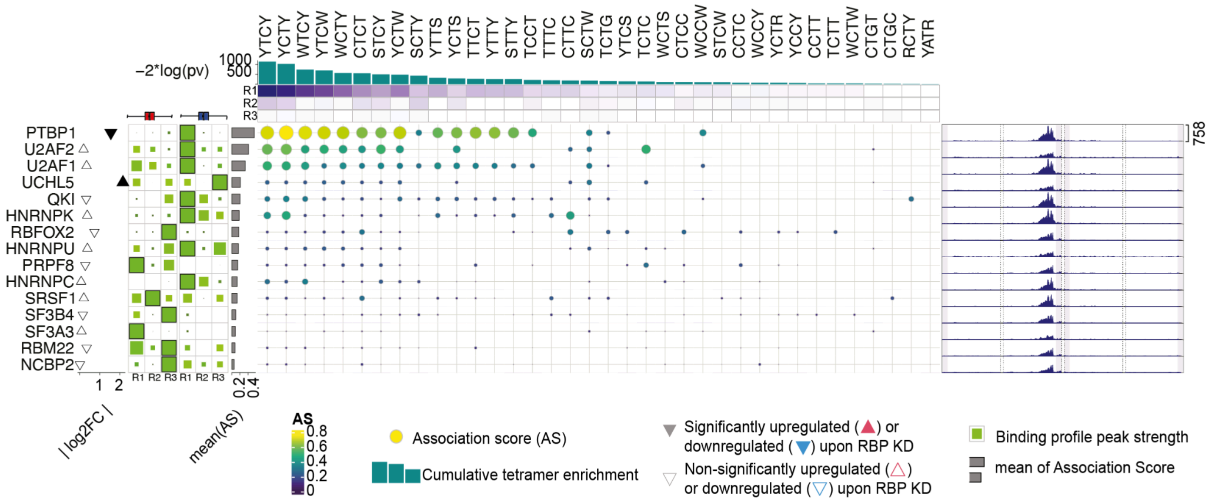

Figure 27; Final visualization heatmap for PTBP1 regulated exons. Enriched tetramers in the columns, RBPs in the rows and matrix values are the association scores. Point size and color gradient are proportional to the AS. Left annotation contains the absolute log2 Fold Change of differential gene expression. Triangle pointing up means up-regulated after the silencing, pointing down means down-regulated after the silencing (sign of gene expression is user defined). Green squares on the left are the BS of each protein (binding peaks are taken from the HepG2) cell line. Left barplot indicates the final RBP score computed as the mean(AS). Top barplot is the cumulative tetramer of the tetramer (Fisher's Method aggregation of p-values from different regions and runs). Upper annotation heatmap refers to the tetramer strength in the three regulatory regions. Tetramer splicing maps on the right in the RNAmotifs run with the optimal parameters of the row protein.

The results for all the proteins are summarized in Figure 28. The number in the cells represents the position in the ranking of each predicted protein, where 1 means that the predicted protein is the best predictor for the considered dataset. Only datasets for which at least one enriched tetramer was found are shown, therefore datasets for which it was possible to compute the association matrix, for a total of 39 records over 56 records. Among the 39 datasets, 17 of them were top-associated with the actual regulatory protein, which is ranked as first. Considering only the regulation type in which the maximum BS is found, 14 out of 24 datasets were perfectly predicted (Figure 28). Nevertheless, 11 datasets presented the actual regulatory protein ranked after the third place.

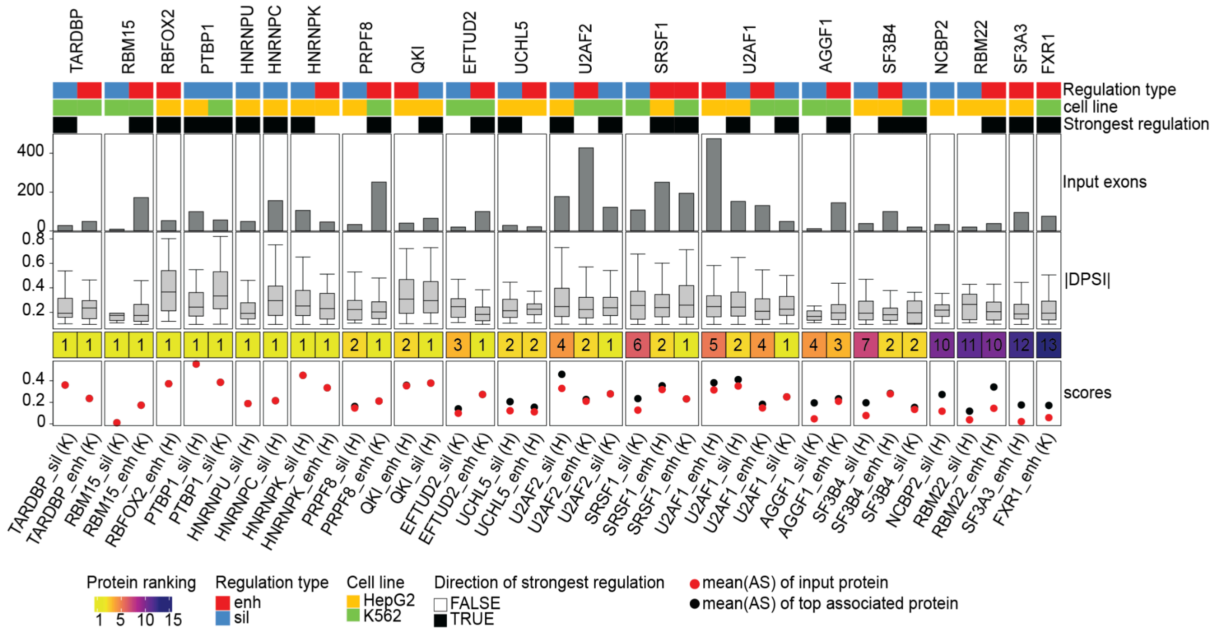

Figure 28: Final scores of each dataset. bottom annotation represents the mean(AS) of the protein regulating the input exons in red and the maximum mean(AS) in black if it is the score of another protein. Numbered cells are the ranking of the proteins in the heatmap, ordered by decreasing mean(AS). Distributions of input exons absolute DPSI depicted by boxplots. Barplot is the total number of input exons. Black cells annotation indicates whether the regulation type is where the maximum BS is. Top annotations indicate cell lines and regulation type.

I wondered whether the low accuracy could be due to the technical features such as the |DPSI| distributions and number of input exons, since the stronger the signal, the easier it might be to find the actual regulators. I also asked whether the motif strength could impact on RNAmotifs tetramers detection and the transcriptional secondary effect of the proteins silencing could pollute the alternative exons with other proteins' targets. To answer this question I built a linear model, fitting the final protein score as a function of mean(|DPSI|), number of input alternative exons, IC of the selected mCROSS PWM and number of differentially expressed RBPs (see Methods). I found that mean(|DPSI|) was the strongest positive predictor of the final protein score (Figure 29A). I also measured the relative importance of regressors, that confirmed the importance of mean(|DPSI|) (Figure 29B). Pearson's correlation analysis also showed a significant correlation between the final score and the mean(|DPSI|) (R=0.52, 3.4-05) (Figure 29C).

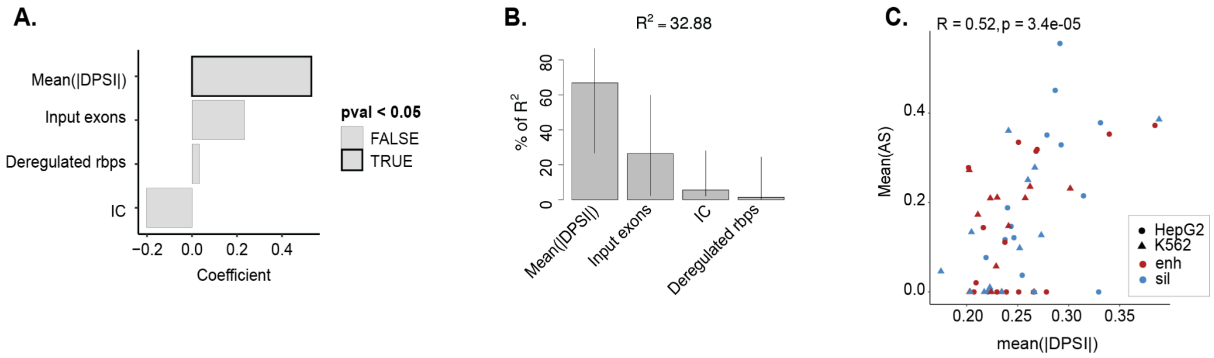

Figure 29: Linear model analysis. **A.** Coefficients of the linear model having as dependent variable the mean(AS) of the protein of the input exons and as features the y-axis variables. **B.** Bootstrap results with percentage of R^2 for each regressor. **C.** Scatter plot between mean(AS) of the input protein and the mean absolute DPSI of the input exons.

Testing RNAmars on HNRNPK silencing in PC3 cell line

Having proved RNAmars accuracy on ENCODE data, I was interested in testing its robustness with unpublished data. Thus, we performed a RBP silencing experiment followed by RNA-seq, to see whether RNAmars was capable of predicting the actual regulator (see Methods). In particular, we were interested in evaluating the RNA binding protein HNRNPK, since we previously discovered that HNRNPK is upregulated by the oncogenic transcription factor FOXA1 (Del Giudice et al. 2022), which is responsible for a wide alternative splicing dysregulation in prostate cancer. Moreover, HNRNPK was reported to be associated with cancer development and proliferation in different cancer subtypes (W. Zhou et al. 2023). We therefore conducted a HNRNPK transient silencing in PC3 cell line (see Methods). A total of 4,641 exons were defined as alternatives and 3,601 as constitutive (Figure 30A). These events were used as input for RNAmars (see Methods).

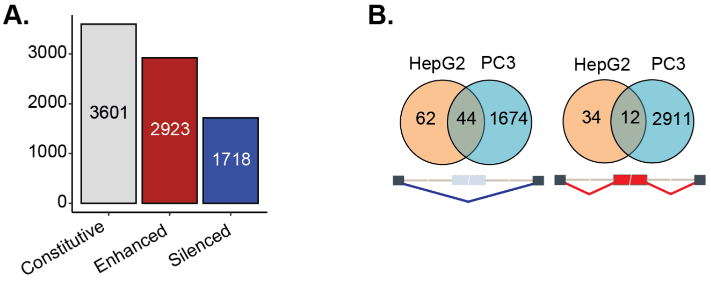

Figure 30: Alternative exons from HNRNPK silencing in PC3 cell line. A. Number of constitutive, enhanced and silenced exons used as input for RNAmars. B. Intersection between HNRNPK regulated exons in PC3 cell line from our experiment and the HNRNPK regulated exons in HepG2 cell line from ENCODE data

Since HNRNPK binding profile was not available for K562 because of the limited number of exons (38 exons) (thus not passing filters (see Methods)), I used the HepG2 eCLIP data from ENCODE in RNAmars pipeline. Only a limited subset of exons was common to HepG2 HNRNPK regulated exons in ENCODE data (only 46 enhanced and 44 silenced are shared with ENCODE dataset, among a total of 152 and 4,641 alternative exons in HepG2 and PC3 respectively), (Figure 30B). The high rate of unseen exons in PC3 cell lines could allow me to test the robustness of RNAmars across different exons of the same regulatory protein.

RNAmars was run and the two resulting heatmaps were retrieved, for silenced and enhanced events respectively. The algorithm accurately associated both silenced and enhanced exons to HNRNPK, which was found at the first place in the RBP ranking (Figure 31A-B). For silenced exons, the top tetramer enrichments were retrieved mostly in the exon region and the upstream region, as expected from the BS (Figure 31B).

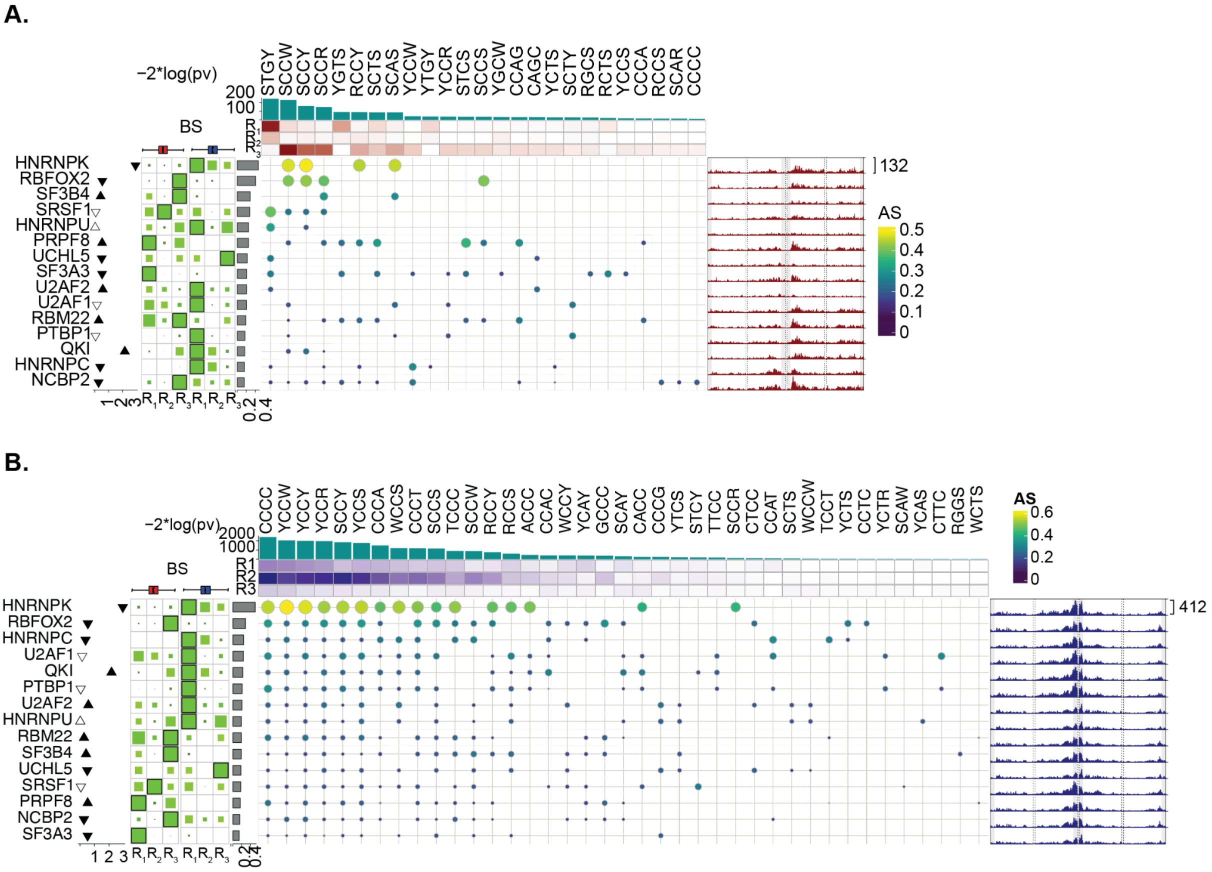

Figure 31: RNAmars results on HNRNPK regulated exons in PC3 cell lines. Annotations are as previously explained in Figure 27. Down-facing triangles indicate down-regulation upon HNRNPK silencing. Up-facing triangles indicate up-regulation upon HNRNPK silencing. **A.** Enhanced exons RNAmars result. **B.** Silenced exons RNAmars result.

Considering only the tetramers enriched in the run with HNRNPK optimal parameters (scrolling window of 35 and enrichment window of 300), in HepG2 and PC3 cell data, they were almost the same for the silenced set of exons (Figure 32A). Tetramers enriched in the enhanced events analysis, instead, were more heterogeneous, nevertheless presenting a C in the second position. Most importantly, the majority enriched tetramers in the silenced exons resembled the canonical HNRNPK motif which is C-rich and flanked by A or T (Figure 32A) (D. Zhou et al. 2021; Feng et al. 2019), in particular, 15 out of 16 tetramers had CC nucleotides in the center of the tetramer.

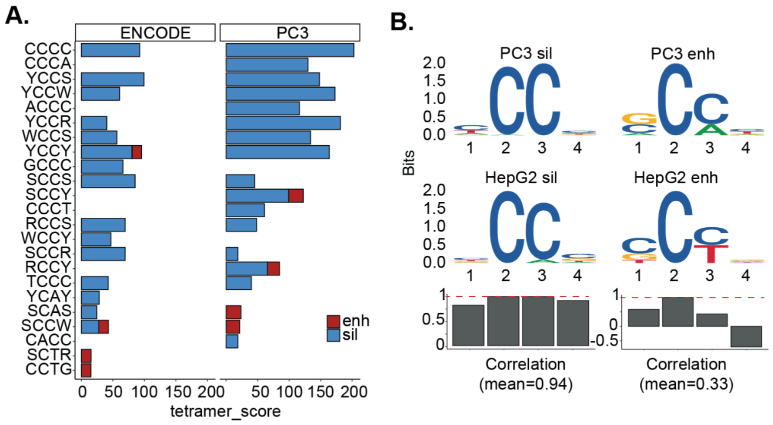

Figure 32: Enriched tetramers of HNRNPK regulated exons in PC3 and HepG2 cell line: **A.** Enrichment scores of tetramers in the two cell lines, separated by regulation type. **B.** Same tetramers collapsed into a single PWM and compared between cell lines using a Pearson's correlation coefficient.

Tetramers were then collapsed into a PWM and compared using Pearson Correlation Coefficient (PCC) (see Methods), which is one of the methods traditionally used to quantify similarity between PWMs (Gupta et al. 2007). Highest correlation was found in the silenced set with a mean correlation of 0.94, while for enhanced was 0.33 (Figure 32B), which is expected since only 1 out of 4 tetramer (SCCW) in enhanced exons is in common between the two cell lines. Considering that HNRNPs proteins have a predominant effect on exons silencing rather than enhancing (Van Nostrand, Freese, et al. 2020), it was expected to find a stronger signal in silenced exons and therefore higher compatibility between the two datasets.

