University of Torino
Doctoral School in Life and Health Sciences
PhD Program in Complex System for Life Sciences

Tissue specific characterization of splicing factor transcriptional regulation

Serena Peirone

Faculty tutor: Prof. Michele Caselle

External tutor: Prof. Matteo Cereda

Cycle XXXIV

A.A. 2020-2021

Abstract

Pre-mRNA alternative splicing (AS) is a fundamental step in the maturation of transcripts. It affects the majority of human genes and is mainly regulated by RNA-binding proteins (RBPs). By profiling thousands of samples, genomic studies have shown that dysregulation of RBP expression and RBPs-mRNA interactions can lead to a variety of diseases including different cancer types. However, disentangling the heterogeneity of cancer transcriptomes to identify defective transcriptional programs reshaping AS remains challenging.

During my Ph.D., I analyzed transcriptomic, epigenomics, chromosome conformation, and protein-mRNA interaction data in 15 distinct tissue types to evaluate the transcriptional regulation of RBPs in cancer. Starting from the gene expression information level, I developed an algorithm, namely GSECA, to identify altered biological processes in RNA-sequencing (RNA-seq) data accounting for their intrinsic heterogeneity. Applied to 8,464 samples from 19 different cancer types, GSECA outperformed conventional algorithms and revealed that AS is the most altered gene set across cancer types as compared to normal tissues. Combining RNA-seq data with information about transcription factor (TF) binding sites and regions of active transcription, I developed a statistical framework to identify candidate TFs that control RBP expression. While each tissue presents specific transcriptional regulation programs, a set of TFs, including the oncogenic MYC, MAX, and FOXA1 emerged as frequently related to RBP expression. To assess these findings, I focused on prostate cancer (PC) which is driven by somatic alterations of MYC and FOXA1. Analyses of 409 primary and 118 castration-resistant PC RNA-seq data confirmed that FOXA1 is the primary manager of AS regulation. To gain mechanistic insights into FOXA1 transcriptional regulation of RBPs, I integrated the statistical framework with 3D chromatin conformation data in PC cell lines. I showed that FOXA1 controls the RBP expression by actively binding to their promoters or interacting enhancer regions. By combining results of RNA-seq and ENCODE experiments, such as assessment of RBP cross-linking sites by eCLIP and splicing changes upon RBP depletion, I found that FOXA1 orchestrates exon usage by modulating the expression of a set of RBPs. Furthermore, FOXA1-mediated AS regulation targets highly evolutionary conserved exons that are functionally relevant for the cell, as nonsense-mediated-determinant exons. Overall, with this work, I revealed that RBP expression is generally altered in tumors and identified the candidate TFs responsible for this

phenomenon. Furthermore, I demonstrated that the pioneer factor FOXA1 directly controls RBP expression in PC, orchestrating the AS of functionally relevant exons.

Organization of the thesis

This thesis describes a study that addresses open questions concerning splicing, its deregulation in cancer and the actors, *i.e.* the transcription factors, that are responsible for the expression of splicing regulators in a tissue-specific manner. Since different aspects, from biological to computational ones, were deeply analyzed I decided to describe each work in a chapter. Each chapter is fully self-consistent and motivates the following study in a unique flow. Each chapter ends with a specific discussion of the findings. The chapter "1" serves as an introduction to immerse the reader in the context of the study, while the Appendices part is meant to aid the curious reader to satisfy the desire to know more about certain topics. A broader conclusion and discussion are given in the final chapter.

Writing style

Research can be now considered mostly a team work and consequently rather ’we’ instead of ’I’. This work was done in collaboration with other researchers who contributed with ideas, knowledge, and support. Therefore, I decided to use the ‘we’ style in writing this thesis.

1. The transcriptional control of alternative splicing

The beginning is the most important part of the work.

Plato

The “central dogma” of biology

Cell is a complex system, in which processes of regulation, transport, construction, transformation of chemical bonds in energy and vice versa take place. In this work we focused on the regulation of gene expression, particularly on the interplay between transcription and splicing factors.

The process of gene expression comprises different molecules: from DNA, to RNA, and to proteins. In 1953, the hypothesis that chromosomal DNA functioned as a mold for RNA molecules that suddenly were transported in the cytoplasm, where they determined the order of amino acids within proteins, was formulated. In 1956, Francis Crick defined this flux of genetic information as the “central dogma” (Figure 1).

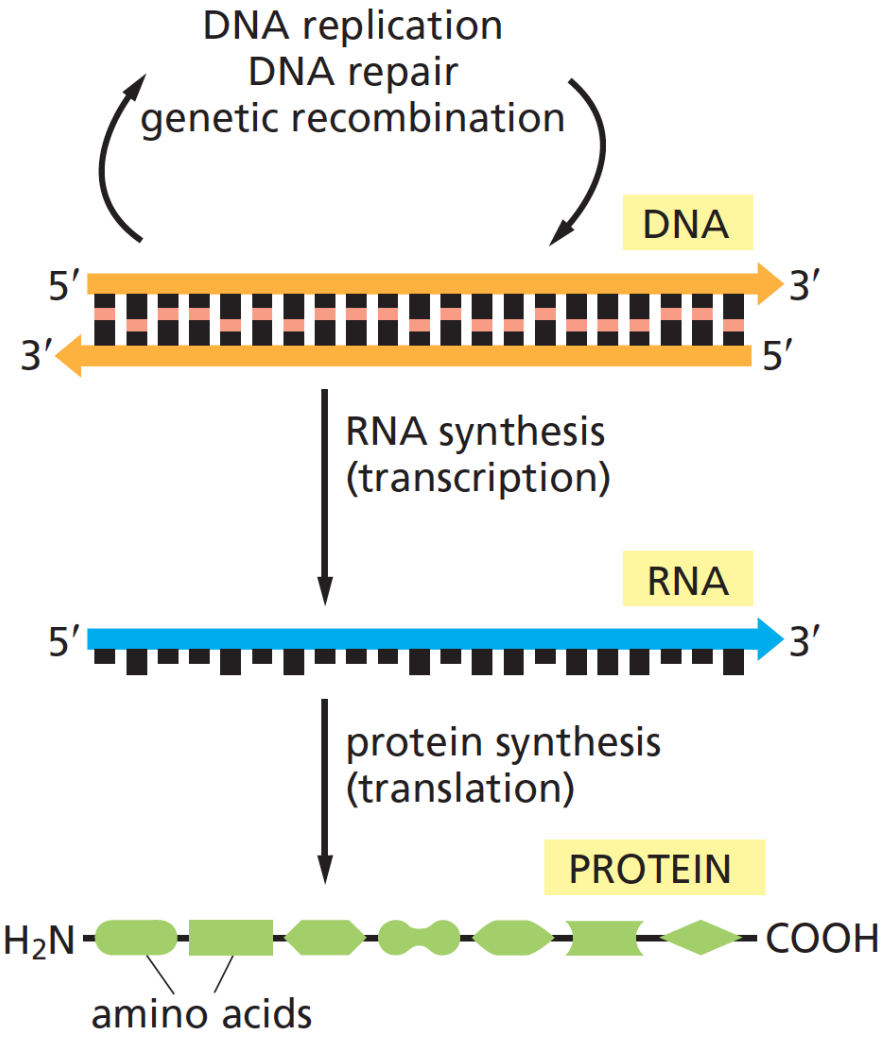


Figure 1. From (Alberts et al. 2002), the pathway from DNA to protein.

DNA is a nucleic acid containing genetic information. It is made up of nucleotides. Each nucleotide comprises a nitrogen base (i.e. cytosine (C), adenine (A), thymine (T) or guanine (G)), a sugar (deoxyribose), and a phosphate group. Bases associate to each other in a strictly complementary way: cytosine with guanine and thymine with adenine. DNA forms a double helix, robust thanks to hydrogen bonds and partially wrapped around histones; it is organized in chromosomes.