Then, I wondered whether the enriched tetramers were located in the same regions of the HepG2 enriched tetramers. In other words, I was interested in assessing the similarity between the tetramer splicing maps deriving from the two datasets. First, I collected the tetramer splicing maps combining all enriched tetramers from RNAmotifs runs with HNRNPK optimal parameters for both PC3 and HepG2. I used as a benchmark splicing map the binding profile of HNRNPK in HepG2. As expected from the binding profile, tetramer splicing maps of both HepG2 and PC3 presented a high enrichment within the exon in silenced events and in the flanking introns of enhanced exons (Figure 33).

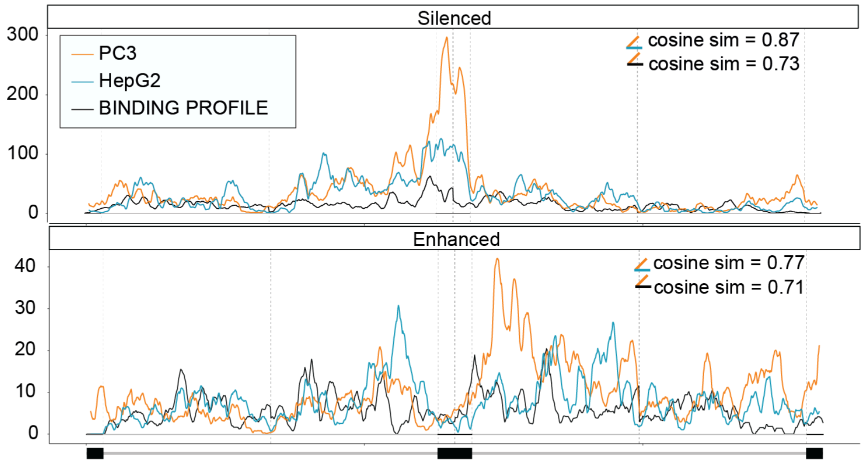

Figure 33: Splicing maps of RNAmotifs tetramers enrichment scores of HNRNPK regulated exons in PC3 (orange) and HepG2 (cyan), and binding profile of eCLIP peaks of HNRNPK in HepG2. Pairwise cosine similarities of the profiles are annotated in the top right corner.

I evaluated the similarity between the splicing maps using the cosine similarity. The agreement between PC3 and HepG2 tetramer splicing maps was particularly high (cosine=0.87) in silenced exons (Figure 33). The PC3 tetramer splicing map resembled the expected binding profile with a cosine similarity of 0.73 for silenced and 0.71 for enhanced. I also pondered on how comparable the ASs related to HNRNPK in the two cell lines were. Therefore, I computed PCC on the final AS associated with HNRNPK on the two datasets Figure 34. On enhanced exons, it was not possible to compute PCC because only one tetramer was in common between the two datasets, while for the silenced exons the correlation was high (r=0.72, pv=0.018). Given the high similarities in terms of sequences, binding and AS between HepG2 and PC3 datasets, despite the fact that the input exons were very different, RNAmars results were promising. In fact, RNAmars was able to predict the actual regulator starting from different cell type and input sets, showing high robustness and reproducibility.

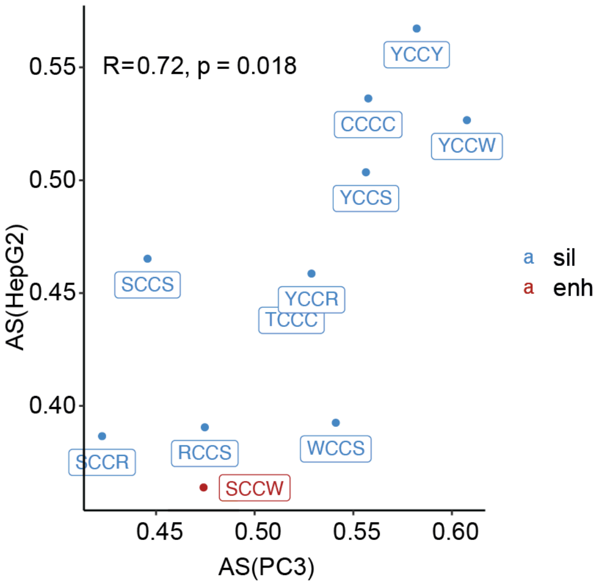

Figure 34: Scatter plot of association scores related to enriched tetramers that were associated to HNRNPK in HepG2 and PC3 cell lines. Color code defines regulation type.

RBFOX2 and U2AF2 regulate alternative splicing in sarcomas

Having confirmed the trustworthiness of RNAmars, it was now suitable to employ it within the Sarcoma dataset.

ASEs were defined as previously described (See Chapter 1). Only the top 2000 constitutive exons ordered by decreasing FDR were kept as controls to avoid imbalance between the two classes (Figure 35).

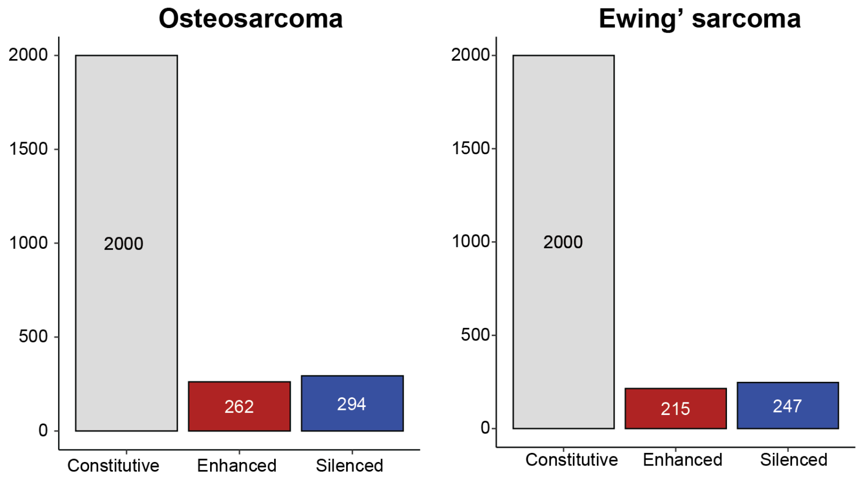

Figure 35: Number of constitutive and alternative exon skipping events in sarcoma data given as inputs to RNAmars algorithm.

These sets of exons were used as input to RNAmars algorithm. Enhanced crosslinking and immunoprecipitation data (eCLIP) from HepG2 cell line was used to compute the Cross-linking Enrichment Score (CES) of all the RBPs to be compared with RNAmotifs Enrichment Score (ES) and retrieve the profile similarity as explained in the Methods section.

EW enhanced exons presented RBFOX2 as top protein which was associated with GC tetramers (Figure 36). In particular, GC tetramers enrichments were mainly within the downstream region R_3, where the BS is maximum, *i.e.* where RBFOX2 is expected to bind. Moreover GC motifs resemble the canonical consensus motif (U)GCAUG of RBFOX2 (Ponthier et al. 2006; Lambert et al. 2014).

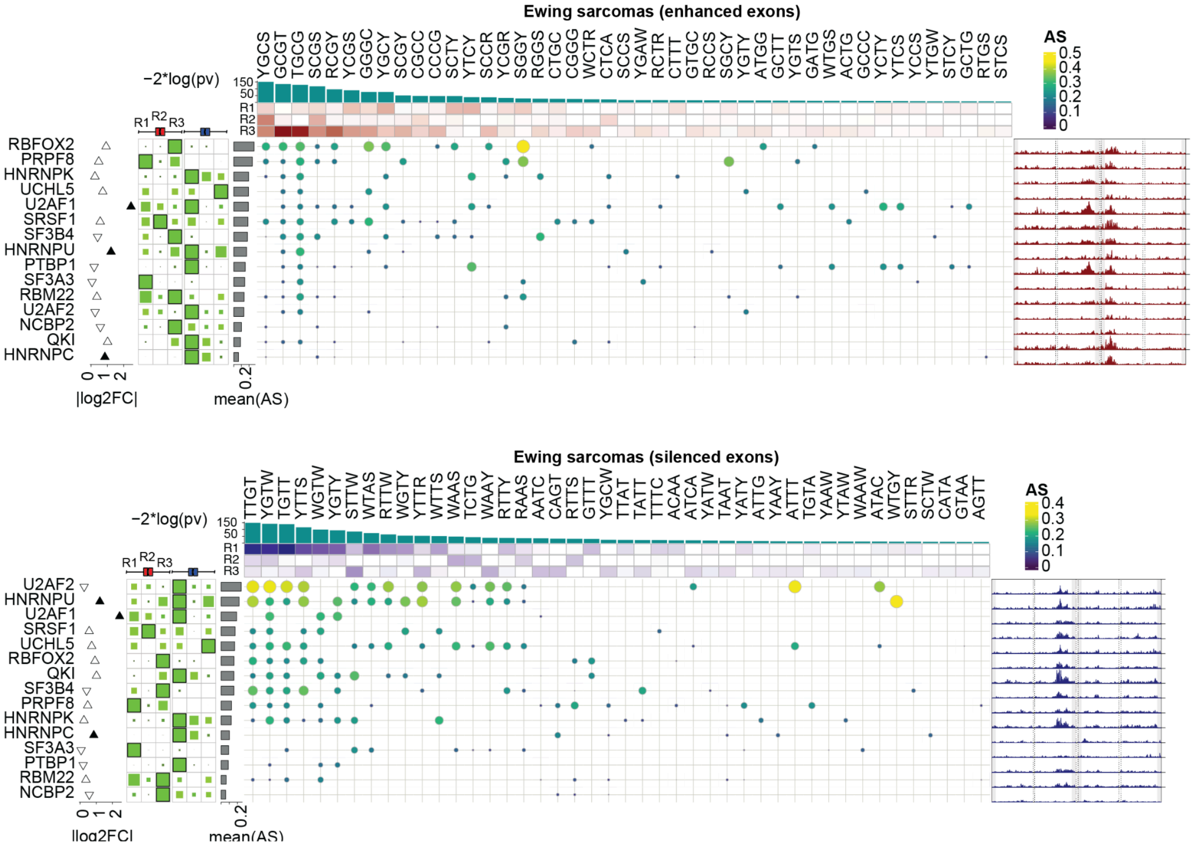

Figure 36: RNAmars results on alternative exons skipping events in EW for enhanced (top heatmap) and silenced (bottom heatmap) exons. Annotations are as previously explained in Figure 27. Log2 Fold Change is positive (triangle pointing up) when the gene is higher expressed in EW and negative (triangle pointing down) when the gene is higher expressed in OB.

In EW silenced exons the top ranked protein was U2AF2 which was highly associated with TT-rich tetramers in the upstream region R_1. This result was in accordance with the expected binding of U2AF2 which exhibits maximum BS in R_1. Moreover, U2AF2 mCROSS PWM presents high information content in two consecutive thymines, validating RNAmars results on the sequence enrichment (Figure 37).

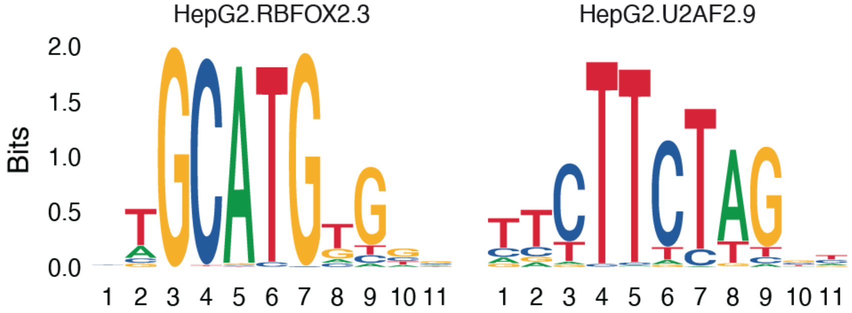

Figure 37: PWMs for RBFOX2 and U2AF2 in HepG2 cell line from mCROSS database.

Also for OS RNAmars predicts RBFOX2 as the top enhancer of alternative exons (Figure 38). However, the multivalent motifs enrichments are less region specific with respect to EW, suggesting the possibility of multiple regulations involving different proteins with different binding preferences. Interestingly, the final protein score, measured as the mean of the association scores (see Methods), for both RBFOX2 and the second most associated protein, HNRNPK, are remarkably similar, with RBFOX2 scoring 0.216 and HNRNPK scoring 0.212. Given the closeness in their rankings, it is advisable to consider both RBFOX2 and HNRNPK as potentially influential factors. Additionally, it is worth noting that previous research has indicated an indirect influence of HNRNPK on RBFOX2 binding (D. Zhou et al. 2021), further underscoring the interplay between these proteins in the context of alternative exon regulation. Enriched motifs in this case are more heterogeneous with respect to EW, presenting enrichments in all three regulatory regions and with a prevalence of CG motifs and are not perfectly resembling the mCROSS PWMs (Figure 37). However, it is a well-established fact that when there is cooperative interaction between RBPs the binding motifs can change considerably (Lang et al. 2021).

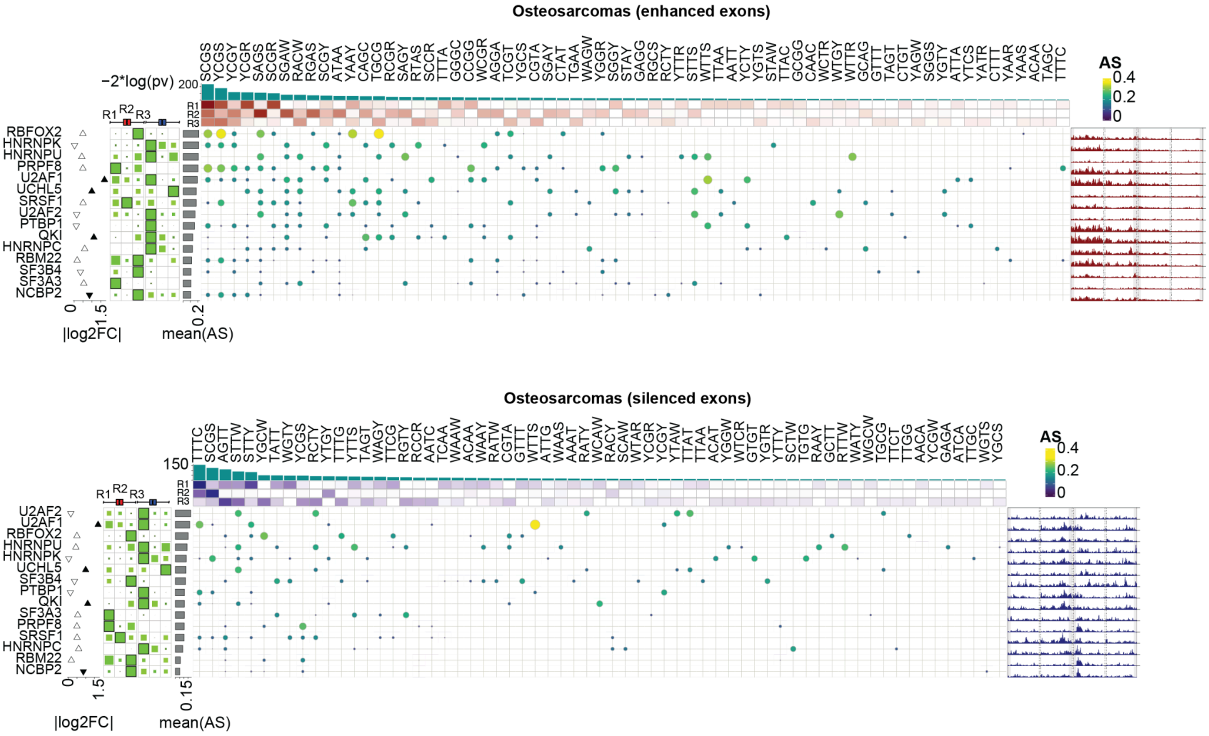

Figure 38: RNAmars results on alternative exons skipping events in OS for enhanced (top heatmap) and silenced (bottom heatmap) exons. Annotations are as previously explained in Figure 27. Log2 Fold Change is positive (triangle pointing up) when the gene is higher expressed in OS and negative (triangle pointing down) when the gene is higher expressed in OB.

Silenced exons were predicted to be regulated by U2AF2 by its association to TT rich sequences. However the most enriched tetramer TTTC is not associated to U2AF2, but to the second top ranked protein U2AF1, which suggests there might be its involvement as well. This is further confirmed by the fact that the highest association score (AS=0.33) happened between WTTS tetramer and U2AF1. This fits with the fact that the two U2AFs cooperate together and share most of the regulated exons (Shao et al. 2014). Moreover, the TT rich motifs are in line with the mCROSS PWM which shows a high preference for consecutive Thymines for both U2AF1 and U2AF2 (Figure 37 and Figure 39).

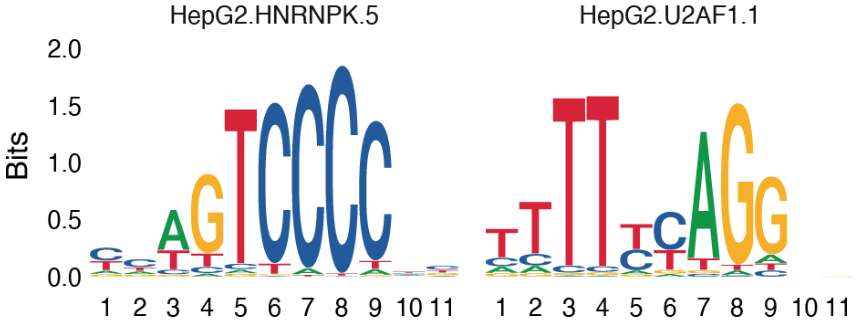

Figure 39: PWMs for HNRNPK and U2AF1 in HepG2 cell line from mCROSS database.

In summary, RBFOX2 and U2AF2 were the top exon enhancer and silencer respectively for both diseases. Furthermore, HNRNPK and U2AF1 appeared to be moderately engaged in splicing regulation in OS. However, it is noteworthy that among the predicted regulators only U2AF1 exhibited differential expression, as indicated by the "log2FC" panel. I wondered whether there might be another source of dysregulation within these proteins, therefore I examined the alternative events occurring within these genes. RBFOX2 displayed three distinct ASEs in EW, specifically A3SS, ES and MXE. OS had only one ASE which coincided with the MXE event observed in EW (Table 2). Intriguingly, this particular event was also overlapping with two RNA Recognition Motifs Domains (RRM_1 and RRM_5) from the PFAM database (Figure 40 and Table 2).

Alternative	Flanking	Event	PFAM	DPSI(EW)	DPSI (OS)
chr22:35756104-35756144	chr22:35752612-35752655 chr22:35759887-35760020	ES	Fox-1_C	-0.121	-
chr22:35781599-35781746 chr22:35781599-35781743	chr22:35809779-35810004	A3SS	-	0.131	-
chr22:35768256-35768349 chr22:35778024-35778078	chr22:35765422-35765483 chr22:35781599-35781743	MXE	RRM_1, RRM_5	0.205	0.152

Table 2: Alternative splicing events occurring within RBFOX2 gene.

This particular MXE event involves two different isoforms, one including exon five and skipping of exon six (isoform 1 in Figure 40) and another including exon 6 and excluding exon five (isoform 2 in Figure 40).

In particular, sarcomas promote higher inclusion isoform 1 and controls isoform 2. It was previously shown that the skipping of exon 6 deletes a portion of the RRM, reducing the binding capability of the protein. Moreover, the isoform 1 and the full length isoform are also thought to compete for the binding to enhance exon inclusion (Damianov and Black 2010). Specifically, isoform 1 has a weaker binding affinity with RNA than the full length isoform including both exon five and exon 6. However, it is yet unknown how isoform 1 interacts with isoform 2 and which one has stronger impact on binding and splicing activity.

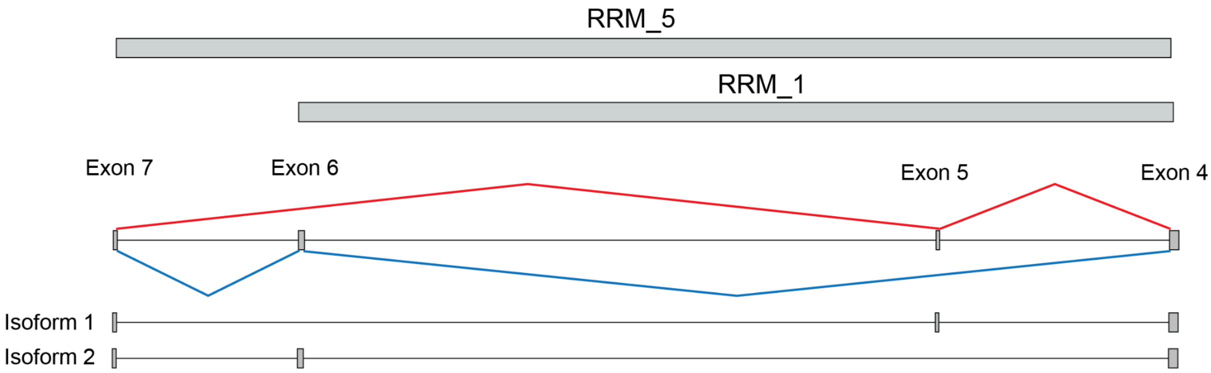

Figure 40: RBFOX2 mutually exclusive exon skipping events. Isoform 1 is the most included form in sarcomas. Above the regions where PFAM domains are found.

Given that sarcomas exhibit higher expression of isoform 1 and RNAmars has detected an increase in RBFOX2-enhanced exons, it is reasonable to presume that skipping exon 6 has a lesser impact on exon inclusion compared to the skipping of exon 5. Expression of RBFOX2 isoform 1 has therefore an overall enhancing of regulated exons.

U2AF1 effects on OS exon silencing can be easily explained by its upregulation in sarcomas. Specifically, the expression of U2AF1 is usually promoting exon exclusion by binding in the 3' splice site (R_1 region), which is coherent with RNAmars results.

HNRNPK presented two significant exon skipping events in OS, however they both impacted the 5'UTR, therefore without affecting the coding sequence and the protein domains. Moreover, neither differential expression nor significant differential splicing in U2AF2 were not found, suggesting that the deficiencies in these proteins rely on other sources, not directly measurable with standard RNA-seq. As an example, RNA modifications can influence the alternative splicing regulation. In particular, N^6-Methyladenosine (m6A) modification disrupts splicing activity by modifying the affinity between RBPs and RNA (Wei et al. 2021; Louloupi et al. 2018). To gain insight in this idea I checked differential expression analysis of a literature curated list of m6A related genes. m6A methylation is a reversible process: there exist some proteins that deposit m6a into RNA (writers), some others that read it (readers) and lastly some that erase it restoring the original nucleotide (erasers) (Jiang et al. 2021). Only the reader RBM15 was found differentially expressed in OS (DESeq2 p-value adjusted = 6.71e-09 & Wilcoxon p-value = 0.018), suggesting that there might be a differential m6A load between cases and controls (Figure 41). Since RBM15 was up-regulated, the m6A load is expected to be higher in sarcomas. However, it is challenging to formulate additional hypotheses in the absence of a comprehensive m6A experiment.

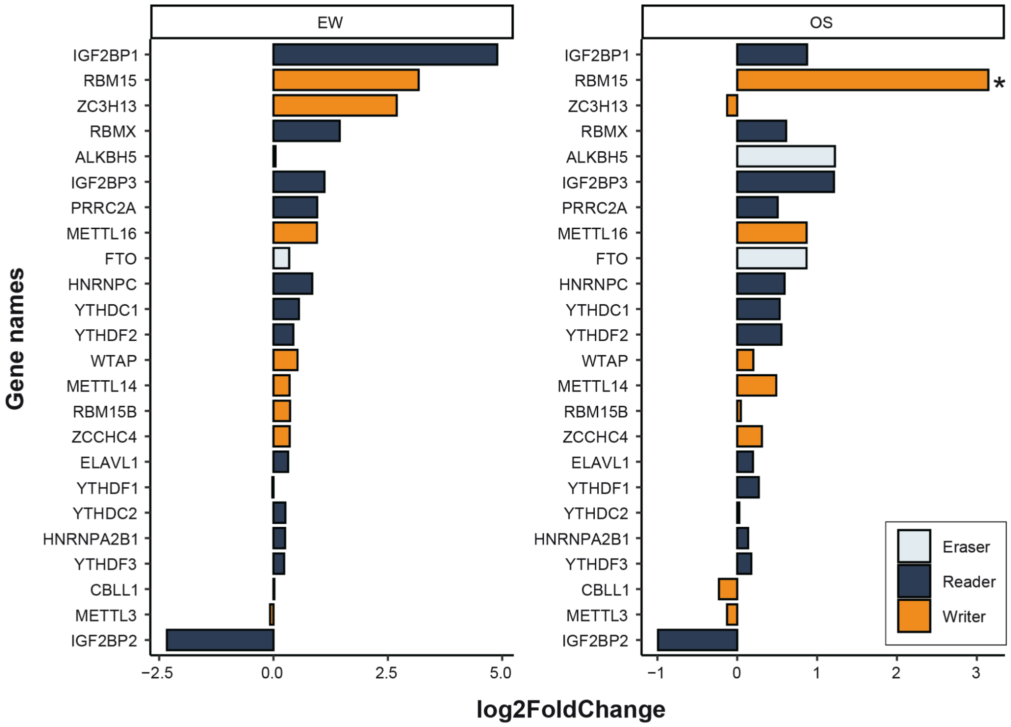

Figure 41: log2 Fold Change of m6A related genes expression, colored by their functions and divided by the two sarcomas subtypes. Star means that the gene is differentially expressed as defined in the previous section.

Methods

RNAmars software employs the RNAmotifs algorithm (Cereda et al. 2014) and it is implemented in the R programming language.

RNAmars algorithm *per se* relies on *a priori* knowledge, in particular on the RBPs binding preferences around their regulated exons. Therefore, preliminary steps are needed to retrieve the binding features for each RBP, in order to train the ultimate algorithm. The workflow can be divided into four major steps, explained in detail in the next paragraphs following Supplementary Figure 1. Briefly, the steps are organized as follows: (i) data retrieval: which types of data have been used and which filtering conditions have been applied to them; (ii) data preprocessing: how to identify the binding preferences of the proteins and evaluate the strength of the signal; (iii) tuning of RNAmotifs parameters: grid search and optimization strategy to identify the optimal parameters for each dataset; (iv) RNAmars algorithm: starting from unseen user defined alternative exons, it describes the strategy to retrieve the putative regulatory protein of AS events.

Selection of RBPs and regulated exons

RBPs cross-linking sites, as iCounts peak instances, from eCLIP experiments in HepG2 and K562 cell lines were collected (Curk, 2019). Concomitantly, differential splicing data upon RBPs depletion (*i.e.* shRNA or CRISPR) deriving from rMATS software (Shen et al. 2014) in the same cell lines were retrieved from ENCODE (Van Nostrand, Freese, et al. 2020) (accession number ENCSR413YAF, Supplementary Table 3). To select RBPs that were most likely to bind the multivalent RNA motifs, a list of 64 and 80 11-nt long position weight matrices (PWMs) eCLIP data derived from HepG2 and K562 respectively were collected from the mCross database (Feng et al., 2019). For each RBP, the PWM with lowest (i) Allelic Interaction p-values and (ii) highest consistency were selected. Only RBPs with an associated PWM were retained for further analyses.

Only cassette exons (CEs) were considered in the analysis, since they are the most frequent alternative splicing events in human cancers (Yangjun Zhang et al. 2021).

The percent spliced in (PSI) value was used as a measure of exon inclusion in the mature mRNA (J. P. Venables et al. 2009).

CEs with (*i*) significant inclusion changes (*i.e.* |DPSI|>0.1 and FDR<0.1) upon RBP depletion, (*ii*) annotated as “cassetteExon” in the UCSC hg19 “knownAlt” table (Navarro

Gonzalez et al. 2021), and (iii) with at least one iCounts peak instance within 300 and 30 nucleotides (nts) into introns and exons, respectively, from the splice sites (if introns and exon were shorter than 600 and 61 nts, respectively, then the whole introns and exon were evaluated) were considered as alternatively spliced, or RBP-regulated exons.

Sizes of putative regulatory regions were defined following previously published indications to best capture the motifs and binding preferences of the RBPs (Barash et al. 2010; Van Nostrand, Freese, et al. 2020). Cassette exons with DPSI>0.1 or DPSI<-0.1 were defined respectively as enhanced and silenced exons (Figure 42).

Constitutive exons were defined as previously proposed ((Attig et al. 2018). Briefly, CEs with (*i*) no significant inclusion changes (*i.e.* $|\Delta\mu(\mathrm{PSI})|<0.01$ and FDR>0.1) upon RBP depletion and (*ii*) annotated in the RefSeq database version 210 (Frankish et al. 2019) but not in the UCSC hg19 “knownAlt” table (Navarro Gonzalez et al. 2021) were considered as constitutively spliced and used as controls.

Since at least 50 exons are required to disentangle RBPs regulatory patterns (Yee et al. 2019), only datasets with more than 50 Alternative Splicing Events (ASEs) were retained for the downstream analyses (Figure 42).

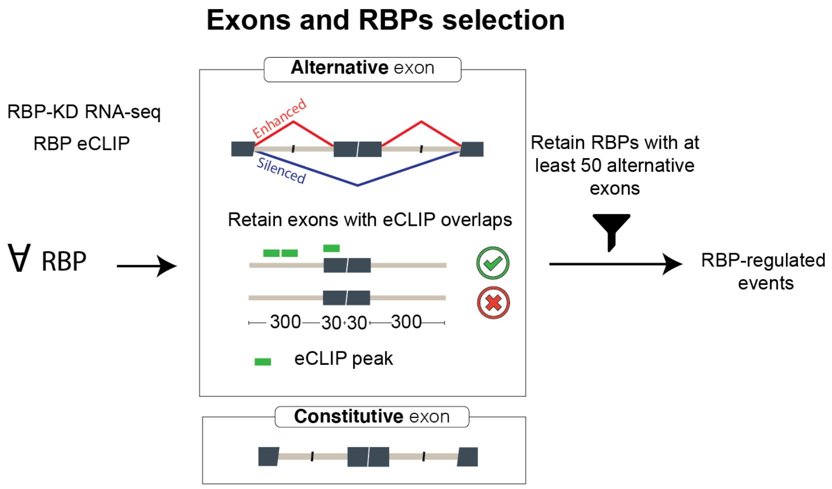

Figure 42: Selection of exons diagram.

RNA splicing maps of crosslinking sites

To assess the binding preferences of each RBP, an enrichment profile of iCounts peaks, hereafter called “splicing map” (Cereda et al. 2014), around the ASEs with respect to the constitutive exons was generated as previously proposed with minor modifications (Del Giudice et al. 2022). First, the regulatory regions for the splicing map were defined considering 300 and 30 nts into introns and exons, respectively (upstream, alternative and downstream exons) as previously reported (Van Nostrand, Freese, et al. 2020), see Figure 43. Since RBPs generally bind at different positions to promote or prevent exon inclusion (Cereda et al. 2014), for each RBP two splicing maps were built separately, for silenced and for enhanced exons respectively, both against constitutive exons.

At each position along the map, the cross-linking enrichment score (CES) was computed comparing the proportion of enhanced, or silenced, and constitutive CEs having at least one iCounts peak of protein *p* using a one-tailed Fisher's exact test:

$$CES(p,a)_i = \; -\,2 \; \cdot \, log(pFis(p,a)_i)$$

where $pFis_i$ is the Fisher's exact test p-value at i-th position along the map, *p* refers to the protein and *a* refers to either enhanced or silenced exons.

The collection of all the CESs along the splicing map defines the binding profile $B(p,a)$ of protein *p* for regulation type *a* (silenced or enhanced).

To account for the ability of RBPs to bind multiple proximal (*i.e.* multivalent) motifs (Cereda et al. 2014), all binding profiles were averaged in a scrolling window of 15 nts using the *filter* function of "stats" R package v4.2.3 and visualized as RNA splicing map.

To evaluate the value of RNA maps in predicting exon inclusion by the corresponding RBP, a bootstrap procedure with downsampling was implemented. For 500 iterations, a percentage *j* of randomly selected CEs ranging from 10% to 100% with 10% increasing step (*j* = {10%, 20%, …, 100%}) was selected. For each subset of exons, a binding profile of protein p (*i.e.* $B(p,a)_j$) was calculated as described above and compared with the observed one (*i.e.* $B(p,a)$) using the cosine similarity as follows:

$$cos(B(p,a)_j,\, B(p,a)_{100\%}) \;=\; \frac{B(p,a)_j \cdot B(p,a)_{100\%}}{||\, B(p,a)_j|| \;||\, B(p,a)_{100\%}||}$$

For each percentage of exons, cosine similarities were averaged. Area under the curve (AUC) was calculated and used to interpret the ability of the eCLIP-based RNA splicing map (*i.e.* the binding profile) to predict RBP-mediated splicing regulation (Figure 43).

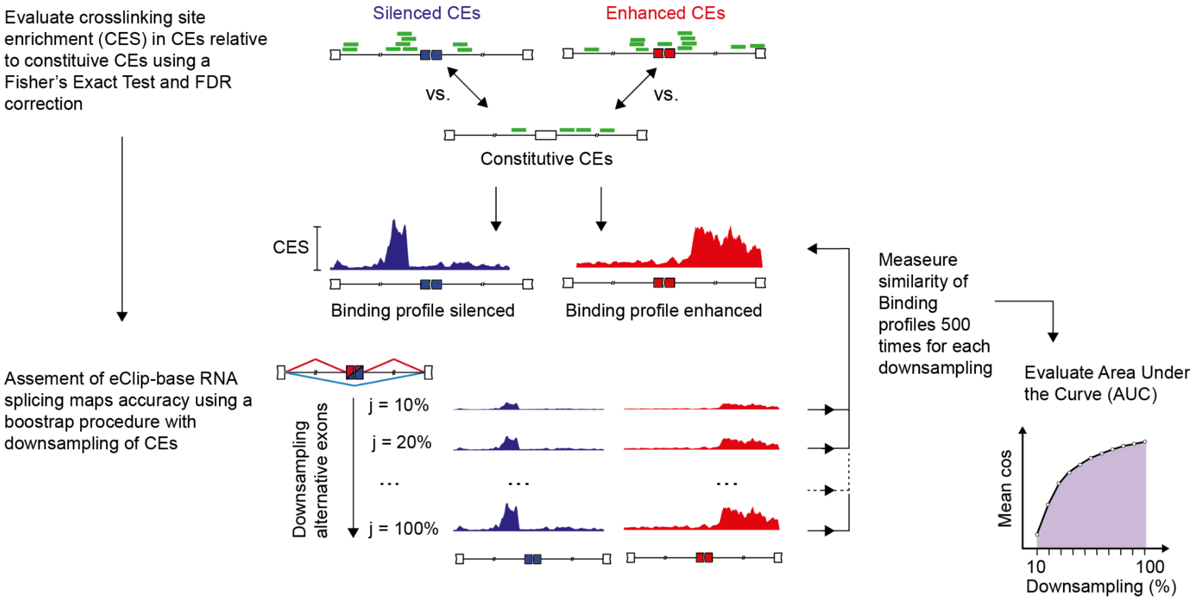
Evaluate crosslinking site enrichment (CES) in CEs relative to constituive CEs using a Fisher's Exact Test and FDR correction
Silenced CEs
Enhanced CEs
vs.
vs.
Constitutive CEs
CES
Binding profile silenced
Binding profile enhanced
Measeure similarity of Binding profiles 500 times for each downsampling
Evaluate Area Under the Curve (AUC)
Assement of eClip-base RNA splicing maps accuracy using a boostrap procedure with downsampling of CEs
Downsampling alternative exons
j = 10%
j = 20%
...
j = 100%
Mean cos
10
100
Downsampling (%)

Figure 43: Diagram showing the procedure to retrieve the signal recovery rate. Firstly, binding profiles are built from silenced and enhanced exons pairwise compared to constitutive exons. Then the binding profile derived from subsets of the original exons are compared with the complete profile with a cosine similarity. When examining the mean cosine similarity on the y-axis against the downsampling percentage on the x-axis, the AUC serves as a metric to quantify the rate of signal recovery.