RNA, like DNA, is a nucleic acid. The only differences are that RNA contains a different sugar, ribose, and uracil instead of thymine. It is generally single stranded, nevertheless it can form secondary structures. There are three principal types of RNA, namely messenger RNA (mRNA), which encodes DNA information for the creation of proteins, ribosomal RNA (rRNA) which encodes for ribosome components and transfer RNA (tRNA), which helps decode a mRNAsequence into a protein.

Proteins are polymers made up of twenty amino acids. Each amino acid includes an amino group, a carboxylic acid group and a specific side chain; it is encoded in DNA by a triplet of bases, called codon. Thanks to different side chain characteristics they reach specific shapes that are fundamental for cell good health. Proteins cover a lot of different tasks within the cell, from structural function to regulation of transcription. Also histones, DNA supports, are a composition of proteins.

When the cell needs a particular protein, the nucleotide sequence of the appropriate portion of the immensely long DNA molecule in a chromosome is first copied into RNA (a process called transcription). These RNAs are used as templates to direct the synthesis of the protein (a process called translation).

Over time the central dogma has been expanded and several other mechanisms were discovered. We focused on transcription factors' regulation and splicing.

Transcriptional regulation of gene expression

Transcription factors (TFs) are conserved proteins that bind on specific binding sites on the genome (Lambert et al. 2018) and are divided into enhancers and silencers on their function. The former have the role of recruiting RNA polymerase II, who is responsible for transcription; conversely, the others avoid it. TFs can bind both in proximity of the target gene (on the promoters) or thousands of base pairs upstream or downstream (on the enhancers). Promoters are specific sequences of nucleotides indicating the starting point for

RNA synthesis. Several TFs can interact between them thanks to the formation of handles in the DNA (Figure 2).

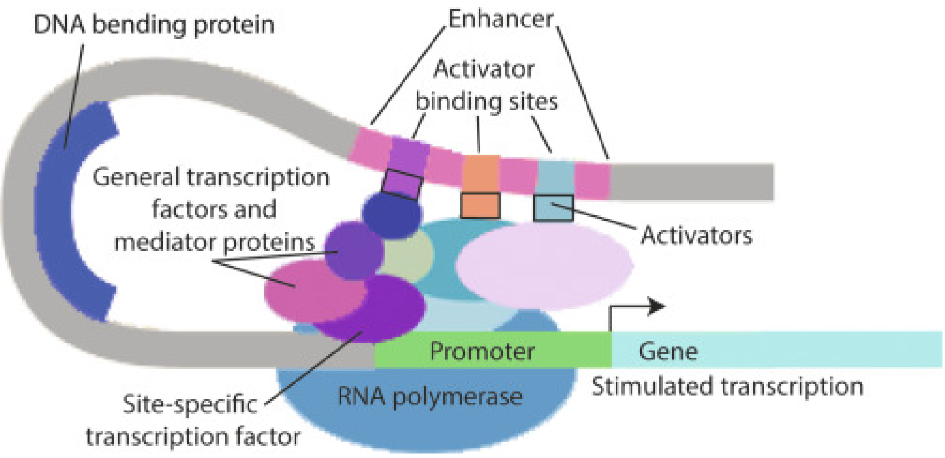


Figure 2. From (Mobley 2019), Scheme of transcriptional regulation.

Some TFs act as pioneer factors binding to nucleosomal target sites and reorganizing local chromatin to increase their accessibility (Donaghey et al. 2018).

TFs that present different types of domain: DNA-binding domains (DBDs) and activation domains (ADs). While the binding ones have been widely studied and contribute to the specificity of TF action by recognising specific motifs on the genome (Rot et al. 2017; Z. Chen et al. 2015), it is not clear how the activation domains influence gene expression. Recently, it has been shown that some TF's ADs form phase-separated condensates with Mediator and this mechanism could explain gene activation (Boija et al. 2018) (Figure 3). In particular, Mediator is a coactivator factor that is required for stabilizing other TF on the promoter for allowing transcription initiation. Bojia et al. have shown that the ability to phase separate with Mediator, which would employ the features of high valency and low affinity characteristic of liquid-liquid phase-separated condensates, operates alongside an ability of some TFs to form high affinity interactions with Mediator. According to the authors, this mechanism can explain the involvement of several TFs in the transcription initiation of genes.

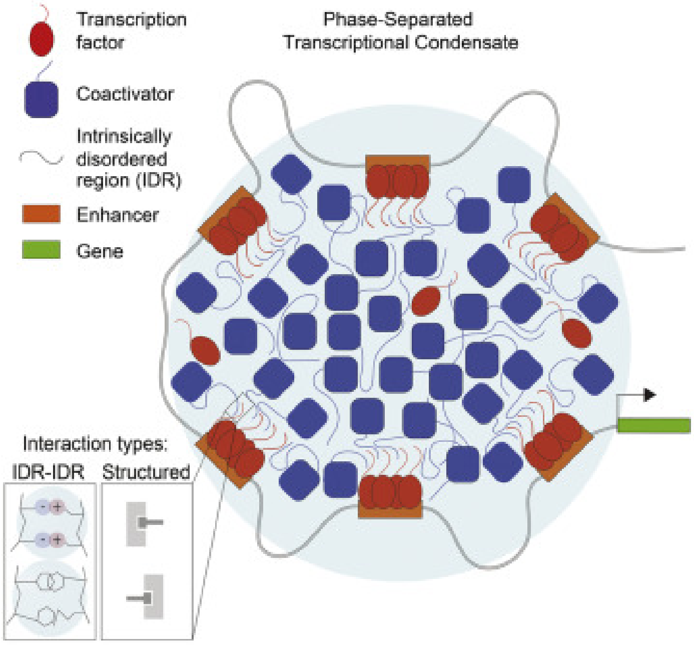
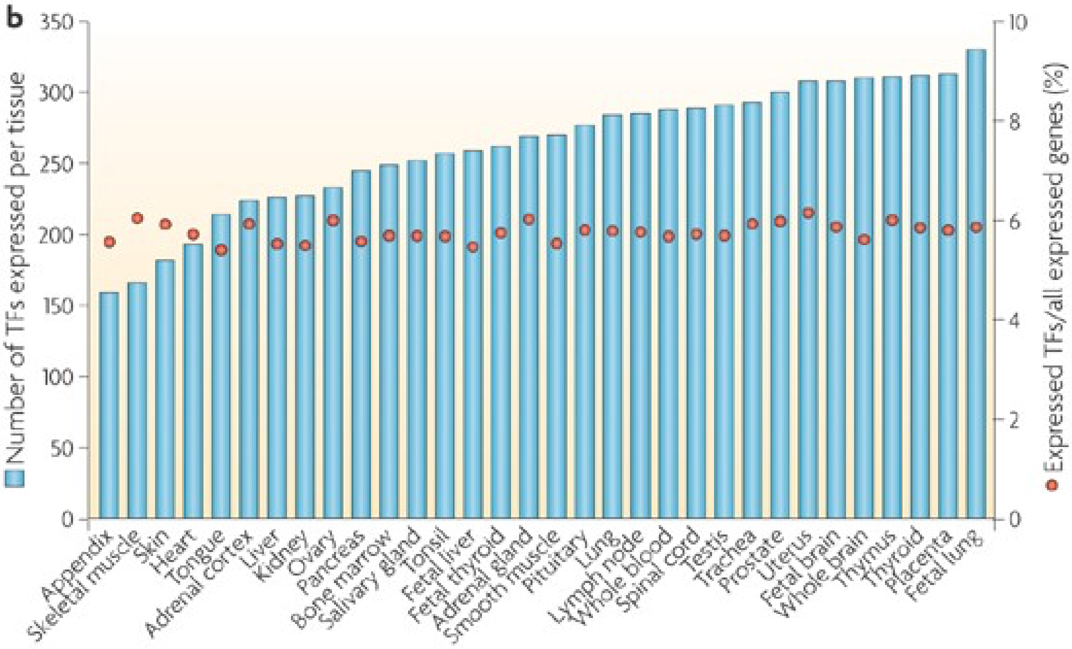

Transcription factor
Coactivator
Intrinsically disordered region (IDR)
Enhancer
Gene
Phase-Separated Transcriptional Condensate
Interaction types:
IDR-IDR
Structured

Figure 3. From (Boija et al. 2018), Phase-separated transcriptional condensate.

The number of expressed transcription factors varies between tissue types, while the proportion of expressed TFs with respect to all genes is almost constant (Figure 4). The number of TFs shared between tissues, instead, follows a U shape, so they tend to be expressed in few tissues or in the majority of them (Vaquerizas et al. 2009) (Figure 4).



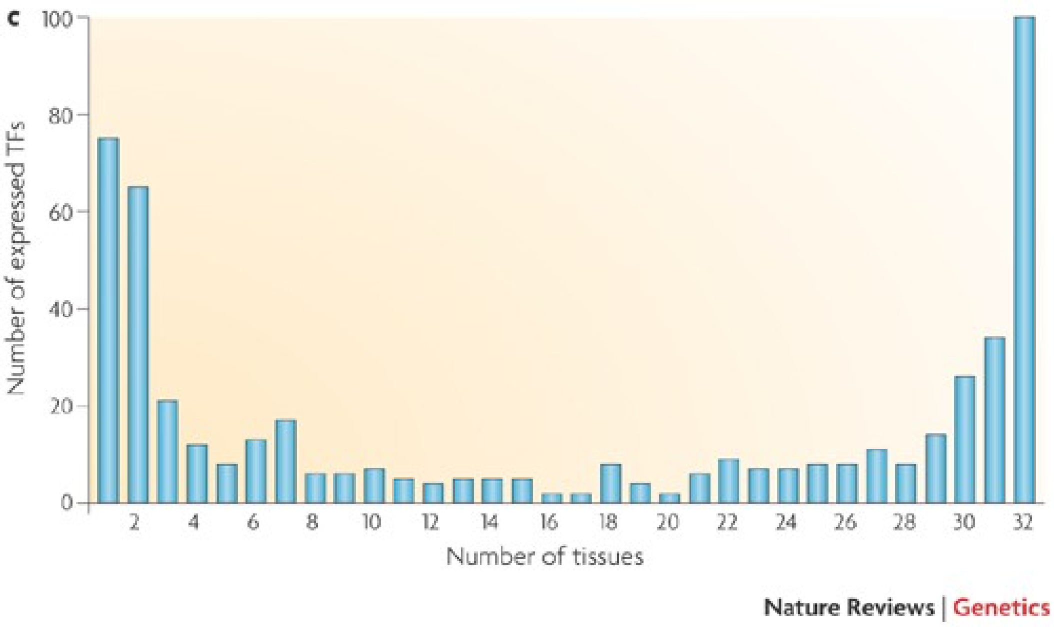


Figure 4. From (Vaquerizas et al. 2009), b. Tissue-specific number of expressed transcription factors. C. Number of shared expressed transcription factors by different tissues.

Nevertheless, once transcribed, pre-mRNA molecules are processed to obtain the final mRNA that can potentially be translated.

Splicing of the pre-mRNA

Among co-transcriptional processes, splicing is a fundamental step in the production of mature mRNA where non-coding sequences (introns) are removed and protein coding sequences (exons) are joined together to form mature RNA.

More than 95% of genes present different isoforms that encode for different proteins, thus amplifying protein diversity (Su et al. 2006) (Figure 5).

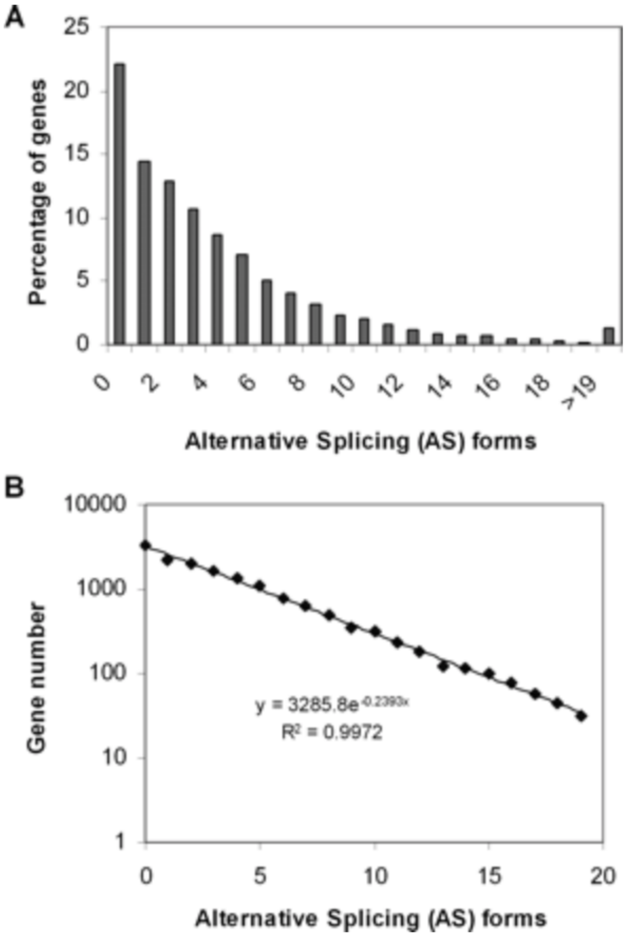


Figure 5. From (Su et al. 2006), A.B. Overviews of the number of isoforms belong to a single gene.

Different isoforms are characterized by different splice sites, namely the boundaries of an exon and an intron. Given the exon-intron structure of a gene, constitutive splice sites can be distinguished from alternative ones. A constitutive splice site is always included into the mature transcript, while an alternative one can be omitted sometimes. Likewise, the terms constitutive and alternative are applied to exons and to the splicing process. As shown in Figure 6, most alternative splice events can be classified into the following basic types:

• the inclusion or exclusion of one (or more) exons, giving rise to a cassette exon,

• the usage of alternative 5’ss or 3‘ss,

• the mutual exclusion of exons, involves the selection of an exon from an array of two or more exon variants,

• the retention of an intron (Z. Wang and Burge 2008; El Marabti and Younis 2018), also known as intron retention (IR)

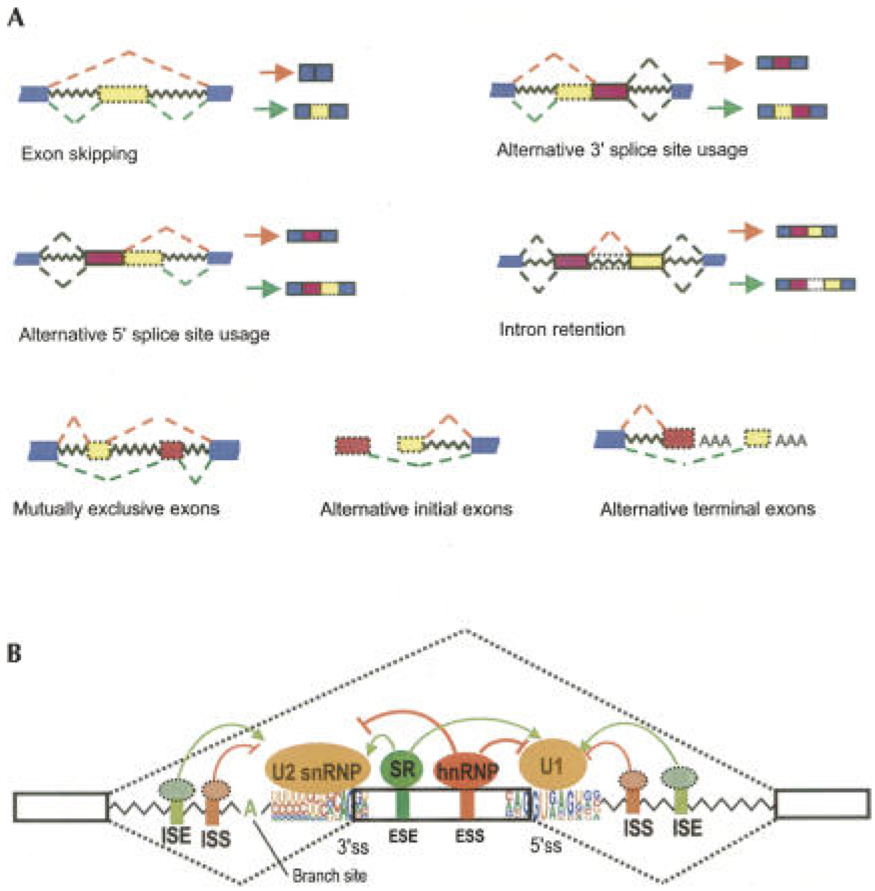


Figure 6. From (Z. Wang and Burge 2008), A. Overview of splicing event types. B. Splicing mechanism.