Quantification of RBP binding activity at regulated exons

Impact of RBP regulation on exon inclusion was evaluated at three canonical regions around the splice sites (ss) of CEs as previously proposed (Cereda et al. 2014). These regions were defined as follows: R1 [-300;0] nts of intronic sequence upstream of the 3′ss; R2 of exonic sequence [1;30] nts downstream of the 3′ss and [-30;-1] nts upstream of the 5′ss; R3 [1;300] nucleotides of intronic sequence downstream of the 5′ss. If introns and exon were shorter than 600 and 61 nts, respectively, then the whole introns and exon were evaluated. A bootstrap procedure was implemented to define an empirical background distribution of CESs at each position along the RNA splicing map (*i.e.* CES_{emp}). For each RBP and each iteration, exon labels (*i.e.* enhanced/silenced and constitutive) were randomly shuffled and the CES was calculated for every position of the map. Positions at regions R1, R2, and R3 with observed CESs greater than the 95th percentile of the corresponding CES_{emp} were considered as significantly bound by the RBP of interest. To define the propensity of RBPs to bind these positions, the difference between the observed and empirical CES (*i.e.* *ΔCES)* was measured:

$$\Delta CES(p,a)_i = CES(p,a)_i - CES(p,a)_{emp,i}$$

where *i* is the i-th position in each region, *p* corresponds to the protein for which iCounts peaks are being examined, and *a* denotes the type of regulation. Finally, to quantify RBP preferential binding around splice sites, the maximum ΔCES_i was evaluated for each region and referred to as “binding strength” (BS):

$$BS(p,a)_r = max\{\Delta CES(a)_i\}$$

where *r* is the regulatory region (*i.e.* R1, R2, R3), and *i* is the i-th position in the region. For each protein, the binding strength is a six dimensional vector collecting BSs within each of the three regulatory regions in both the regulation types *a.*

To allow comparisons across different RBPs, the BSs were normalized to the maximum one.

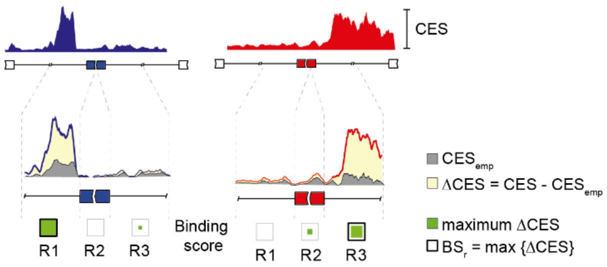

Figure 44: Diagram showing the idea behind the BS and the region of interest

Identification of multivalent RNA motifs

Since different RBPs can have different binding properties, tuning RNAmotifs enrichment window *e* and scrolling window *n* parameters to best suit each considered protein was needed. In the original RNAmotifs algorithm, the parameters were set to n=15 and e=30 as they well described the Nova1 and Nova2 binding preferences (C. Zhang and Darnell 2011). The best *n* and *e* windows of each RBP were defined through a grid search considering *n*={5, 15, 25, 35} and *e*={30, 50, 100, 200, 300} for a total of 20 combinations (Supplementary Figure 1C). For each set of RBP-regulated exons, in each cell line, and for each combination of parameters, RNAmotifs was run with 10,000 bootstrap iterations for empirical p-value (Cereda et al. 2014). In order to widen the splicing map, RNAmotifs script was modified to consider intronic regions of 300 nts and exonic regions of 30 nts. Each RNAmotifs result consists of 512 tetramers, each of them having six p-values, three related to R_1, R_2 and R_3 in enhanced exons and three for the same regulatory regions for silenced exons. Only tetramers with one-tailed Fisher's Exact Test p-values ≤ 0.05 (or ≤ the 1st percentile of the p-value distribution was used as threshold when the first percentile was lower than 0.05) and empirical p-values ≤ 0.0005 were considered as significantly enriched and retained for further analysis.

Non-significant region-specific enrichment p-values were set equal to 1.

A tetramer was denoted as t_e or t_s if it was enriched in enhanced exons or silenced exons respectively. In case it was significant in both enhanced and silenced, it got double annotation separately.

Given all tetramers, region-specific enrichment p-values were combined into one goodness-of-fit (χ^2) statistics for each region *R* producing a six dimensional vector representing the global tetramers enrichment in the three regulatory regions R_i in both enhanced and silenced exons. This vector was called tetramer strength (TS).

$$TS_i = -2\sum_t log(pFis_{t|i})$$

where t is an enriched tetramer, $p_{t|i}$ is the Fisher's p-value of tetramer t in region $i \in \{R_{1,s}, R_{2,s}, R_{3,s}, R_{1,e}, R_{2,e}, R_{3,e}\}$ (where the subscript *e* refers to the region of enhanced exons and *s* to silenced exons).

The overall tetramer score was defined as the Fisher method's aggregation of the TS, representing the overall enrichment score of the input dataset.

The cumulative tetramer score (CTS) was defined as the sum of tetramer scores across the six regions and considered as a unique value for the tetramer.

$$CTS(t) = -2\sum_i log(pFis_{t|i})$$

where *i* is the regulatory region ($i \in \{R_{1,s}, R_{2,s}, R_{3,s}, R_{1,e}, R_{2,e}, R_{3,e}\}$), t is the enriched tetramer, $p_{t|i}$ is the Fisher's p-value of tetramer t in region *i*.

RNAmotifs enrichment scores (ES) representing the preferential location of tetramers around alternative exons were converted into two distinct scores, representing RNA splicing maps for enhanced and silenced exons separately (Figure 45A), as follows:

$$ES(t_a)_i = -2\sum_{t\in r} log(pFis_i(t_a))$$

where t is the tetramer of interest, *i* is the position along the map, *a* refers to either enhanced or silenced CEs. Denote the collection of the $ES(t_a)_i$ along all the position *i* in the map as $M(t_a)$.

The tuning object function had to maximize the compatibility between the splicing maps of tetramer t_a, $M(t_a)$, where $a \in \{s, e\}$, and the binding profile of protein *p* restricted to exons for which t_a is found enriched, $B(p,\ t_a)$. The more the two splicing maps are similar the more probable it is that the protein *p* is responsible for AS of the input exons.
The function used to measure the similarity between tetramer t_a and protein *p* was derived separately for each set of parameters as follows. For each enriched tetramer t_a, in the region R_i of the splicing map, where *i*={1, 2, 3}, the subset of the input exons containing at least one occurrence in R_i was retrieved. In case the tetramer was enriched in more than one region, the union of exons was considered. Starting from this set of exons, the binding profile was created as described above, resulting in the binding profile of protein *p* restricted to exons containing tetramer t_a, $B(p, t_a)$.
For each tetramer t_a and protein *p*, the tetramer splicing map and the restricted binding profile were then compared using a cosine similarity (Figure 45B), resulting in a *profile similarity score* $\beta(p, t_a)$:

$$\beta(p, t_a) \quad = cosine(M(t_a),\ B(p,\ t_a)) = \frac{M(t_a) \cdot B(p,t_a)}{||\, M(t_a) \,||\ ||\, B(p,t_a) \,||} = \frac{\sum_i M(t_a)_i B(p, t_a)_i}{\sqrt{\sum_i M(t_a)_i^2}\sqrt{\sum_i B(p, t_a)_i^2}}$$

where *i* represents the position along the map and *a* indicates the regulation type (*i.e.* silenced or enhanced). The cosine similarity was chosen as it is invariant to amplification of the signal ("Learning Similarity with Cosine Similarity Ensemble" 2015), *i.e.* it favors the shape of two vectors rather than their norm.

The profile similarity score $\beta(p, t_a)$ was then weighted by the signal recovery rate $\alpha(p)_a$ to penalize proteins for which subsets of exons did not give a robust signal. The final score was called association score $AS(p, t_a)$ and calculated as follows (Figure 45C):

$$AS(\,p, t_a) \ = AS_a = \alpha(p)_a \ \cdot \beta(p, t_a)$$

$$a \ \in \{s, e\}$$

The association score describes the strength of the association between tetramer t_a and protein *p.* In particular, the higher the score, the higher the probability that the protein is regulating the input set of exons by binding to that motif.
This procedure was carried out for all the combinations of parameters, yielding to 40 different AS matrices (20 for silenced and 20 for enhanced exons) for each RNAmotifs run (Cereda et al. 2014) and for each set of RBP-regulated exons in the two cell lines, separately.

To evaluate the optimal set of parameters for each dataset (RBP-regulated exons in one of the two cell lines), the sensitivity and specificity in associating the enriched tetramer to the actual regulatory protein were calculated using a Receiver Operating Characteristic curve strategy (ROC). In particular, for each dataset of input exons controlled by RBP_{input}, and for each RNAmotifs run *r (Cereda et al. 2014)*, two matrices of association scores, one for the enhanced exons, AS_e, and one for the silenced exons, AS_s, were produced. For each cell line, the matrix showed on the columns the enriched tetramers found by RNAmotifs in RBP_{input} regulated exons (Cereda et al. 2014), and on the rows the available proteins for that cell line. In the dataset of exons regulated by RBP_{input}, ASs relative to that protein are expected to be high. ASs are normalized to range from zero to one for each association score matrix in run *r*, defining the normalized AS (ASn):

$$ASn(p, t_a)_r = \left[\frac{AS(p,t_a) - min(AS_a)}{max(AS_a) - min(AS_a)}\right]_r$$

Given a threshold $L \in [0, 1]$ the protein *p* is defined as associated to tetramer *t* if $ASn(p, t_a)_r$ > L, ranging between 0 and 1 with 0.01 step. As this procedure can be viewed as a classification algorithm, performance metrics were defined. Given the protein RBP_{input}, consider the AS matrix resulting from its own regulated exon (either silenced or enhanced). Metrics can be defined as:

- Positives: length of $ASn(p, t_a)$, where p = RBP_{input}, the number of enriched tetramers (*i.e.* number of columns of the AS matrix).
- Negatives: length of $ASn(p, t_a)$, where p ≠ RBP_{input}, *i.e.* the number of associations where *p* ≠ RBP_{input}.
- True Positives (TP): number of tetramers associated with P_A with a score greater or equal to *L* ($ASn(p, t_a) \geq L$ *with p =* RBP_{input}).

- False Positives (FP): number of tetramers associated with the other proteins with a score greater or equal to *L* $(ASn(p, t_a)$≥ *L with p ≠ RBP$_{input}$*)

Since these quantities were computed separately for the two cell lines, the overall accuracy metrics were obtained by averaging the two metrics from the cell lines.

The ROC curve was then built starting from the metrics, making L ranging from 0 to 1 with a step of 0.01 (Figure 45D). This process was performed for each RNAmotifs (Cereda et al. 2014) run and separately for each dataset of RBP-regulated exons.

The set of parameters that maximized the Area Under the ROC curve (AUROC) was selected as the optimal one. Since two ROC curves for each RBP were evaluated (already aggregated across cell lines), for the enhanced and silenced exons respectively, the regulation type for which the region of interest was found was considered (*i.e.* where the maximum BS is located). In other words, the accuracy metrics and respective AUROC curves were taken from the regulation type where the maximum BS is found. In cases in which the two cell lines had different regions of interest, the optimal set of parameters was defined as the one that presented the maximum AUROC across the two regulation types. Moreover, when AUROC was less than 0.5, meaning that the prediction performed worse than a random classifier, the default RNAmotifs parameters were considered as optimal parameters (n=15, e=30) (Cereda et al. 2014).

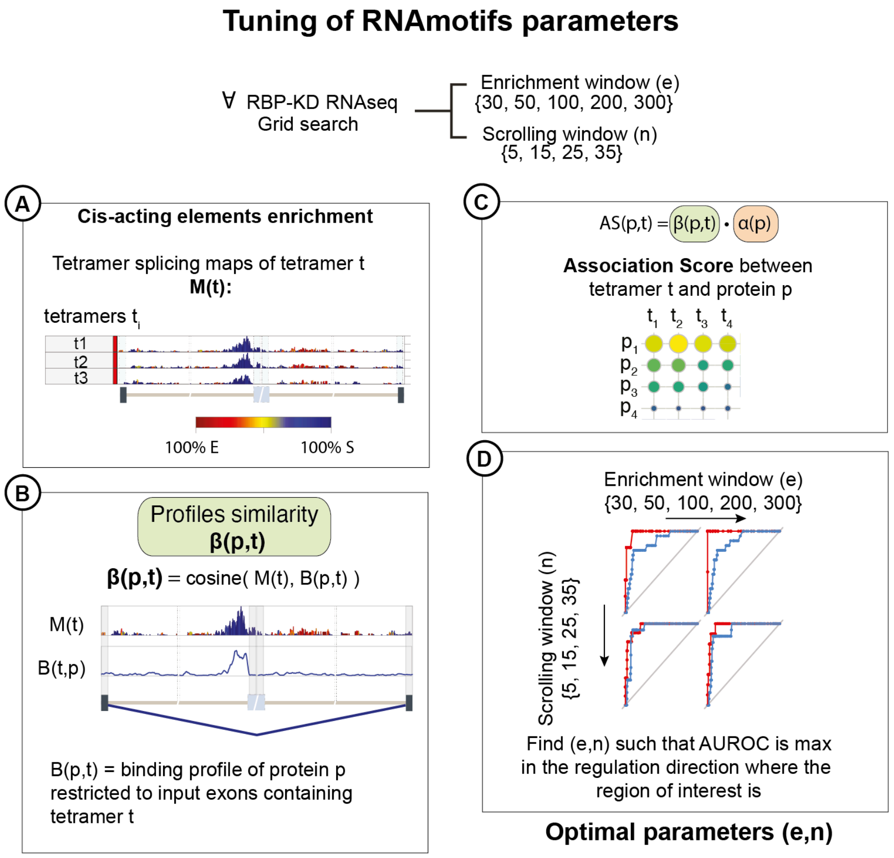
Tuning of RNAmotifs parameters
∀ RBP-KD RNAseq
Grid search
Enrichment window (e)
{30, 50, 100, 200, 300}
Scrolling window (n)
{5, 15, 25, 35}
A
Cis-acting elements enrichment
Tetramer splicing maps of tetramer t
M(t):
tetramers t_i
t1
t2
t3
100% E
100% S
B
Profiles similarity
β(p,t)
β(p,t) = cosine(M(t), B(p,t))
M(t)
B(t,p)
B(p,t) = binding profile of protein p restricted to input exons containing tetramer t
C
AS(p,t) = β(p,t) • α(p)
Association Score between tetramer t and protein p
t1 t2 t3 t4
p1
p2
p3
p4
D
Enrichment window (e)
{30, 50, 100, 200, 300}
Scrolling window (n)
{5, 15, 25, 35}
Find (e,n) such that AUROC is max in the regulation direction where the region of interest is
Optimal parameters (e,n)

Figure 45: VIsual explanation of RNAmotifs parameters tuning procedure. **A.** RNAmotifs splicing map splicing map. **B.** Profile similarity between binding profile and RNAmotifs splicing map. **C.** Association scores between tetramers and proteins. **D.** Evaluation of ROC curves across different combinations of RNAmotifs parameters.

RNAmars algorithm

Given the set of optimal parameters and the signal recovery rate for each protein, representing the base knowledge used by the tool, RNAmars *per se* can be used to identify which proteins regulate the alternative splicing process of differentially included exons retrieved from an RNAseq experiment. Required inputs are: (i) a list of coordinates of exons with their differential inclusion direction (1 for differentially included, -1 for differentially excluded and 0 for constitutive), (ii) the cell line to be used for the eCLIP peaks data (K562 or HepG2).

A differential gene expression table in DESeq2 format (Love, Huber, and Anders 2014) can also be supplied to be used in the summary visualization in the DESeq2 format. Since gene expression is not used for the purpose of retrieving the associations between tetramers and proteins, the latter is totally optional.

RNAmars workflow is composed of three main parts.

(i) First, RNAmotifs (Cereda et al. 2014) is run with the parameters *n* and *e* equal to all the optimal parameters previously found, since *a priori* the regulatory protein of splicing events is still unknown.

(ii) Given a tetramer t_a enriched in at least one run of RNAmotifs (Cereda et al. 2014), to evaluate its association with protein p, the $AS(p,\ t_a)$ is computed as described above, considering only the run matching the optimal parameters of *p*. In other words:

- $AS(\,p, t_a) = AS_a = \alpha(p)_a \cdot \beta(p, t_a)$ if t_x is enriched in run with optimal parameters of protein p
- $AS(\,p, t_a) = 0$ if t_x is not enriched in run with optimal parameters of protein p

(iii) All the AS are then collected into a final AS matrix, in which each row is a different RBP and columns are the tetramers that were found as enriched in at least one of the runs of RNAmotifs (Cereda et al. 2014) under different parameter configurations. The AS matrix is the main output of RNAmars. The AS matrix is also visualized through a heatmap, where cell color and cell size represent the association score between the row protein and the column tetramer (Figure 46). The cell is empty if the column tetramer is not enriched in the run with corresponding optimal parameters of the row protein. The heatmap is built using the 'Complex Heatmap' R package v2.14.0 (Gu, Eils, and Schlesner 2016). The barplot on top depicts a score that combines p-values from the different runs of RNAmotifs (Cereda et al. 2014) through the Fisher's method.

$$Tetramer\ score\ =\ -\sum_{r\,\in\,runs}\ \sum_{i\,\in\,\{R_1,R_2,R_3\}}\left[2\ log\left(pv_{i,r}\right)\right]$$

where i is one of the three regulatory regions R_1, R_2 and R_3 of RNAmotifs (Cereda et al. 2014) and *r* is a RNAmotifs run with a combination of (e, n) parameters.