This process, called alternative splicing, explains the great quantity of proteins present in the cell with respect to the number of genes that code for them. Splicing is finely regulated by RNA binding proteins (RBPs), also known as *trans*-acting factors, that bind on different sites on the transcript, the *cis*-regulatory elements: intron splicing silencers (ISS), intron splicing enhancers (ISE), exon splicing silencers (ESS) and exon splicing enhancers (ESE), and, as these names point, show to the spliceosome machinery (a great complex of proteins that make the action of cut and paste) if sequences they are bound on are to be retained or skipped away (Pozzoli and Sironi 2005; Ule and Blencowe 2019; Z. Wang and Burge 2008). There are also other regulatory sequences necessary for splicing: 5’ and 3’ splice sites, which are located at intron-exon junctions, and the branch point region near the 3’ splice site, where a small spliceosome ribonucleoprotein binds.

RBPs regulate splicing according to position-dependent principles and bind on clusters of short and degenerate sequences (Cereda et al. 2014; Ray et al. 2013).

Often, RBPs work in complexes, to maintain the perfect balance of isoforms into the cell (Papasaikas et al. 2015).

Alternative Splicing is tissue-specific and is influenced by cell health state, phase of life and external conditions and is one of the finest regulated mechanisms of the gene's expression chain. Moreover, splicing is highly interconnected with other cellular processes, like transcription, DNA and RNA modifications (Gehring and Roignant 2021; Ule and Darnell 2006) .

Genetic alterations of transcription factors, RNA binding proteins, and their interactions promotes cancer

Deregulations in both transcription tuning and co-transcriptional processes can lead to several diseases, including cancer.

Mutations in both TFs and their binding sites can bring to aberrant regulation of their targets. Mutations can also disrupt or modify their interaction network (Lambert et al. 2018). Also, their aberrant expression has frequently been linked to tumor progression and invasion (Kajita, McClinic, and Wade 2004; M. Qiu et al. 2014; Robinson et al. 2014; Pelengaris, Khan, and Evan 2002; Xu et al. 2017).

For what concerns alternative splicing, instead, recent studies have revealed multiple ways by which splicing is pathologically altered to promote the initiation and/or maintenance of cancer (Ghigna, Valacca, and Biamonti 2008; Shiraishi et al. 2018; Paschalis et al. 2018; Y. Zhang et al. 2021). Alterations in alternative splicing lead to the creation of aberrant isoforms that could promote several pro-tumorigenic mechanisms, including proliferation, apoptosis, invasion, metabolism boosting, angiogenesis, DNA damage, and also, drug resistance and immune response (Anczuków and Krainer 2016) (Figure 7).

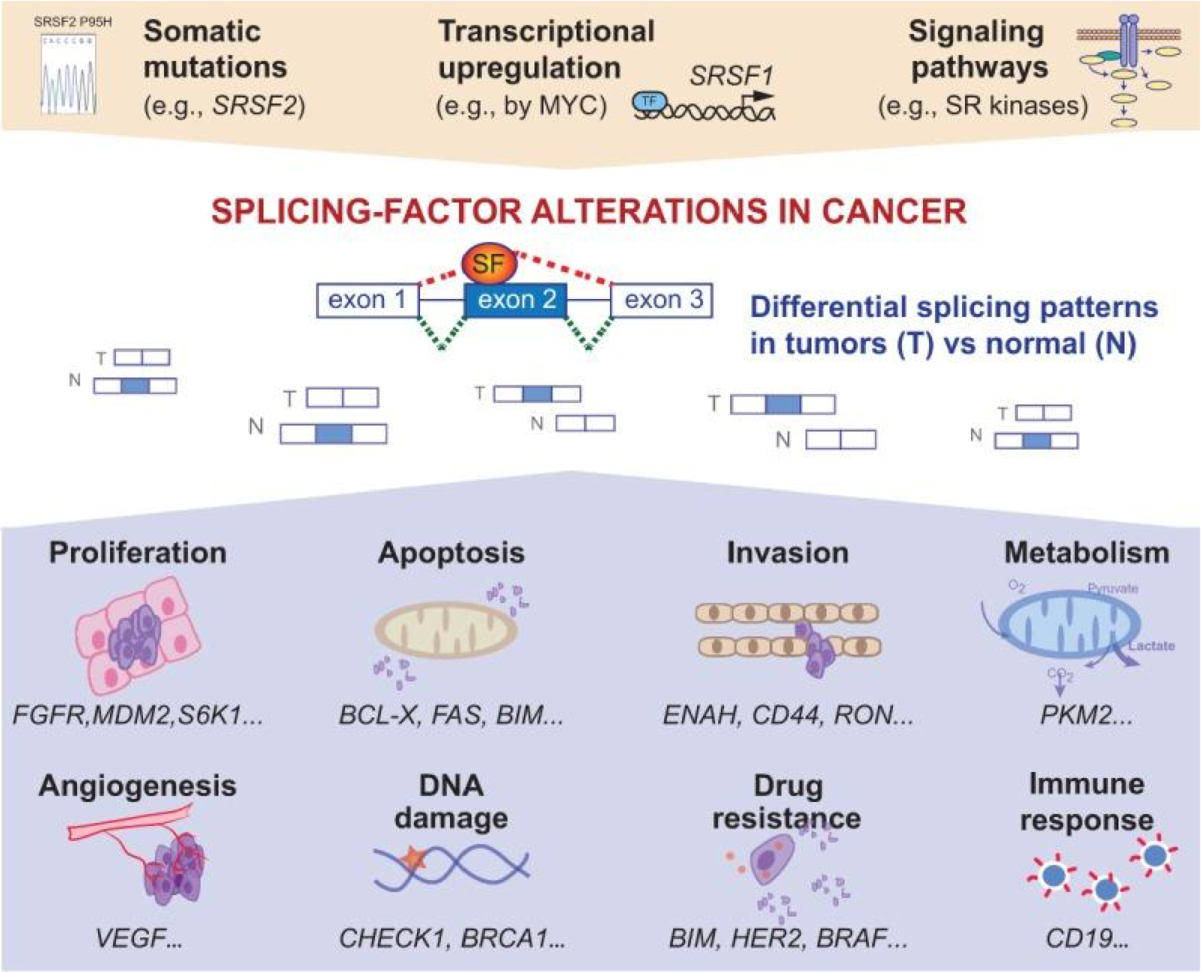


Figure 7. From (Anczuków and Krainer 2016), Splicing-factor alterations in cancer.

Basic mechanisms for normal splicing require *cis*-regulatory elements (splicing enhancers and silencers) and *trans*-acting factors that bind to these elements to promote splicing and/or recruitment of spliceosome (Z. Wang and Burge 2008). Each of these elements may be subjected to various forms of dysregulation during the course of tumorigenesis.