The resulting vector containing tetramer scores is also provided as an output to the user.

The three-rows heatmap under the barplot represents the aggregation of tetramer enrichments of different runs by Fisher's method, separately for the three regulatory regions (Figure 46). The heatmap with differently sized green squares on the left represents the BS and, hence, in which region the binding of the row protein is expected. Rows are sorted by values reported in the barplot on the left. Such values are the row-wise averages of the association scores of tetramers found enriched in the optimal parameters of the row protein (Figure 46). They represent the final protein scores, interpreted as the level of confidence that the protein is related to the input alternative exons. If the differential expression table is provided as input by the user, the log2 Fold Changes of the genes coding for the row proteins are plotted aside. The sign of the log2 Fold Change is given by the shape and color of the triangles: upwards red triangles for positive log2 Fold Changes and downwards blue triangles for negative ones. The triangles are filled if p-value adjusted is lower than 0.05, empty otherwise. On the right hand side of the heatmap the tetramer splicing maps resulting from the RNAmotifs runs with optimal parameters of the corresponding protein are depicted.

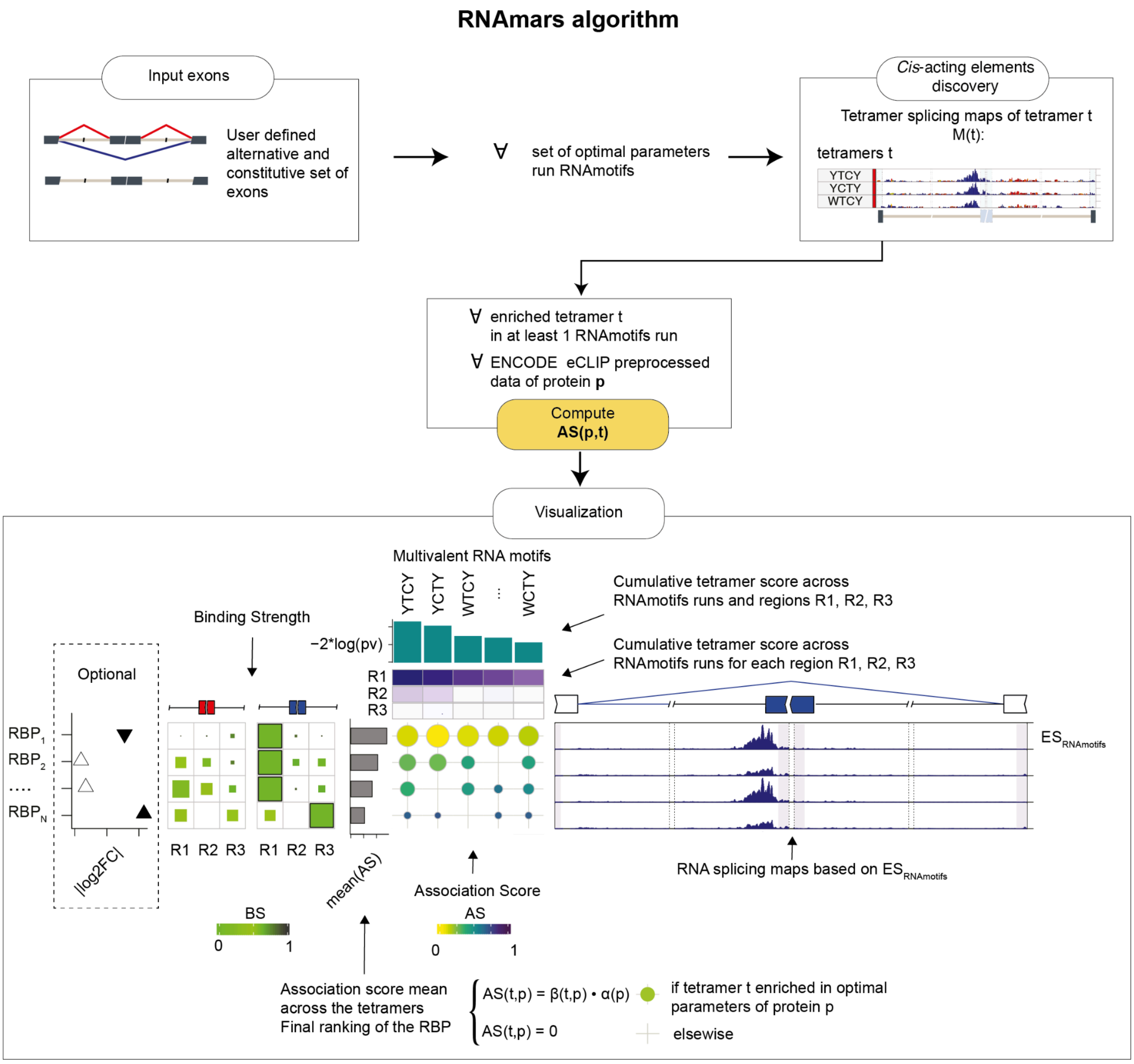
RNAmars algorithm
Input exons
User defined alternative and constitutive set of exons
∀ set of optimal parameters run RNAmotifs
Cis-acting elements discovery
Tetramer splicing maps of tetramer t M(t):
tetramers t
YTCY
YCTY
WTCY
∀ enriched tetramer t in at least 1 RNAmotifs run
∀ ENCODE eCLIP preprocessed data of protein p
Compute AS(p,t)
Visualization
Multivalent RNA motifs
YTCY
YCTY
WTCY
...
WCTY
Cumulative tetramer score across RNAmotifs runs and regions R1, R2, R3
Cumulative tetramer score across RNAmotifs runs for each region R1, R2, R3
Binding Strength
−2*log(pv)
R1
R2
R3
Optional
RBP1
RBP2
....
RBPN
|log2FC|
R1 R2 R3 R1 R2 R3
mean(AS)
ES RNAmotifs
RNA splicing maps based on ES RNAmotifs
Association Score
BS
0
1
AS
0
1
Association score mean across the tetramers
Final ranking of the RBP
AS(t,p) = β(t,p) • α(p)
if tetramer t enriched in optimal parameters of protein p
AS(t,p) = 0
elsewise

Figure 46: Visual explanation of RNAmars algorithm. First alternative and constitutive exons are defined by the user. RNAmotifs are run with the combinations of optimal parameters. Then, Association Scores are computed comparing RNAmotifs splicing maps with binding profiles from HepG2 or K562 eCLIP data. Visualization step: enriched tetramers are in the columns, RBPs are in

the rows and matrix values are the association scores. Point size and color gradient are proportional to the AS. Left annotation contains the absolute log2 Fold Change of differential gene expression. Triangle pointing up means up-regulated , pointing down means down-regulated (sign of log2 fold change of gene expression is user defined). Green squares on the left are the BS of each protein cell line. Left barplot indicates the final RBP score computed as the mean(AS). Top barplot is the cumulative tetramer of the tetramer (Fisher's Method aggregation of p-values from different regions and runs). Upper annotation heatmap refers to the tetramer strength in the three regulatory regions. Tetramer splicing maps on the right in the RNAmotifs run with the optimal parameters of the row protein.

Information content of PWMs

The Information Content (IC) was measured at each of the 11 positions *i* following:

$$IC_i = 2 + \sum_{j \in \{A,C,T,G\}} p_j \cdot log_2(p_j)$$

where p_j is the probability of having the letter *j* in the position *i.* IC ranges from 0 (*i.e.* all the nucleotides have the same probabilities to occur) to 2 (*i.e.* certainty of a given nucleotide in a specific position). The overall IC of the protein in the cell line was defined as the mean across all positions.

PWM of multivalent RNA motifs

To collapse a list of enriched tetramers into a singular PWM two steps were used:

1. Each tetramer was converted into a probability matrix made of 4 columns (the nucleotide of the tetramers) and 4 rows (the nucleotide dictionary). In the case the tetramer was non-degenerate (containing exclusively A, C, T, G), values were set to 1 in correspondence of the position for each letter of the tetramer. For degenerate tetramers (containing Y, R, W, S), instead, a value of 0.5 was given to both the nucleotides composing the degeneracy.
2. Given the list of probability matrices, they were collapsed together by performing a weighted average of the letter frequencies at each position. The weight used was the cumulative tetramer score of the tetramer as defined before. The resulting matrix was composed by occurrences $f_{i,j}$ of nucleotide *i* at position *j.*

Tetramer PWMs were compared between each other using Pearson Correlation Coefficient (PCC). Specifically, at each position PCC of nucleotide frequencies was computed between the two PWMs, resulting in a four dimensional vector of correlations. In particular, for each position *j* in the PWM (*i.e*. the column index):

$$PCC(X_j, Y_j)$$

where $j \in \{1, 2, 3, 4\}$ and X_j and Y_j are four dimensional vectors of nucleotide weights *f* in the position *j* $(f_{A,j}, f_{C,j}, f_{G,j}, f_{T,j})$. The mean of these correlations was then used as a measure of similarities between the two lists of tetramers.

Multivariable covariance analysis

Relative contributions of the technical features to the correlation with a response variable (e.g. the mean score of the input regulatory protein) were measured using the following approach. The technical features, or regressors, considered were: |DPSI| levels, number of input exons, IC of PWMs and number of deregulated RBPs. The number of dysregulated RBPs was determined by identifying splicing-related genes within the dataset of interest that exhibited an adjusted p-value of ≤ 0.05. This was done among the selection of RBPs considered for this study. (15 RBPs for HepG2 and 13 for K562, see Supplementary table 4). Regressor values were normalized using a near-zero variance filter, Yeo-Johnson transformation, centering around their mean, and scaling by their standard deviation using the *preProcess* function in the R 'caret' package v6.0-94 (Kuhn 2008) with parameters method = c("center", "scale", "YeoJohnson", "nzv"). The response variable mean score was set to 0 for datasets without enriched tetramers. A generalized linear regression model (GLM) was fitted to the response variable based on the normalized values of regressors using the *glm* function in the R 'stats' package v4.2.3. Relative importance of each regressor to the correlation measured by the model was calculated using the function *calc.relimp* in the R 'relaimpo' package v2.2-6. In practice, the coefficient of determination R^2 was divided into the contribution of each regressor using the averaging over orderings method (Lindeman, Merenda, and Gold 1980). Confidence intervals were measured using a bootstrapping procedure implemented in the function *boot.relimp*. For 1,000 iterations, the full observation vectors were resampled and the regressor contributions were calculated.

HNRNPK splicing analysis

HNRNPK was silenced in PC3 cell line and RNA-seq library preparation and sequencing was performed as previously described (Del Giudice et al. 2022). Raw sequencing reads from HNRNPK silencing RNAseq were aligned to the human genome reference GENCODE

GRCh37 version 28 (Frankish et al. 2019) using STAR v.2.7.3a (Dobin et al. 2013) in basic two-pass mode using the following parameters: --alignInsertionFlush Right --outSAMstrandField intronMotif --outSAMattributes NH HI NM MD AS XS --peOverlapNbasesMin 20 --peOverlapMMp 0.25 --chimSegmentMin 12 --chimJunctionOverhangMin 8 --chimOutJunctionFormat 1 --chimMultimapScoreRange 3 --chimScoreJunctionNonGTAG -4 --chimMultimapNmax 20 --chimNonchimScoreDropMin 10 --outFilterIntronStrands RemoveInconsistentStrands --outFilterMultimapNmax 1 --bamRemoveDuplicatesType UniqueIdentical.

Alternative splicing events were detected using rMATS v4.1.1 (Dobin et al. 2013; Shen et al. 2014). Cassette exons with FDR<0.01 and |DPSI|>0.1 were defined as alternatively included. Events with FDR>0.1, |DPSI|<0.005, belonging to the gencode comprehensive annotation (Frankish et al. 2019) and not labeled as ‘alt’ in UCSC hg19 database (Navarro Gonzalez et al. 2021) were considered as constitutive.

Chapter 3: Splicing derived neoepitopes

Immunotherapy is the process by which engineered T cells reject tumor cells by binding to antigenic peptides. These peptides are characterized by high affinity with the major histocompatibility complex (MHC) which presents them on the cell surface. Traditionally, therapies have focused mainly on mutations-derived neoantigens. However, expanding the neoepitopes repertoire to alternative splicing derived neoantigens results in an increased immunotherapy target space.

The standard approach to find splicing derived neoepitopes involves several steps (Figure 47). Firstly, it required the identification of tumor-specific splicing junctions (neojunctions). To achieve this, all splicing junctions are compared with junctions from a comprehensive database of human healthy tissues, including the matched normal samples when available. Next, the peptides encoded by the neojunctions are tested for their binding affinity with MHC-I and MHC-II complexes. This process can be done by exploiting prediction tools such as NetMHC and NetMHCpan (Andreatta and Nielsen 2016; Jurtz et al. 2017). The neoepitopes with the highest binding affinities are then selected for validation through mass spectrometry and T cell activation or cytotoxicity assays. Finally, the remaining validated neoepitopes can be employed for vaccination to stimulate a T-cell response in the patient, or they can be utilized to identify antigen-specific T cell receptors (TCRs) for engineering T cells for direct cancer therapy in patients.

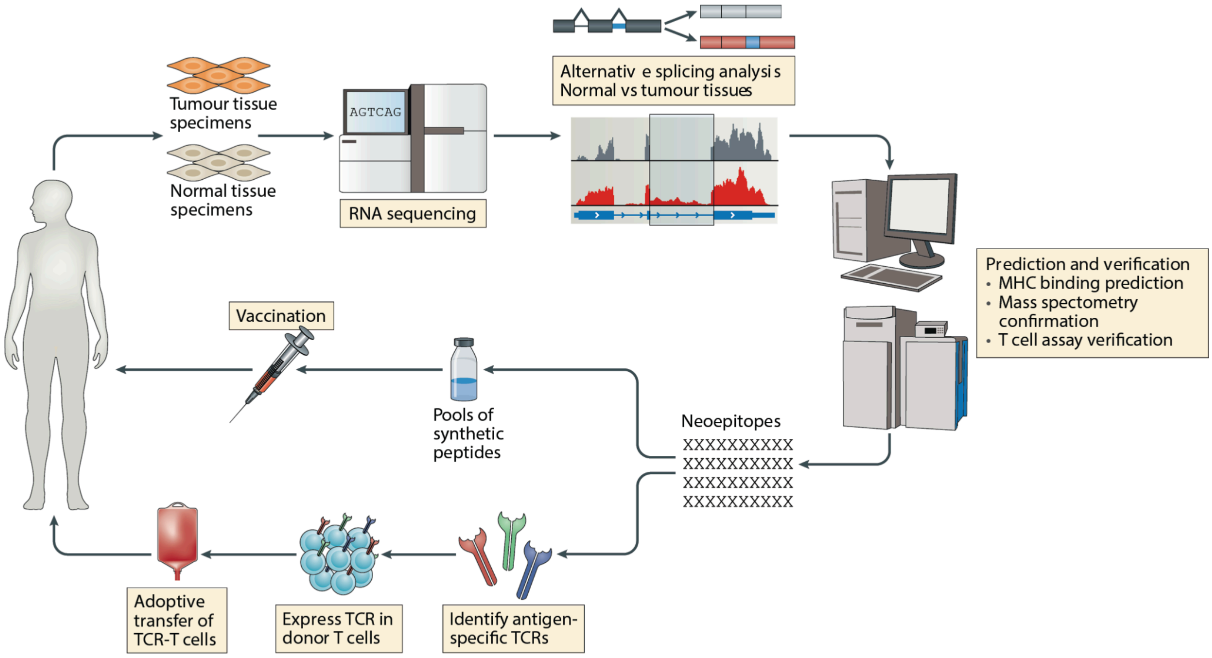
Tumour tissue specimens
Normal tissue specimens
AGTCAG
RNA sequencing
Alternativ e splicing analysi s
Normal vs tumour tissues
Prediction and verification
• MHC binding prediction
• Mass spectometry confirmation
• T cell assay verification
Neoepitopes
XXXXXXXXXX
XXXXXXXXXX
XXXXXXXXXX
XXXXXXXXXX
Pools of synthetic peptides
Vaccination
Identify antigen-specific TCRs
Express TCR in donor T cells
Adoptive transfer of TCR-T cells

Figure 47: Schematic workflow of the identification of splicing derived neoepitopes and the following steps for therapies design. Taken from (Frankiw, Baltimore, and Li 2019)

Recently, some researchers opened the way in this direction: the analysis of 8,705 patients from TCGA showed that 68% of tumors contained at least one alternative splicing derived neoepitopes, while only 30% of tumors contained somatic mutation derived neoepitopes (Kahles et al. 2018). In a study on melanoma patients it was experimentally shown through mass spectrometry immunopeptidome analysis that intron retention events lead to the production of neoepitopes which were presented on the surface of cancer cells (Smart et al. 2018). Moreover Jayasinghe et al. suggested that the neoantigens derived from mis-spliced junctions are more immunogenic than the missense mutations (Jayasinghe et al. 2018). They also showed that a substantial number of neojunctions are recurrent across patients, including events in cancer driver genes such as TP53 and PTEN. In mouse models, the pharmacologic perturbation of RNA splicing has been shown to induce splicing-derived neoepitopes which enhanced the anti-tumor immunity (Lu et al. 2021).

In the case of pediatric EW and OS where the tumor mutational burden is low (Gröbner et al. 2018), studying mRNA alterations can be a promising strategy for expanding the immunotherapy target space. It is crucial to widen the neoantigen space because it was shown that high neoantigen burden implies an increased immune response (Turajlic et al. 2017). In this chapter I will delve into an extensive examination of splicing-derived neoepitopes within the OS and EW cohort.

Results

Alternative splice sites and exon skipping are major sources of neoepitopes

In chapter 1 it was shown that the alternative events generate also to novel junctions that are not found in the genome annotation. Therefore, there could be margin to identifying tumor-specific junctions that potentially become neoepitopes to target with immunotherapy. To do so I exploited ISOTOPE software (J. L. Trincado et al. 2021) which effectively identifies junctions absent in the annotation as well as in control samples, and for which the encoded peptides exhibit high affinity with the MHC. Being sample specific, ISOTOPE helps at finding patient-specific candidates, holding significant promise for advancing the field of personalized medicine.

Not all the novel junctions can lead to an isoform Open Reading Frame (ORF) change if the junction does not fall into the coding part of the transcript. Even if the ORF changes, the novel Peptide affinity is tested against MHC and only those with enough affinity (less than 500nM) are kept. Figure 48 shows the number of novel junctions categorized by these three groups: (i) Junctions where the novel isoform does not result in an ORF change, hence no neoepitopes are generated; (ii) cases where the isoform lead to an ORF change, but the peptides lack the affinity with the patient's MHC; (iii) the isoform undergoes a ORF change and the peptides exhibit affinity with the patient's MHC, These junctions are defined as neoepitopes related events. On average, only 8% of all the novel junctions are found to harbor neoepitopes, while 38% lead to an ORF change but do not produce peptides with high MHC affinity (Figure 48). The remaining 54% correspond to isoforms where the ORF remains unchanged. The proportion of the neojunctions harboring neoepitopes is compatible with what was already found in different cell lines types (J. L. Trincado et al. 2021).

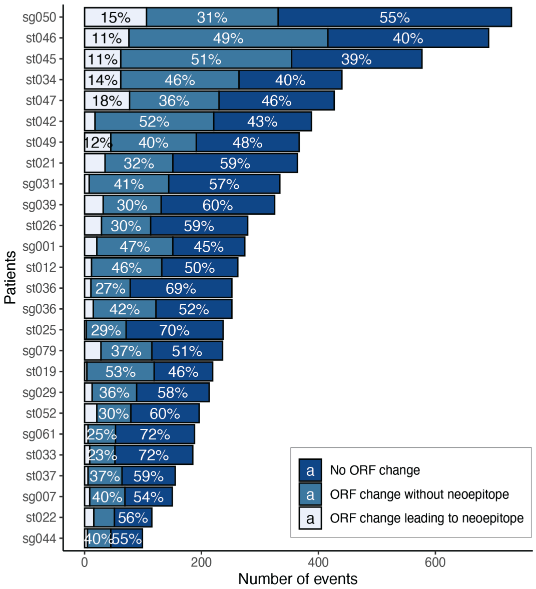

Figure 48: Number of events per sample stratified by their ORF change and their predisposition to form neoepitopes.

The junctions identified by ISOTOPE are divided into four categories: intron retention, alternative 3’ splice site or 5’ splice site (a3a5), exonization (*i.e*. creation of a novel exon) and neoskipping (*i.e.* a novel exon skipping event).

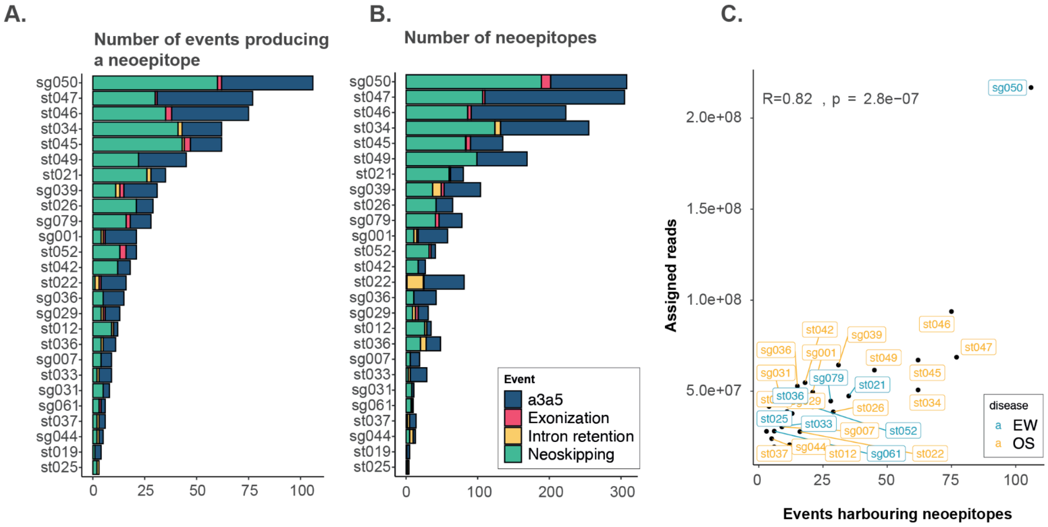

Figure 49: Number of detected neoepitopes and related neojunctions. **A.** Number of alternative events per patient harboring a neoepitope. **B.** Number of neoepitopes per patient. **C.** Scatter plot of assigned reads to genes versus number of events harboring neoepitopes. R is Pearson’s correlation coefficient

The most frequent neoepitope related junction types in samples were either neoskipping or a3a5 (Figure 49). Each junction was on average harboring three neoepitopes. The patient sg050 had the highest number of junctions harboring neoepitopes, with a total of 106 events.

The number of discovered neoepitope-related junctions depends strongly on coverage of the samples. In particular the number of assigned reads (*i.e.* sum of all the counts assigned to genes) is correlated with the number of neoepitope-related junctions (Figure 49). Overall there were 413 unique genes with 546 novel junctions harboring a total of 1698 different neoepitopes across patients.

Neoepitopes are patient specific

The majority of junctions was private, i.e. specific only to that sample, emphasizing the underlying heterogeneity among the patients (Figure 50). Nonetheless, it was crucial for me to prioritize the pursuit of a strong consensus among patients. This meant focusing on identifying candidate neoepitopes that were commonly shared across multiple patients, thus increasing the likelihood of developing a broadly applicable immunotherapy for pediatric sarcomas. The two most shared neoepitopes were found in seven different samples of both subtypes (Figure 50), and they arose from the same alternative event. It is a 26 nucleotides long alternative 5' splice site within the ATF6B gene.

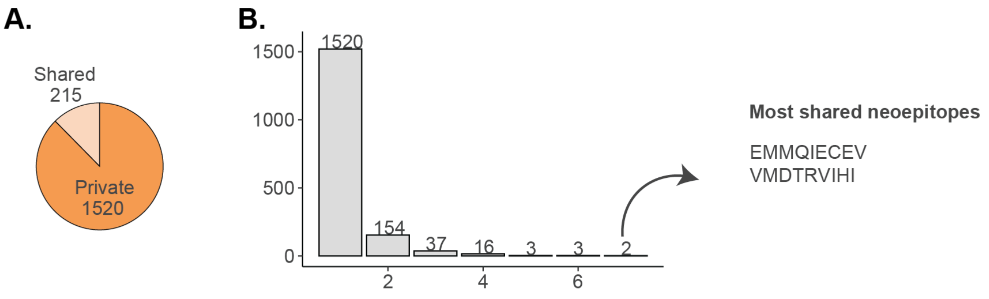

Figure 50: Neoepitopes specificity. A. Pie chart indicating the amount of neoepitopes that are shared and that are patient specific (private). B. Histogram showing the number of neoepitopes shared by N patients, with N takes values in x-axis.

To decipher which biological pathways were involved in genes containing neoepitopes I conducted an over-representation analysis of gene ontology terms. Intriguingly, I observed that the most enriched pathways overlapped with those identified as the top enriched pathways in the down-regulated genes within sarcomas, as detailed in the transcriptional alteration section. Specifically, the top enriched term “External encapsulating structure organization”, had 32 genes harboring neoepitopes in OS. This pathway was also among the top pathways for down-regulated genes in OS.

However, it is important to note that the genes falling within this category exhibited only modest absolute fold changes, with only four of them demonstrating significant downregulation (one gene) or upregulation (three genes), as illustrated in Figure 51.

Consequently, this pathway appeared to be affected by both an overall gene expression downregulation and the formation of splicing-derived neoepitopes, although there was limited overlap in terms of targeted genes between these two mechanisms (only four genes out of the 32).

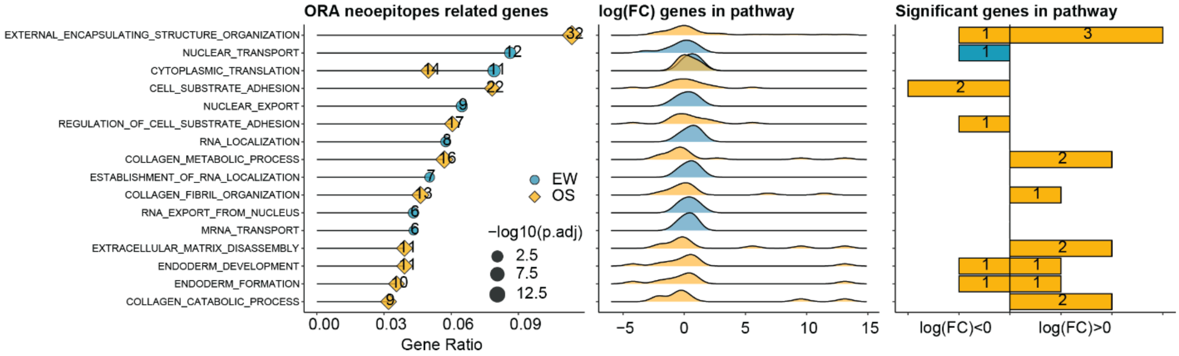

Figure 51: Over-representation analysis of neoepitopes related genes, divided by disease subtype. Biological processes enrichments in the left panel, point size proportional to p-value adjusted and the number over the point defines the number of genes belonging to that pathway. Central panel is a density of the log2 Fold Change distribution from DESeq2 results of the genes within that pathway. In the right hand panel there are the number of significant differentially expressed genes in the corresponding pathway.

Inspecting the number of differentially expressed genes within the neoepitopes related genes I observed that the majority (82% for EW and 90% for OS) were not differentially expressed (Figure 52).

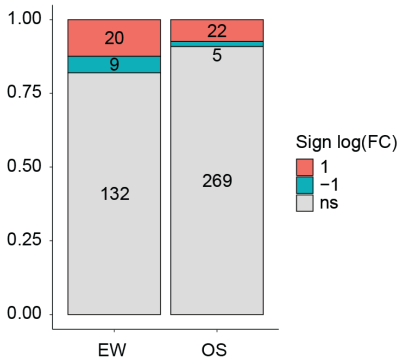

Figure 52: Proportion and number of up-regulated (salmon), down-regulated (cyan) and non significant differentially gene expressed neoepitopes related genes.

Therefore, the origin of these neoepitopes is not due to a different gene activation, considering the low change in expression, between the two tissues, but rather from variations in junction patterns between the tumor and normal tissues. This finding aligns with the observations in Chapter 1, underscoring the substantial dissimilarity between the targets of splicing modifications and those of transcriptional alterations. This disparity extends to splicing changes resulting in neoepitope formation.

Methods

Neoepitopes discovery

HLA genotyping at four digits resolution was estimated for each patient from RNA-seq Fastq files using seq2HLA (Boegel et al. 2012) which internally uses bowtie (v. 0.12).

Junction files from STAR output were merged together using Junckey ("GitHub - comprna/Junckey: Collection of Scripts for Computing PSI of Junction Clusters" n.d.). Transcript counts were computed using Salmon (Patro et al. 2017). ISOTOPE pipeline (J. L. Trincado et al. 2021) was then run to classify splicing-derived neoepitopes in four types of novel junctions: exonizations, neoskippings, retained introns and alternative 3' or 5' splice sites (A3A5). Junctions were classified as novel if all these conditions were satisfied (i) they had at least 20 supporting reads in at least one tumor sample (ii) either one or both of the splice-sites were not present in the human annotation (GENCODE complete annotation GRCh38.v33) (iii) did not appear in any of the normal samples from comprehensive datasets like Intropolis and CHESS-2.2 (iv) the junctions were not present in control samples (with a 20 reads coverage threshold). Since matched normal samples have not been extracted, publicly available osteoblast data were used as controls (Moriarity et al. 2015) (Accession code GSE57925).

Suppa generateEvents function was used to retrieve splicing events from CHESS GTF (Juan L. Trincado et al. 2018). Since Intropolis junctions were only available in hg19, the custom function liftover_intropolis.py from GitHub repository of (Nellore et al. 2016) was used to convert the coordinates to hg38. To retrieve intron retention the reference annotation was widened to also include introns coordinates using KMA software with parameter --extend 40. Raw reads were then mapped to this augmented annotation with Bowtie2 with parameters -k 200 --rdg 6,5 --rfg 6,5 --score-min L,-.6,-.4 and intron retentions were quantified with eXpress -1.5.1 (Roberts and Pachter 2013).

Repeated elements intervals were downloaded from UCSC portal (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/)

Discussion

This project has been conducted with the aim of broadening the repertoire of suitable targets for osteosarcomas (OS) and Ewing (EW) pediatric sarcomas using a statistical approach. Given the limited effectiveness of current therapies for these tumor types and their high recurrence rates, there is a need to explore new therapeutic strategies using a more personalized and genomic approach. The conventional approach for addressing these tumors involves identifying somatic mutations that act as cancer drivers, guiding targeted therapies (Yu, O'Toole, and Trent 2015). However, pediatric sarcomas are notably lowly mutated cancers (Gröbner et al. 2018). Therefore, the objective of this thesis was to uncover additional irregularities within the transcriptomic landscape as potential avenues.

The project tackles the issue from three different perspectives: (i) Exon levels: finding a set of poor survival related events (Chapter 1); (ii) Protein level: identifying the proteins that are majorly involved in alternative splicing dysregulation (Chapter 2); and (iii) Cell level: detecting tumor specific splicing derived neoantigens (Chapter 3).

From a preliminary inspection of transcriptomic data, both OS and EW showed a divergent landscape between transcriptional and splicing alterations targets, highlighting two different levels of regulations.

The integration of The Cancer Genome Atlas TCGA clinical data and the respective survival analysis identified seven alternative splicing events strictly significantly related to a poor patient prognosis in both diseases. In particular, an exon skipping within the tumor suppressor gene NF1 stands out. The skipped exon is 63 nts long and falls within a RasGAP domain, whose activity is to inactivate the oncogenic RAS proteins function. Depletion of this domain could therefore lead to a loss of function of NF1 protein provoking an activation of RAS proliferation pathway. Considering their strong correlation with patient survival, these events are promising candidates for targeted intervention using Splice Switching antisense Oligonucleotides (SSO) (Yuanjiao Zhang et al. 2021). Such technology can potentially lead to a degradation of the damaging isoform, holding the promise of improving patient survival.

In order to pinpoint the key RNA binding proteins (RBPs) that were majorly associated with alternative splicing, RNAmars algorithm was developed. This statistical approach assesses the association of an RBP's involvement in alternative splicing by comparing the positional enrichments of *cis*-acting elements with those of *trans*-acting factors exploiting RBP depletion followed by RNA-seq and eCLIP data from ENCODE database (Van Nostrand, Freese, et al. 2020). The results generated by RNAmars are collected within an interpretable heatmap. This heatmap not only includes the ranking of the proteins that are primarily responsible for splicing dysregulation but also provides valuable details regarding pre-mRNA sequences, binding preferences, and differential gene expression.

RNAmars was tested on unpublished data by transiently silencing HNRNPK in PC3 cell lines. RNAmars was able to identify HNRNPK as the top regulator of its exons, as expected. It also retrieved the expected CC rich motifs of HNRNPK downstream the alternative region for enhanced exons and within the exon itself for silenced ones. This analysis revealed how RNAmars successfully identifies RBPs that are highly probable to bind to the identified motifs, thus playing a pivotal role in the regulation of alternative splicing.

RNAmars was used to analyze alternative exons in sarcomas and it identified RBFOX2 as the major exon enhancer and U2AF2 as the main silenced in both OS and EW. Notably, in OS, HNRNPK and U2AF1 also exhibited some level of involvement, albeit with a lesser degree of impact, influencing cassette exon inclusion and exclusion, respectively.

RBFOX2 in was associated mainly to Serine rich tetramers (-GC-, -CG-, or -GG-), while U2AFs were predicted to be associated mostly with T-rich motifs in both diseases.

With the insights gained from RNAmars results, it becomes feasible to design a small interfering RNA (siRNA) strategy aimed at suppressing the identified deleterious RBPs and restore normal splicing regulation. Furthermore, leveraging the ability of RNAmars to associate RBPs to enriched tetramers, it is also possible to design a SSO to modulate inclusion of RBPs splicing targets by interfering with their binding.

Lastly, the antigens analysis unveiled the patient-specific nature of splicing derived neoepitopes, particularly arising from alternative splice sites of novel exon skipping events. The highest degree of overlap between different samples were two neoepitopes, which were shared among seven patients and originated from an alternative 5’ splice site of ATF6B gene. This analysis on splicing derived neoepitopes expands the repertoire of possible targets for engineered T-cells targeted immunotherapy.

The major limitation of this study stems from the absence of matched normal controls within the clinical trial. The only samples that were available from the patients, apart from the tumor biopsies, were blood specimens utilized for the DNA analysis. Nonetheless, employing blood as control in differential transcriptomic analysis lacks validity due to significant expression level variations between vastly distinct tissues.

The second option was to use publicly available data of bone cell types, such as osteoblasts, osteocytes, osteoclasts and bone lining cells. Unfortunately, I did not find any data from these cell types in databases such as GTEx (The GTEx Consortium* 2013). I was only able to find osteoblast data that had already been used as controls in a study about osteosarcomas (Moriarity et al. 2015).