First of all there are mutations within exons and/or introns of the target gene or in genes that code for spliceosome parts or RBPs (Anna and Monika 2018). Mutations in exons or introns can disrupt or create de novo splicing silencers and enhancers (rendering the site unrecognizable by the sequence-specific RNA binding protein that is required for splicing and, for instance, recruiting other RBPs) or create cryptic splice sites (splice sites that are created by mutations). Moreover, they can also disrupt a RNA secondary structure that has a regulatory function.

Other disruptions are copy number variations that can cause a differential expression of genes (S. C. Lee and Abdel-Wahab 2016). Other causes of aberrant constitutive or alternative splicing are quantitative changes of proteins that regulate splicing. Several RBPs have been shown to be over-or under-expressed in cancer (S. C. Lee and Abdel-Wahab 2016). This deregulation can be due either to disrupted signaling pathways or to the aberrant expression regulation by transcription factors. Recently, a genomic study on the aberrant expression of the MYC transcription factor in Eμ-myc transgenic mice lymphomagenesis showed that splicing factors are among those genes upregulated by MYC (Koh et al. 2015). In particular, MYC promotes transcription of the core small nuclear ribonucleoprotein particle (SNRNP) assembly genes, including the RBP Prmt5 that is essential for cell survival and proliferation. Another work enlightened the role of MYC in controlling RBP expression in human glioma (David et al. 2010). Similarly, MYC has been related to RBP expression in triple negative breast cancer (Cieśla et al. 2021) and prostate cancer (Phillips et al. 2020), leading to aberrant splicing. As the direct targeting of MYC has been clinically unsuccessful, targeting its downstream effector pathways seems a valid opportunity. Given the strict relation between MYC and splicing in promoting cancer invasiveness, it has been recently proposed that methods targeting RBPs or their downstream splicing targets may be a potential avenue for treatment (Urbanski et al. 2021).

Aim of the thesis

Although there is much evidence of TF aberrant regulation on splicing factors in cancer, a comprehensive analysis of AS transcriptional regulation is still missing. The project aims at assessing the AS deregulation in cancer, focusing on the regulation of RBP expression by TFs. Our goal is to elucidate the transcriptomic regulation of AS and gain functional insights into the defective AS in cancer. Furthermore, the project will evaluate whether the mechanisms that regulate cancer-promoting AS events are common to different cancer types.

2. Alteration of RBP expression across cancer types

Inequality is the cause of all local movements. There is no rest without equality.

Leonardo da Vinci

To evaluate whether RBP expression is deregulated across all cancer types, processed RNA-seq expression data were retrieved from RNAseqDB (Q. Wang et al. 2018) for both normal and cancer tissues.

RNAseqDB has been developed to integrate RNA-seq data generated from two different sequencing studies while accounting for batch effects. In particular, transcriptomic data from The Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga) and Genotype Tissue Expression project (GTEx) ("The Genotype-Tissue Expression (GTEx) Project" 2013) were realigned to the human genome reference, corrected for batch artifacts, and quantified in terms of gene and isoform expression levels (Figure 8).

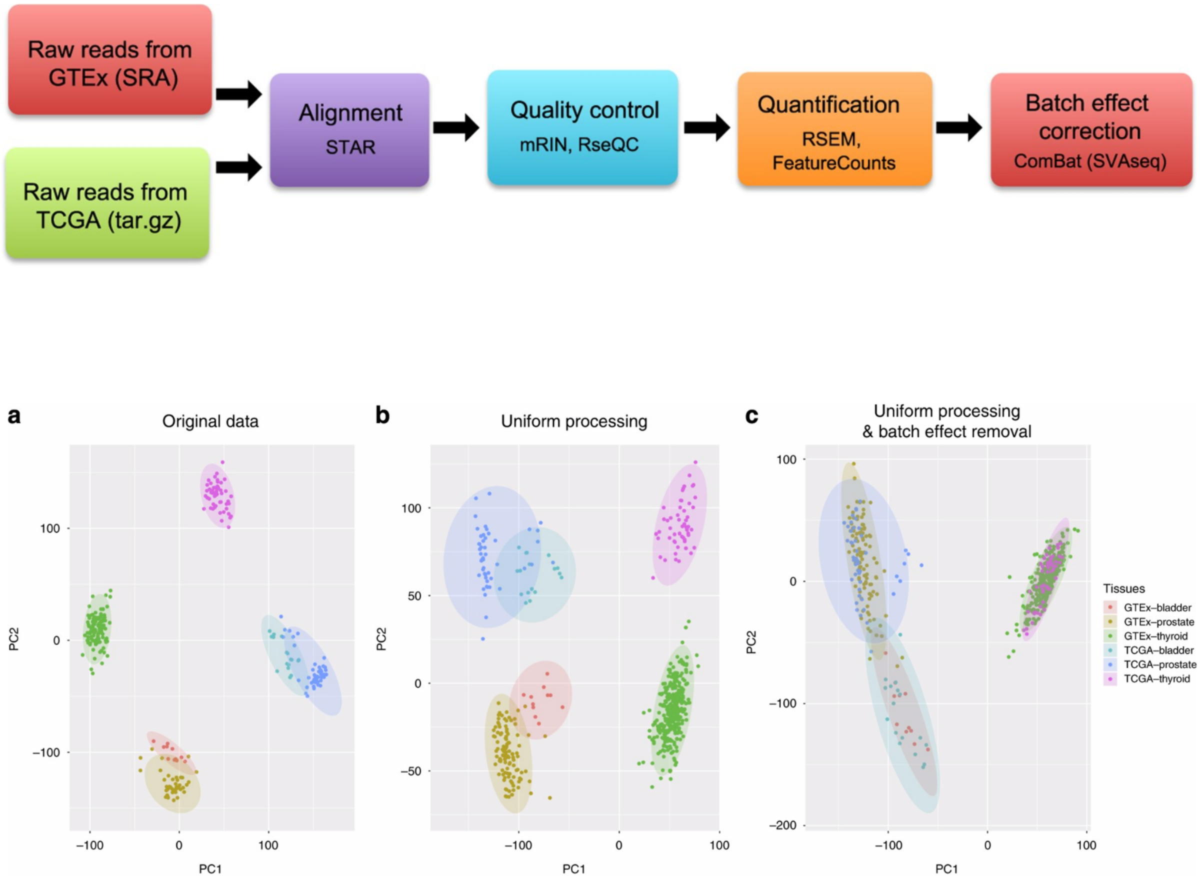


Figure 8. From (Q. Wang et al. 2018), **Upper panel.** RNAseqDB pipeline for uniforming TCGA and GTEx data. **Bottom panel.** PCAs comparing data from TCGA and GTEx before uniforming procedure (a), after uniforming procedure (b) and after uniforming procedure and batch correction (c).

Overall, the RNAseqDB dataset comprises 6,142 tumors of 19 cancer types and 2,322 normal samples from 15 tissues (Figure 9).