Additionally, this study could benefit from a larger sample size especially given that there are only 26 tumor samples which can be limiting for robust statistical analyses. However it is worth noting that the clinical trial is ongoing with an estimated completion date in the next few years (ClinicalTrial.gov id:NCT04621201), and new patients are currently being enrolled.

Conclusions

In this thesis, alternative splicing has been demonstrated as a valuable source of potential targetable candidates at the isoform, protein and cellular level. The computational and statistical approaches employed in this project have helped not only enhance the comprehension of transcriptomic abnormalities in sarcomas, but have also identified three different intervention points and their corresponding candidates for potential personalized therapeutic strategies.

References

Afşar, C. U., I. O. Kara, B. K. Kozat, H. Demiryürek, B. B. Duman, and F. Doran. 2013. "Neurofibromatosis Type 1, Gastrointestinal Stromal Tumor, Leiomyosarcoma and Osteosarcoma: Four Cases of Rare Tumors and a Review of the Literature." *Critical Reviews in Oncology/hematology* 86 (2). https://doi.org/10.1016/j.critrevonc.2012.11.001.

Aghamirzaie, Delasa, Eva Collakova, Song Li, and Ruth Grene. 2016. "CoSpliceNet: A Framework for Co-Splicing Network Inference from Transcriptomics Data." *BMC Genomics* 17 (1): 845.

Ajiro, Masahiko, Rong Jia, Yanqin Yang, Jun Zhu, and Zhi-Ming Zheng. 2016. "A Genome Landscape of SRSF3-Regulated Splicing Events and Gene Expression in Human Osteosarcoma U2OS Cells." *Nucleic Acids Research* 44 (4): 1854–70.

Andreatta, M., and M. Nielsen. 2016. "Gapped Sequence Alignment Using Artificial Neural Networks: Application to the MHC Class I System." *Bioinformatics* 32 (4). https://doi.org/10.1093/bioinformatics/btv639.

An, Ning, Xue Yang, Shujun Cheng, Guiqi Wang, and Kaitai Zhang. 2015. "Developmental Genes Significantly Afflicted by Aberrant Promoter Methylation and Somatic Mutation Predict Overall Survival of Late-Stage Colorectal Cancer." *Scientific Reports* 5 (December): 18616.

Attig, J., F. Agostini, C. Gooding, A. M. Chakrabarti, A. Singh, N. Haberman, J. A. Zagalak, et al. 2018. "Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA Processing." *Cell* 174 (5). https://doi.org/10.1016/j.cell.2018.07.001.

Bailey, Timothy L. 1994. *Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers*.

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. 2009. "MEME SUITE: Tools for Motif Discovery and Searching." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkp335.

Barash, Yoseph, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe, and Brendan J. Frey. 2010. "Deciphering the Splicing Code." *Nature* 465 (7294): 53–59.

Biayna, Josep, Helena Mazuelas, Bernat Gel, Ernest Terribas, Gabrijela Dumbovic, Inma Rosas, Juana Fernández-Rodriguez, et al. 2021. "Using Antisense Oligonucleotides for the Physiological Modulation of the Alternative Splicing of NF1 Exon 23a during PC12 Neuronal Differentiation." *Scientific Reports* 11 (1): 3661.

Boegel, Sebastian, Martin Löwer, Michael Schäfer, Thomas Bukur, Jos de Graaf, Valesca Boisguérin, Ozlem Türeci, Mustafa Diken, John C. Castle, and Ugur Sahin. 2012. "HLA Typing from RNA-Seq Sequence Reads." *Genome Medicine* 4 (12): 102.

Bradley, Robert K., and Olga Anczuków. 2023. "RNA Splicing Dysregulation and the Hallmarks of Cancer." *Nature Reviews. Cancer* 23 (3): 135–55.

Brems, H., E. Beert, T. de Ravel, and E. Legius. 2009. "Mechanisms in the Pathogenesis of Malignant Tumours in Neurofibromatosis Type 1." *The Lancet Oncology* 10 (5). https://doi.org/10.1016/S1470-2045(09)70033-6.

Carazo, Fernando, Marian Gimeno, Juan A. Ferrer-Bonsoms, and Angel Rubio. 2019. "Integration of CLIP Experiments of RNA-Binding Proteins: A Novel Approach to Predict Context-Dependent Splicing Factors from Transcriptomic Data." *BMC Genomics* 20 (1): 521.

Carazo, Fernando, Juan P. Romero, and Angel Rubio. 2019. "Upstream Analysis of Alternative Splicing: A Review of Computational Approaches to Predict Context-Dependent Splicing Factors." *Briefings in Bioinformatics* 20 (4): 1358–75.

Cereda, Matteo, Uberto Pozzoli, Gregor Rot, Peter Juvan, Anthony Schweitzer, Tyson Clark, and Jernej Ule. 2014. "RNAmotifs: Prediction of Multivalent RNA Motifs That Control Alternative Splicing." *Genome Biology* 15 (1): R20.

Chowdhry, M., C. Hughes, R. J. Grimer, V. Sumathi, S. Wilson, and L. Jeys. 2009. “Bone Sarcomas Arising in Patients with Neurofibromatosis Type 1.” *The Journal of Bone and Joint Surgery. British Volume* 91-B (9): 1223–26.

“Comprehensive Pancancer Genomic Analysis Reveals (RTK)-RAS-RAF-MEK as a Key Dysregulated Pathway in Cancer: Its Clinical Implications.” 2019. *Seminars in Cancer Biology* 54 (February): 14–28.

Damerell, Victoria, Michael S. Pepper, and Sharon Prince. 2021. “Molecular Mechanisms Underpinning Sarcomas and Implications for Current and Future Therapy.” *Signal Transduction and Targeted Therapy* 6 (1): 246.

Damianov, Andrey, and Douglas L. Black. 2010. “Autoregulation of Fox Protein Expression to Produce Dominant Negative Splicing Factors.” *RNA* 16 (2): 405–16.

Del Giudice, Marco, John G. Foster, Serena Peirone, Alberto Rissone, Livia Caizzi, Federica Gaudino, Caterina Parlato, et al. 2022. “FOXA1 Regulates Alternative Splicing in Prostate Cancer.” *Cell Reports* 40 (13): 111404.

Del Giudice, Marco, Serena Peirone, Sarah Perrone, Francesca Priante, Fabiola Varese, Elisa Tirtei, Franca Fagioli, and Matteo Cereda. 2021. “Artificial Intelligence in Bulk and Single-Cell RNA-Sequencing Data to Foster Precision Oncology.” *International Journal of Molecular Sciences* 22 (9). https://doi.org/10.3390/ijms22094563.

Deng, Lei, Youzhi Liu, Yechuan Shi, Wenhao Zhang, Chun Yang, and Hui Liu. 2020. “Deep Neural Networks for Inferring Binding Sites of RNA-Binding Proteins by Using Distributed Representations of RNA Primary Sequence and Secondary Structure.” *BMC Genomics* 21 (Suppl 13): 866.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21.

Dominguez, Daniel, Peter Freese, Maria S. Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, et al. 2018. “Sequence, Structure, and Context Preferences of Human RNA Binding Proteins.” *Molecular Cell* 70 (5): 854–67.e9.

Dressler, L., M. Bortolomeazzi, M. R. Keddar, H. Misetic, G. Sartini, A. Acha-Sagredo, L. Montorsi, et al. 2022. “Comparative Assessment of Genes Driving Cancer and Somatic Evolution in Non-Cancer Tissues: An Update of the Network of Cancer Genes (NCG) Resource.” *Genome Biology* 23 (1). https://doi.org/10.1186/s13059-022-02607-z.

Farahani, E., H. K. Patra, J. R. Jangamreddy, I. Rashedi, M. Kawalec, Rao Pariti Rk, P. Batakis, and E. Wiechec. 2014. “Cell Adhesion Molecules and Their Relation to (cancer) Cell Stemness.” *Carcinogenesis* 35 (4). https://doi.org/10.1093/carcin/bgu045.

Feng, Huijuan, Suying Bao, Mohammad Alinoor Rahman, Sebastien M. Weyn-Vanhentenryck, Aziz Khan, Justin Wong, Ankeeta Shah, Elise D. Flynn, Adrian R. Krainer, and Chaolin Zhang. 2019. “Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites.” *Molecular Cell* 74 (6): 1189–1204.e6.

Fernandez, Karen S., Michelle L. Turski, Avanthi Tayi Shah, Boris C. Bastian, Andrew Horvai, Steven Hardee, and E. Alejandro Sweet-Cordero. 2019. “Ewing Sarcoma in a Child with Neurofibromatosis Type 1.” *Molecular Case Studies* 5 (5): a004580.

Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, et al. 2019. “GENCODE Reference Annotation for the Human and Mouse Genomes.” *Nucleic Acids Research* 47 (D1): D766–73.

Frankiw, L., D. Baltimore, and G. Li. 2019. “Alternative mRNA Splicing in Cancer Immunotherapy.” *Nature Reviews. Immunology* 19 (11). https://doi.org/10.1038/s41577-019-0195-7.

“Functional Genomic Screening Reveals Splicing of the EWS-FLI1 Fusion Transcript as a Vulnerability in Ewing Sarcoma.” 2016. *Cell Reports* 14 (3): 598–610.

Fu, Xiang-Dong, and Manuel Ares Jr. 2014. “Context-Dependent Control of Alternative Splicing by RNA-Binding Proteins.” *Nature Reviews. Genetics* 15 (10): 689–701.

Gerstberger, S., M. Hafner, and T. Tuschl. 2014. “A Census of Human RNA-Binding Proteins.” *Nature Reviews. Genetics* 15 (12). https://doi.org/10.1038/nrg3813.

Ghanbari, Mahsa, and Uwe Ohler. 2020. “Deep Neural Networks for Interpreting RNA-Binding Protein Target Preferences.” *Genome Research* 30 (2): 214–26.

“GitHub - comprna/Junckey: Collection of Scripts for Computing PSI of Junction Clusters.” n.d. GitHub. Accessed March 13, 2023. https://github.com/comprna/Junckey.

Gröbner, Susanne N., Barbara C. Worst, Joachim Weischenfeldt, Ivo Buchhalter, Kortine Kleinheinz, Vasilisa A. Rudneva, Pascal D. Johann, et al. 2018. “The Landscape of Genomic Alterations across Childhood Cancers.” *Nature* 555 (7696): 321–27.

Grünewald, T. G. P., F. Cidre-Aranaz, D. Surdez, E. M. Tomazou, E. de Álava, H. Kovar, P. H. Sorensen, O. Delattre, and U. Dirksen. 2018. “Ewing Sarcoma.” *Nature Reviews. Disease Primers* 4 (1). https://doi.org/10.1038/s41572-018-0003-x.

Guo, Jihua, Jun Jia, and Rong Jia. 2015. “PTBP1 and PTBP2 Impaired Autoregulation of SRSF3 in Cancer Cells.” *Scientific Reports* 5 (September): 14548.

Gupta, Shobhit, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. 2007. “Quantifying Similarity between Motifs.” *Genome Biology* 8 (2): R24.

Gu, Zuguang, Roland Eils, and Matthias Schlesner. 2016. “Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data.” *Bioinformatics* 32 (18): 2847–49.

Hinman, Melissa N., Alok Sharma, Guangbin Luo, and Hua Lou. 2014. “Neurofibromatosis Type 1 Alternative Splicing Is a Key Regulator of Ras Signaling in Neurons.” *Molecular and Cellular Biology* 34 (12): 2188–97.

Huether, Robert, Li Dong, Xiang Chen, Gang Wu, Matthew Parker, Lei Wei, Jing Ma, et al. 2014. “The Landscape of Somatic Mutations in Epigenetic Regulators across 1,000 Paediatric Cancer Genomes.” *Nature Communications* 5 (1): 1–7.

Hu, Yin, Yan Huang, Ying Du, Christian F. Orellana, Darshan Singh, Amy R. Johnson, Anaïs Monroy, et al. 2013. “DiffSplice: The Genome-Wide Detection of Differential Splicing Events with RNA-Seq.” *Nucleic Acids Research* 41 (2): e39.

Izquierdo, José María, Nuria Majós, Sophie Bonnal, Concepción Martínez, Robert Castelo, Roderic Guigó, Daniel Bilbao, and Juan Valcárcel. 2005. “Regulation of Fas Alternative Splicing by Antagonistic Effects of TIA-1 and PTB on Exon Definition.” *Molecular Cell*. https://doi.org/10.1016/j.molcel.2005.06.015.

Janiszewska, Michalina, Marina Candido Primi, and Tina Izard. 2020. “Cell Adhesion in Cancer: Beyond the Migration of Single Cells.” *The Journal of Biological Chemistry* 295 (8): 2495–2505.

Jayasinghe, Reyka G., Song Cao, Qingsong Gao, Michael C. Wendl, Nam Sy Vo, Sheila M. Reynolds, Yanyan Zhao, et al. 2018. “Systematic Analysis of Splice-Site-Creating Mutations in Cancer.” *Cell Reports* 23 (1): 270–81.e3.

Jiang, Xiulin, Baiyang Liu, Zhi Nie, Lincan Duan, Qiuxia Xiong, Zhixian Jin, Cuiping Yang, and Yongbin Chen. 2021. “The Role of m6A Modification in the Biological Functions and Diseases.” *Signal Transduction and Targeted Therapy* 6 (1): 1–16.

Jolma, Arttu, Jilin Zhang, Estefania Mondragón, Ekaterina Morgunova, Teemu Kivioja, Kaitlin U. Laverty, Yimeng Yin, et al. 2020. “Binding Specificities of Human RNA-Binding Proteins toward Structured and Linear RNA Sequences.” *Genome Research* 30 (7): 962–73.

Jurtz, V., S. Paul, M. Andreatta, P. Marcatili, B. Peters, and M. Nielsen. 2017. “NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data.” *Journal of Immunology* 199 (9). https://doi.org/10.4049/jimmunol.1700893.

Kahles, André, Kjong-Van Lehmann, Nora C. Toussaint, Matthias Hüser, Stefan G. Stark, Timo Sachsenberg, Oliver Stegle, et al. 2018. “Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients.” *Cancer Cell* 34 (2): 211–24.e6.

Kanehisa, M., and S. Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28 (1). https://doi.org/10.1093/nar/28.1.27.

Keene, Jack D., Jordan M. Komisarow, and Matthew B. Friedersdorf. 2006. “RIP-Chip: The Isolation and Identification of mRNAs, microRNAs and Protein Components of Ribonucleoprotein Complexes from Cell Extracts.” *Nature Protocols*.

https://doi.org/10.1038/nprot.2006.47.

Kim, Young-Kook. 2022. “RNA Therapy: Rich History, Various Applications and Unlimited Future Prospects.” *Experimental & Molecular Medicine* 54 (4): 455–65.

König, Julian, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J. Turner, Nicholas M. Luscombe, and Jernej Ule. 2010. “iCLIP Reveals the Function of hnRNP Particles in Splicing at Individual Nucleotide Resolution.” *Nature Structural & Molecular Biology* 17 (7): 909–15.

Kuhn, Max. 2008. “Building Predictive Models in R Using the Caret Package.” *Journal of Statistical Software* 28 (November): 1–26.

Lambert, N., A. Robertson, M. Jangi, S. McGeary, P. A. Sharp, and C. B. Burge. 2014. “RNA Bind-N-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins.” *Molecular Cell* 54 (5). https://doi.org/10.1016/j.molcel.2014.04.016.

Lang, Benjamin, Jae-Seong Yang, Mireia Garriga-Canut, Silvia Speroni, Moritz Aschern, Maria Gili, Tobias Hoffmann, Gian Gaetano Tartaglia, and Sebastian P. Maurer. 2021. “Matrix-Screening Reveals a Vast Potential for Direct Protein-Protein Interactions among RNA Binding Proteins.” *Nucleic Acids Research* 49 (12): 6702–21.

“Learning Similarity with Cosine Similarity Ensemble.” 2015. *Information Sciences* 307 (June): 39–52.

Lee, Stanley Chun-Wei, and Omar Abdel-Wahab. 2016. “Therapeutic Targeting of Splicing in Cancer.” *Nature Medicine* 22 (9): 976–86.

Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. “The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection.” *Cell Systems* 1 (6): 417.

Li, Hongshuai, Jie Yang, Guohui Yang, Jia Ren, Yu Meng, Peiyi Qi, and Nan Wang. 2021. “Identification of Prognostic Alternative Splicing Events in Sarcoma.” *Scientific Reports* 11 (1): 1–9.

Lindeman, Richard Harold, Peter Francis Merenda, and Ruth Z. Gold. 1980. *Introduction to Bivariate and Multivariate Analysis*. Pearson Scott Foresman.

Liu, Yongting, Bin Xie, and Qiong Chen. 2023. “RAS Signaling and Immune Cells: A Sinister Crosstalk in the Tumor Microenvironment.” *Journal of Translational Medicine* 21 (1): 1–14.

Lizio, M., I. Abugessaisa, S. Noguchi, A. Kondo, A. Hasegawa, C. C. Hon, M. de Hoon, et al. 2019. “Update of the FANTOM Web Resource: Expansion to Provide Additional Transcriptome Atlases.” *Nucleic Acids Research* 47 (D1). https://doi.org/10.1093/nar/gky1099.

Llorian, Miriam, Schraga Schwartz, Tyson A. Clark, Dror Hollander, Lit-Yeen Tan, Rachel Spellman, Adele Gordon, et al. 2010. “Position-Dependent Alternative Splicing Activity Revealed by Global Profiling of Alternative Splicing Events Regulated by PTB.” *Nature Structural & Molecular Biology* 17 (9): 1114–23.

Louloupi, Annita, Evgenia Ntini, Thomas Conrad, and Ulf Andersson Vang Ørom. 2018. “Transient N-6-Methyladenosine Transcriptome Sequencing Reveals a Regulatory Role of m6A in Splicing Efficiency.” *Cell Reports* 23 (12): 3429–37.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Lu, Sydney X., Emma De Neef, James D. Thomas, Erich Sabio, Benoit Rousseau, Mathieu Gigoux, David A. Knorr, et al. 2021. “Pharmacologic Modulation of RNA Splicing Enhances Anti-Tumor Immunity.” *Cell* 184 (15): 4032–47.e31.