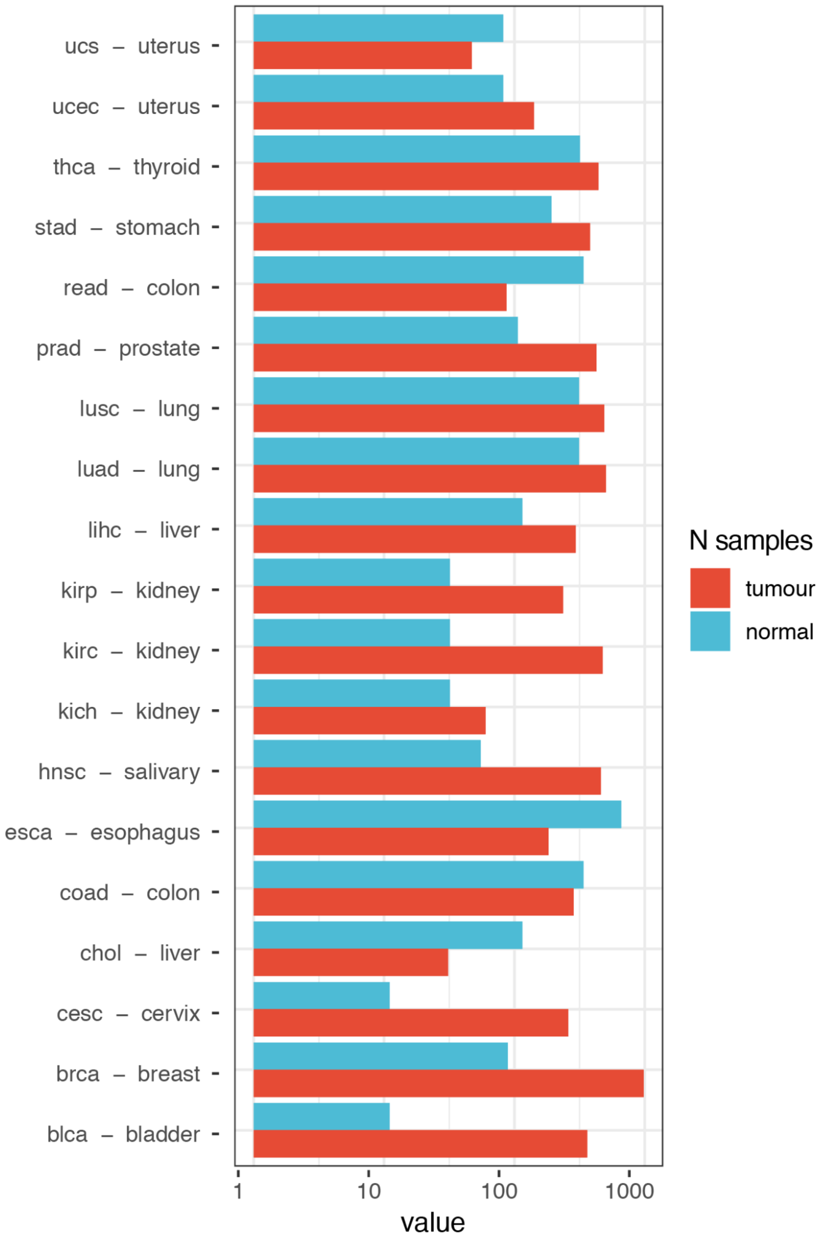


Figure 9. Barplots representing the number of samples available for each tumor type (red) and for the associated normal tissues (light blue).

Gene expression profiles presented a high degree of heterogeneity, especially between samples of the same tumor (Figure 10). Therefore, dissecting the contribution of heterogeneity to gene expression levels is of central relevance to obtain genes that are commonly deregulated in tumors.

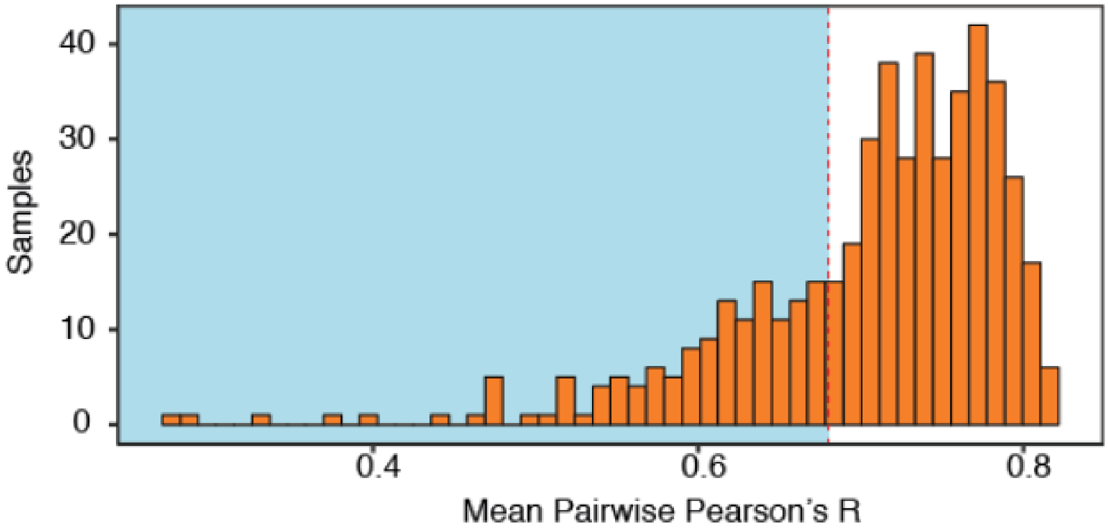


Figure 10. From (Lauria et al. 2020), Histogram showing the distributions of mean Pearson's Correlation of pairwise comparisons, for all PC samples.

In this sense the concept of "pathway" instead of "single gene" alteration has been exploited (A. Subramanian et al. 2005; Lauria et al. 2020). In recent years, several gene set analysis algorithms have been developed (A. Subramanian et al. 2005; Barbie et al. 2009; Hänzelmann, Castelo, and Guinney 2013; E. Lee et al. 2008; Tomfohr, Lu, and Kepler 2005). They can be divided into two groups, the "self-contained" and the "competitive" ones, depending on whether they identify altered gene sets while ignoring or not genes that are outside the gene sets of interest, with the first more powerful than the latter (A. Subramanian et al. 2005; Lauria et al. 2020). Nevertheless, all the available methods were developed to handle microarray data and, also, do not consider inter-sample heterogeneity, so we developed a new method that could handle both RNA-seq data and heterogeneity.

For what concerns RNA-seq distribution, we know from literature that they are characterized by a bimodal behavior, reflecting the presence of two major subpopulations of genes in cells (*i.e.* lowly and highly expressed genes) (Hebenstreit et al. 2011). Moreover, several works exploiting machine learning approaches applied on Big Data, had recently shown that the division of numerical features into a limited number of non-overlapping intervals (*i.e.* discretization) improved the accuracy of the algorithms (Liu et al. 2002; "Machine Learning on Big Data: Opportunities and Challenges" 2017). Discretization had been used in computational biology to explore gene regulatory networks (Demichelis et al. 2006), and, as a pre-processing step, to improve classification accuracy using microarray data (Helman et al. 2004).

On this basis we developed a Gene Set Enrichment Class Analysis (GSECA) (Lauria et al. 2020) algorithm to identify altered gene sets between two cohorts of samples in heterogeneous data. In a few words GSECA implements a finite mixture modeling approach to identify the bimodal distribution of each RNA-seq profile and then applies a discretization on the obtained curve to increase the signal-to noise ratio. Discretized data are then evaluated through a statistical framework to identify altered gene sets.

Gene Set Enrichment Class Analysis (GSECA)

GSECA models the distributions of protein coding gene expression levels of each sample i as a mixture of two Gaussian probability density functions (Hebenstreit et al. 2011) (Figure 11):

$$f\left(x_i\right) = \lambda_1 \phi\left(x_1; \mu_1, \sigma_1\right) + \lambda_2 \phi\left(x_1; \mu_2, \sigma_2\right)$$

where λ is the mixing proportion, μ and σ are the mean and the standard deviation, respectively . This choice is based on the biological evidence that cells express two major populations of genes at high and low levels (McLachlan and Peel 2000). To estimate the parameters of the two components the Expectation-Maximization algorithm is used (McLachlan and Peel 2000; “Royal Statistical Society Publications” n.d.). The algorithm runs iteratively until the maximum likelihood is reached. The mean of the first component (*i.e.* highly expressed genes) is required to be greater than the one of the second component (*i.e.* lowly expressed genes). After the fitting step, GSECA evaluated the probabilities τ_1 and τ_2 of each gene to belong to the two distributions defined by the two components.

A data discretization approach is then applied, fulfilling a biological and a statistical requirement. First, lowly and highly expressed genes must be divided. Second, the discretization of the bimodal distribution of gene expression levels of a sample should provide an adequate distribution of genes among expression classes (ECs) to ensure a similar statistical power for the subsequent tests performed for each class. In particular, for each sample genes are considered as not expressed (NE) or not detected if their expression level in FPKM is lower than 0.01; lowly expressed (LE) if the probability τ_2 of belonging to the second component of the mixture is greater than 0.9; highly expressed (HE) if the probability τ_1 of belonging to the first component is greater than 0.9; or medium expressed (ME) if both the probabilities τ_1 and τ_2 are <0.9. To ensure an adequate distribution of genes among expression classes and retain as much information from the original continuous attribute, HE genes are, in turn, divided into four classes according to the percentiles of the expression level distribution. Then, a statistical framework is applied in order to identify altered gene sets between the two cohorts. For each gene g in each expression class (EC) c the number of samples in which it is or not assigned to the class, n and r respectively, is calculated in the two cohorts:

$$\forall g \text{ and } c,\quad n_{g,c} = \sum_i g \in c;\quad r_{g,c} = \sum_i g \notin c;$$

where i are the samples in the cohorts A and B (Figure 11).

For each gene set G the total number of samples with genes of G that are and are not in each expression class across samples A and B, N and R, respectively, are calculated as:

$$\forall G \text{ and } c,\quad N_{G,c} = \sum_{g\in G} n_{g,c};\ R_{G,c} = \sum_{g\in G} r_{g,c};$$

Then, a Fisher's Exact test is implemented to evaluate the enrichment or depletion of genes of a gene set G in an EC c in A as compared to cohort B. In particular, GSECA tests the null hypothesis that the cumulative proportions of genes of a gene set in each EC across samples are not different between A and B:

$$\forall G \text{ and } c,\quad H_0\colon \left(N_{G,c}; R_{G,c}\right)_A = \left(N_{G,c}; R_{G,c}\right)_B$$

Each of the seven classes are, at this point, characterized by a P-value representative of the expression alterations in gene sets. Given the contingency table defined by N and R for the two cohorts, the algorithm simulates the table under two independent binomial distributions and performs a two-tailed Fisher's Exact Test. All genes of the gene set that are not in the EC under testing are considered, regardless of their class membership. In this way tests are not interdependent. If more than one gene set is considered, Benjamini-Hockberg method is used to correct p-values for multiple testing.

As each class is tested independently from the others, for each gene set a comprehensive P-value that summarizes the alterations across ECs is also evaluated through the Fisher's method. The final Association Score (AS) (Littell and Folks 1971) is obtained as:

$$\Psi =- 2\sum_{c} log(p(c))$$

$$AS(G) = P_{comb} = 1.0 - P\chi^{2}_{2k}(\Psi)$$

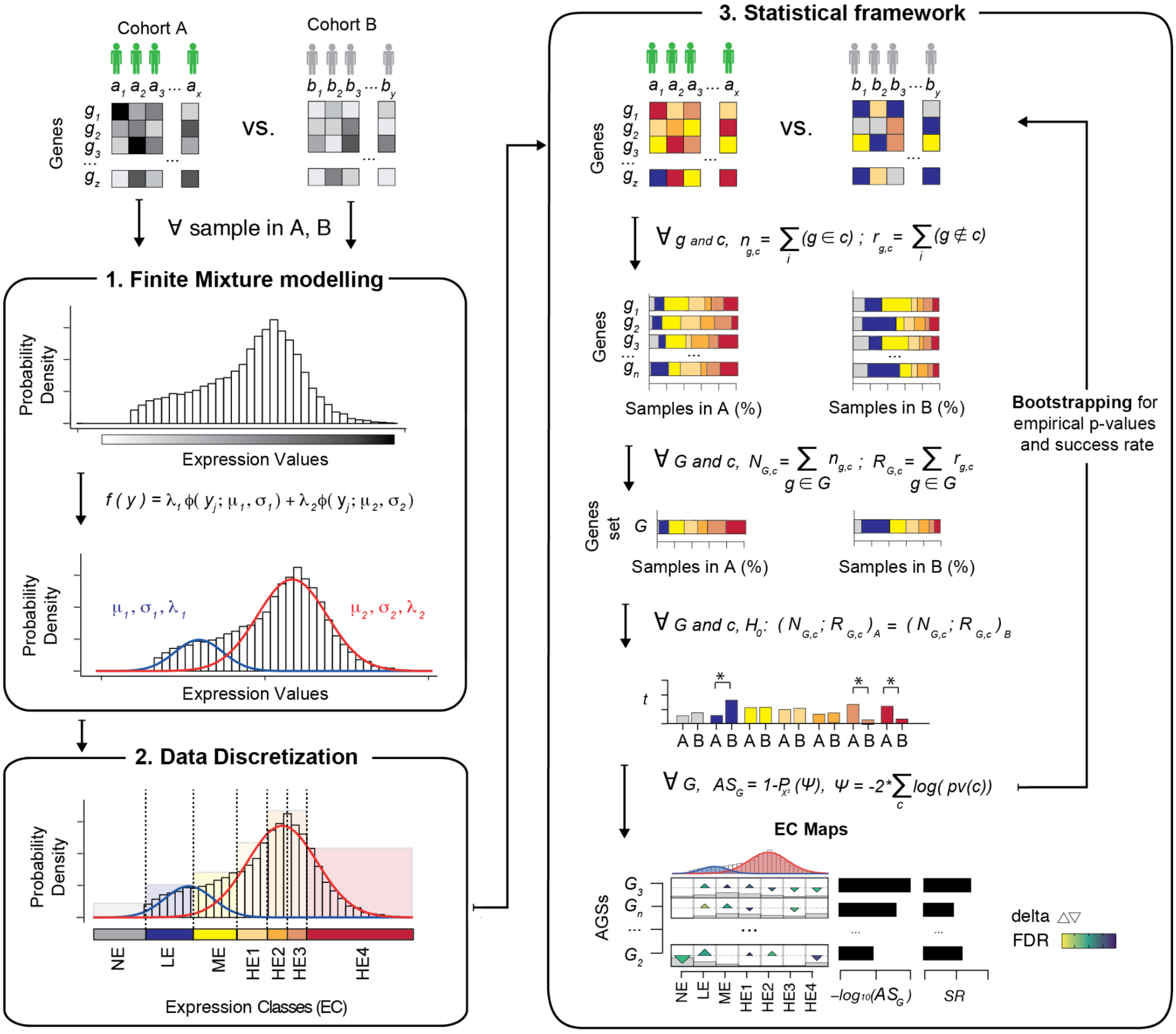
where Ψ is the combined test statistic and χ^{2}_{2k} is a Chi-squared distribution with $2k$ degrees of freedom (k=number of ECs), p is the P-value and c is the expression class.

Also, the significance level of the AS is evaluated by performing a bootstrapping procedure (random sampling with replacement) (Cereda et al. 2014): for 1,000 times sample labels are shuffled and the AS is evaluated for each gene set. At the end the empirical P-value is calculated as:

$$p_{emp}\left(AS_{G}\right) = \frac{1+\left(\sum_{i} AS_{G,i} < AS_{G}\right)}{1+\#iteration}$$

To take into consideration possible sample size differences between the two cohorts, we also implemented a procedure to evaluate the success rate (SR). The algorithm performs a bootstrapping procedure downsampling for 1,000 times the bigger cohort to the dimension of the other and repeating the analysis at each iteration. At the end of these iterations, for each

AS, the SR, or the proportion of significant enrichments (P-value < 0.01, two-tailed Fisher's Exact Test) over the total number of comparisons is calculated as previously described (Gambardella et al. 2017).