McLeod, Clay, Alexander M. Gout, Xin Zhou, Andrew Thrasher, Delaram Rahbarinia, Samuel W. Brady, Michael Macias, et al. 2021. “St. Jude Cloud: A Pediatric Cancer Genomic Data-Sharing Ecosystem.” *Cancer Discovery* 11 (5): 1082–99.

Moore, S. W. 2009. “Developmental Genes and Cancer in Children.” *Pediatric Blood & Cancer* 52 (7). https://doi.org/10.1002/pbc.21831.

Moriarity, Branden S., George M. Otto, Eric P. Rahrmann, Susan K. Rathe, Natalie K. Wolf, Madison T. Weg, Luke A. Manlove, et al. 2015. “A Sleeping Beauty Forward Genetic Screen Identifies New Genes and Pathways Driving Osteosarcoma Development and Metastasis.” *Nature Genetics* 47 (6): 615–24.

Navarro Gonzalez, Jairo, Ann S. Zweig, Matthew L. Speir, Daniel Schmelter, Kate R. Rosenbloom, Brian J. Raney, Conner C. Powell, et al. 2021. “The UCSC Genome Browser Database: 2021 Update.” *Nucleic Acids Research* 49 (D1): D1046–57.

Nellore, Abhinav, Andrew E. Jaffe, Jean-Philippe Fortin, José Alquicira-Hernández, Leonardo Collado-Torres, Siruo Wang, Robert A. Phillips III, et al. 2016. “Human Splicing Diversity and the Extent of Unannotated Splice Junctions across Human RNA-Seq Samples on the Sequence Read Archive.” *Genome Biology* 17 (1): 266.

Nwabo Kamdje, Armel Herve, Paul Takam Kamga, Richard Tagne Simo, Lorella Vecchio, Paul Faustin Seke Etet, Jean Marc Muller, Giulio Bassi, et al. 2017. “Developmental Pathways Associated with Cancer Metastasis: Notch, Wnt, and Hedgehog.” *Cancer Biology & Medicine* 14 (2): 109–20.

Oka, Miho, Liu Xu, Toshihiro Suzuki, Toshiaki Yoshikawa, Hiromi Sakamoto, Hayato Uemura, Akiyasu C. Yoshizawa, et al. 2021. “Aberrant Splicing Isoforms Detected by Full-Length Transcriptome Sequencing as Transcripts of Potential Neoantigens in Non-Small Cell Lung Cancer.” *Genome Biology* 22 (1): 9.

Pan, Xiaoyong, Peter Rijnbeek, Junchi Yan, and Hong-Bin Shen. 2018. “Prediction of RNA-Protein Sequence and Structure Binding Preferences Using Deep Convolutional and Recurrent Neural Networks.” *BMC Genomics* 19 (1): 511.

Passacantilli, Ilaria, Paola Frisone, Elisa De Paola, Marco Fidaleo, and Maria Paola Paronetto. 2017. “hnRNPM Guides an Alternative Splicing Program in Response to Inhibition of the PI3K/AKT/mTOR Pathway in Ewing Sarcoma Cells.” *Nucleic Acids Research* 45 (21): 12270–84.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.” *Nature Methods* 14 (4): 417–19.

Paz, Inbal, Amir Argoetti, Noa Cohen, Niv Even, and Yael Mandel-Gutfreund. 2022. “RBPmap: A Tool for Mapping and Predicting the Binding Sites of RNA-Binding Proteins Considering the Motif Environment.” *Methods in Molecular Biology* 2404: 53–65.

Peirone, Serena, Elisa Tirtei, Anna Campello, Caterina Parlato, Simonetta Guarrera, Katia Mareschi, Elena Marini, et al. 2023. “Impaired Neutrophil-Mediated Cell Death Drives Ewing’s Sarcoma in a Two Years Old Child with Down Syndrome.” *medRxiv*. https://doi.org/10.1101/2023.05.30.23289664.

Pervouchine, Dmitri, Yaroslav Popov, Andy Berry, Beatrice Borsari, Adam Frankish, and Roderic Guigó. 2019. “Integrative Transcriptomic Analysis Suggests New Autoregulatory Splicing Events Coupled with Nonsense-Mediated mRNA Decay.” *Nucleic Acids Research* 47 (10): 5293–5306.

Phillips, John W., Yang Pan, Brandon L. Tsai, Zhijie Xie, Levon Demirdjian, Wen Xiao, Harry T. Yang, et al. 2020. “Pathway-Guided Analysis Identifies Myc-Dependent Alternative Pre-mRNA Splicing in Aggressive Prostate Cancers.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (10): 5269–79.

Ponthier, Julie L., Christina Schluepen, Weiguo Chen, Robert A. Lersch, Sherry L. Gee, Victor C. Hou, Annie J. Lo, et al. 2006. “Fox-2 Splicing Factor Binds to a Conserved Intron Motif to Promote Inclusion of Protein 4.1R Alternative Exon 16.” *The Journal of Biological Chemistry* 281 (18): 12468–74.

Ray, Debashish, Hilal Kazan, Esther T. Chan, Lourdes Peña Castillo, Sidharth Chaudhry, Shaheynoor Talukder, Benjamin J. Blencowe, Quaid Morris, and Timothy R. Hughes. 2009. “Rapid and Systematic Analysis of the RNA Recognition Specificities of RNA-Binding Proteins.” *Nature Biotechnology* 27 (7): 667–70.

Roberts, Adam, and Lior Pachter. 2013. “Streaming Fragment Assignment for Real-Time Analysis of Sequencing Experiments.” *Nature Methods* 10 (1): 71–73.

Saulière, Jérôme, Alain Sureau, Alain Expert-Bezançon, and Joëlle Marie. 2006. “The Polypyrimidine Tract Binding Protein (PTB) Represses Splicing of Exon 6B from the

Beta-Tropomyosin Pre-mRNA by Directly Interfering with the Binding of the U2AF65 Subunit.” *Molecular and Cellular Biology* 26 (23): 8755–69.

Sebestyén, E., B. Singh, B. Miñana, A. Pagès, F. Mateo, M. A. Pujana, J. Valcárcel, and E. Eyras. 2016. “Large-Scale Analysis of Genome and Transcriptome Alterations in Multiple Tumors Unveils Novel Cancer-Relevant Splicing Networks.” *Genome Research* 26 (6). https://doi.org/10.1101/gr.199935.115.

Shao, Changwei, Bo Yang, Tongbin Wu, Jie Huang, Peng Tang, Yu Zhou, Jie Zhou, et al. 2014. “Mechanisms for U2AF to Define 3’ Splice Sites and Regulate Alternative Splicing in the Human Genome.” *Nature Structural & Molecular Biology* 21 (11): 997–1005.

Shen, S., J. W. Park, Z. X. Lu, L. Lin, Henry, Y. N. Wu, Q. Zhou, and Y. Xing. 2014. “rMATS: Robust and Flexible Detection of Differential Alternative Splicing from Replicate RNA-Seq Data.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (51). https://doi.org/10.1073/pnas.1419161111.

Siddaway, Robert, Scott Milos, Arun Kumaran Anguraj Vadivel, Tara H. W. Dobson, Jyothishmathi Swaminathan, Scott Ryall, Sanja Pajovic, et al. 2022. “Splicing Is an Alternate Oncogenic Pathway Activation Mechanism in Glioma.” *Nature Communications* 13 (1): 1–14.

Smart, Alicia C., Claire A. Margolis, Harold Pimentel, Meng Xiao He, Diana Miao, Dennis Adeegbe, Tim Fugmann, Kwok-Kin Wong, and Eliezer M. Van Allen. 2018. “Intron Retention Is a Source of Neoepitopes in Cancer.” *Nature Biotechnology* 36 (11): 1056–58.

Stanley, Robert F., and Omar Abdel-Wahab. 2022. “Dysregulation and Therapeutic Targeting of RNA Splicing in Cancer.” *Nature Cancer* 3 (5): 536–46.

Sterne-Weiler, Timothy, Robert J. Weatheritt, Andrew Best, Kevin C. H. Ha, and Benjamin J. Blencowe. n.d. “*Whippet*: An Efficient Method for the Detection and Quantification of Alternative Splicing Reveals Extensive Transcriptomic Complexity.” https://doi.org/10.1101/158519.

Stitzinger, Simon H., Salma Sohrabi-Jahromi, and Johannes Söding. 2023. “Cooperativity Boosts Affinity and Specificity of Proteins with Multiple RNA-Binding Domains.” *NAR Genomics and Bioinformatics* 5 (2): lqad057.

The GTEx Consortium*. 2013. “The Genotype-Tissue Expression (GTEx) Project.” *Nature Genetics* 45 (6): 580.

Therneau, Terry M., and Patricia M. Grambsch. 2013. *Modeling Survival Data: Extending the Cox Model*. Springer Science & Business Media.

Tirtei, Elisa, Matteo Cereda, Elvira De Luna, Paola Quarello, Sebastian Dorin Asaftei, and Franca Fagioli. 2020. “Omic Approaches to Pediatric Bone Sarcomas.” *Pediatric Blood & Cancer* 67 (2): e28072.

Tomazini, Atilio, and Julia M. Shifman. 2023. “Targeting Ras with Protein Engineering.” *Oncotarget* 14 (January): 672–87.

Trincado, J. L., M. Reixachs-Solé, J. Pérez-Granado, T. Fugmann, F. Sanz, J. Yokota, and E. Eyras. 2021. “ISOTOPE: ISOform-Guided Prediction of epiTOPEs in Cancer.” *PLoS Computational Biology* 17 (9). https://doi.org/10.1371/journal.pcbi.1009411.

Trincado, Juan L., Juan C. Entizne, Gerald Hysenaj, Babita Singh, Miha Skalic, David J. Elliott, and Eduardo Eyras. 2018. “SUPPA2: Fast, Accurate, and Uncertainty-Aware Differential Splicing Analysis across Multiple Conditions.” *Genome Biology* 19 (1): 40.

Turajlic, Samra, Kevin Litchfield, Hang Xu, Rachel Rosenthal, Nicholas McGranahan, James L. Reading, Yien Ning S. Wong, et al. 2017. “Insertion-and-Deletion-Derived Tumour-Specific Neoantigens and the Immunogenic Phenotype: A Pan-Cancer Analysis.” *The Lancet Oncology* 18 (8): 1009–21.

Ule, Jernej, and Benjamin J. Blencowe. 2019. “Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution.” *Molecular Cell* 76 (2): 329–45.

Ule, Jernej, Kirk B. Jensen, Matteo Ruggiu, Aldo Mele, Aljaz Ule, and Robert B. Darnell. 2003. “CLIP Identifies Nova-Regulated RNA Networks in the Brain.” *Science* 302 (5648): 1212–15.

Underwood, Jason G., Paul L. Boutz, Joseph D. Dougherty, Peter Stoilov, and Douglas L.

Black. 2005. “Homologues of the Caenorhabditis Elegans Fox-1 Protein Are Neuronal Splicing Regulators in Mammals.” *Molecular and Cellular Biology* 25 (22): 10005–16.

Van Nostrand, Eric L., Peter Freese, Gabriel A. Pratt, Xiaofeng Wang, Xintao Wei, Rui Xiao, Steven M. Blue, et al. 2020. “A Large-Scale Binding and Functional Map of Human RNA-Binding Proteins.” *Nature* 583 (7818): 711–19.

Van Nostrand, Eric L., Gabriel A. Pratt, Brian A. Yee, Emily C. Wheeler, Steven M. Blue, Jasmine Mueller, Samuel S. Park, et al. 2020. “Principles of RNA Processing from Analysis of Enhanced CLIP Maps for 150 RNA Binding Proteins.” *Genome Biology* 21 (1): 90.

Venables, J. P., R. Klinck, C. Koh, J. Gervais-Bird, A. Bramard, L. Inkel, M. Durand, et al. 2009. “Cancer-Associated Regulation of Alternative Splicing.” *Nature Structural & Molecular Biology* 16 (6). https://doi.org/10.1038/nsmb.1608.

Venables, Julian P., Jamal Tazi, and François Juge. 2012. “Regulated Functional Alternative Splicing in Drosophila.” *Nucleic Acids Research* 40 (1): 1–10.

Venkataramany, A. S., K. M. Schieffer, K. Lee, C. E. Cottrell, P. Y. Wang, E. R. Mardis, T. P. Cripe, and D. S. Chandler. 2022. “Alternative RNA Splicing Defects in Pediatric Cancers: New Insights in Tumorigenesis and Potential Therapeutic Vulnerabilities.” *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 33 (6): 578–92.

Wang, Ting-You, Qi Liu, Yanan Ren, Sk Kayum Alam, Li Wang, Zhu Zhu, Luke H. Hoeppner, Scott M. Dehm, Qi Cao, and Rendong Yang. 2021. “A Pan-Cancer Transcriptome Analysis of Exitron Splicing Identifies Novel Cancer Driver Genes and Neoepitopes.” *Molecular Cell* 81 (10): 2246–60.e12.

Wei, G., M. Almeida, G. Pintacuda, H. Coker, J. S. Bowness, J. Ule, and N. Brockdorff. 2021. “Acute Depletion of METTL3 Implicates N 6-Methyladenosine in Alternative Intron/exon Inclusion in the Nascent Transcriptome.” *Genome Research* 31 (8). https://doi.org/10.1101/gr.271635.120.

Xue, Yuanchao, Yu Zhou, Tongbin Wu, Tuo Zhu, Xiong Ji, Young-Soo Kwon, Chao Zhang, et al. 2009. “Genome-Wide Analysis of PTB-RNA Interactions Reveals a Strategy Used by the General Splicing Repressor to Modulate Exon Inclusion or Skipping.” *Molecular Cell* 36 (6): 996–1006.

Yang, Xia, Wen-Ting Huang, Rong-Quan He, Jie Ma, Peng Lin, Zu-Cheng Xie, Fu-Chao Ma, and Gang Chen. 2019. “Determining the Prognostic Significance of Alternative Splicing Events in Soft Tissue Sarcoma Using Data from The Cancer Genome Atlas.” *Journal of Translational Medicine* 17 (1): 283.

Yap, Yoon-Sim, John R. McPherson, Choon-Kiat Ong, Steven G. Rozen, Bin-Tean Teh, Ann S. G. Lee, and David F. Callen. 2014. “The NF1 Gene Revisited – from Bench to Bedside.” *Oncotarget* 5 (15): 5873–92.

Yee, Brian A., Gabriel A. Pratt, Brenton R. Graveley, Eric L. Van Nostrand, and Gene W. Yeo. 2019. “RBP-Maps Enables Robust Generation of Splicing Regulatory Maps.” *RNA* 25 (2): 193.

Yuan, Jimin, Xiaoduo Dong, Jiajun Yap, and Jiancheng Hu. 2020. “The MAPK and AMPK Signalings: Interplay and Implication in Targeted Cancer Therapy.” *Journal of Hematology & Oncology* 13 (1): 1–19.

Yu, Bing, Sandra A. O’Toole, and Ronald J. Trent. 2015. “Somatic DNA Mutation Analysis in Targeted Therapy of Solid Tumours.” *Translational Pediatrics* 4 (2): 125.

Zarnack, Kathi, Julian König, Mojca Tajnik, Iñigo Martincorena, Sebastian Eustermann, Isabelle Stévant, Alejandro Reyes, Simon Anders, Nicholas M. Luscombe, and Jernej Ule. 2013. “Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements.” *Cell* 152 (3): 453–66.

Zhang, Chaolin, and Robert B. Darnell. 2011. “Mapping in Vivo Protein-RNA Interactions at Single-Nucleotide Resolution from HITS-CLIP Data.” *Nature Biotechnology* 29 (7): 607–14.

Zhang, Jidong, Bo Liu, Zhihan Wang, Klaus Lehnert, and Mark Gahegan. 2022. “DeepPN: A Deep Parallel Neural Network Based on Convolutional Neural Network and Graph

Convolutional Network for Predicting RNA-Protein Binding Sites." *BMC Bioinformatics* 23 (1): 257.

Zhang, Yangjun, Xiangyang Yao, Hui Zhou, Xiaoliang Wu, Jianbo Tian, Jin Zeng, Libin Yan, et al. 2021. "OncoSplicing: An Updated Database for Clinically Relevant Alternative Splicing in 33 Human Cancers." *Nucleic Acids Research* 50 (D1): D1340–47.

Zhang, Yiming, Ran Zhou, and Yuan Wang. 2022. "Sashimi.py: A Flexible Toolkit for Combinatorial Analysis of Genomic Data." *bioRxiv*. https://doi.org/10.1101/2022.11.02.514803.

Zhang, Yuanjiao, Jinjun Qian, Chunyan Gu, and Ye Yang. 2021. "Alternative Splicing and Cancer: A Systematic Review." *Signal Transduction and Targeted Therapy* 6 (1): 1–14.

Zhou, Delong, Sonia Couture, Michelle S. Scott, and Sherif Abou Elela. 2021. "RBFOX2 Alters Splicing Outcome in Distinct Binding Modes with Multiple Protein Partners." *Nucleic Acids Research* 49 (14): 8370–83.

Zhou, Weiwei, Qiuling Jie, Tao Pan, Jingyi Shi, Tiantongfei Jiang, Ya Zhang, Na Ding, Juan Xu, Yanlin Ma, and Yongsheng Li. 2023. "Single-Cell RNA Binding Protein Regulatory Network Analyses Reveal Oncogenic HNRNPK-MYC Signalling Pathway in Cancer." *Communications Biology* 6 (1): 82.

Zhu, Yiran, Liyuan Zhu, Xian Wang, and Hongchuan Jin. 2022. "RNA-Based Therapeutics: An Overview and Prospectus." *Cell Death & Disease* 13 (7): 644.