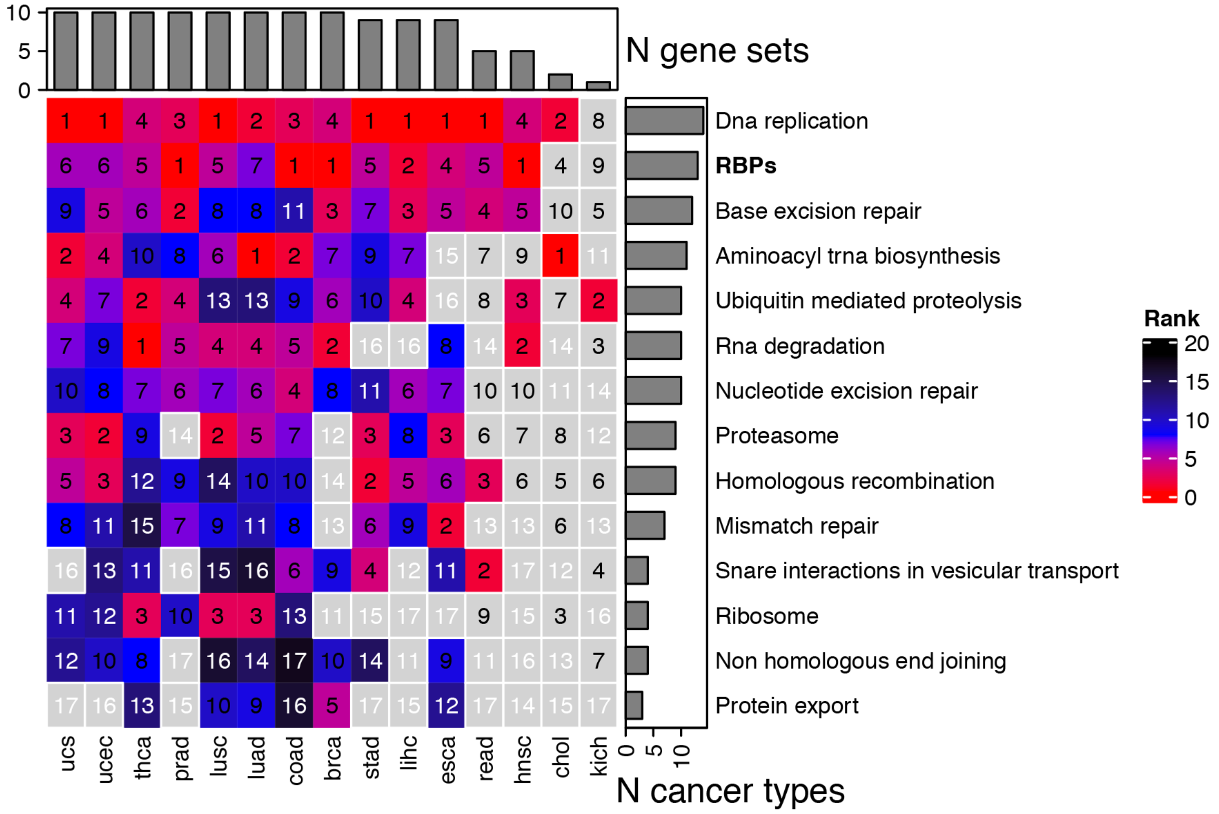
Cohort A
Cohort B
$a_1\, a_2\, a_3 \cdots a_x$
$b_1\, b_2\, b_3 \cdots b_y$
Genes
g_1 g_2 g_3 … g_z
vs.
∀ sample in A, B
1. Finite Mixture modelling
Probability Density
Expression Values
$f(y) = \lambda_1 \phi(y_j; \mu_1, \sigma_1) + \lambda_2 \phi(y_j; \mu_2, \sigma_2)$
$\mu_1, \sigma_1, \lambda_1$
$\mu_2, \sigma_2, \lambda_2$
2. Data Discretization
NE LE ME HE1 HE2 HE3 HE4
Expression Classes (EC)
3. Statistical framework
$\forall$ g and c, $n_{g,c} = \sum_i (g \in c)$; $r_{g,c} = \sum_i (g \notin c)$
g_n
Samples in A (%)
Samples in B (%)
$\forall$ G and c, $N_{G,c} = \sum_{g \in G} n_{g,c}$; $R_{G,c} = \sum_{g \in G} r_{g,c}$
Genes set
G
$\forall$ G and c, H_0: $(N_{G,c}; R_{G,c})_A = (N_{G,c}; R_{G,c})_B$
t
*
A B A B A B A B A B A B A B
$\forall$ G, $AS_G = 1 - P_{\chi^2}(\Psi)$, $\Psi = -2 * \sum_c \log(pv(c))$
Bootstrapping for empirical p-values and success rate
EC Maps
AGSs
G_3 G_n … G_2
NE LE ME HE1 HE2 HE3 HE4
$-\log_{10}(AS_G)$
SR
delta △▽
FDR

Figure 11. From (Lauria et al. 2020), Schematic representation of GSECA algorithm. GSECA requires as input normalized gene expression data of two groups of samples and , and a list of gene sets . The algorithm proceeds through three sequential steps: (*i*) the sample-specific finite mixture modeling of gene expression distribution; (*ii*) the sample-specific discretization of expression values into seven categorical expression classes; and (*iii*) the statistical identification of altered gene sets (AGSs) obtained by comparing the cumulative proportion of genes of a gene set in each EC between the two cohorts using a Fisher's exact test. The expression perturbation is summarized into an association score (AS), corrected with two bootstrapping procedures for false discoveries (empirical p-value) and different sample sizes of the cohorts (success rate, SR). The AGSs are visualized as EC maps. The EC maps display the difference of the cumulative proportion of the genes of a gene set in the seven ECs between the two cohorts as triangles, whose sizes are proportional to such difference. The upper and the lower vertex of the triangles represent enrichment and depletion in cohort A as compared to B, respectively. EC maps depict the proportion *N* of genes in the gene set in each EC as gray bars. GSECA orders AGSs according to their AS, thus obtaining the list of the most altered processes associated with the phenotype of interest.

Detailed analysis and technical considerations about the algorithm are available in the “Appendix A. Gene Set Enrichment Class Analysis”.

GSECA reveals a broad alteration of RBP expression across cancer types

GSECA was applied to the 19 tumor types using as controls their corresponding normal tissues. The chosen gene sets were the 16 Genetic Information Processing (GIP) sets, as they cover all fundamental steps of the cell manipulation of information embedded in the genome, from DNA replication, through transcription, to protein degradation.



	ucs	ucec	thca	prad	lusc	luad	coad	brca	stad	lihc	esca	read	hnsc	chol	kich
Dna replication	1	1	4	3	1	2	3	4	1	1	1	1	4	2	8
RBPs	6	6	5	1	5	7	1	1	5	2	4	5	1	4	9
Base excision repair	9	5	6	2	8	8	11	3	7	3	5	4	5	10	5
Aminoacyl trna biosynthesis	2	4	10	8	6	1	2	7	9	7	15	7	9	1	11
Ubiquitin mediated proteolysis	4	7	2	4	13	13	9	6	10	4	16	8	3	7	2
Rna degradation	7	9	1	5	4	4	5	2	16	16	8	14	2	14	3
Nucleotide excision repair	10	8	7	6	7	6	4	8	11	6	7	10	10	11	14
Proteasome	3	2	9	14	2	5	7	12	3	8	3	6	7	8	12
Homologous recombination	5	3	12	9	14	10	10	14	2	5	6	3	6	5	6
Mismatch repair	8	11	15	7	9	11	8	13	6	9	2	13	13	6	13
Snare interactions in vesicular transport	16	13	11	16	15	16	6	9	4	12	11	2	17	12	4
Ribosome	11	12	3	10	3	3	13	11	15	17	17	9	15	3	16
Non homologous end joining	12	10	8	17	16	14	17	10	14	11	9	11	16	13	7
Protein export	17	16	13	15	10	9	16	5	17	15	12	17	14	15	17

Figure 12. Heatmap showing results of pancancer GSECA analysis. For each tumor type gene sets are ranked according to their combined p-value (from the most significant). Only significant (combined p-value≤10^-3, empirical p-value≤0.3, AS=1 and rank≤10) are colored according to their rank, while not significant associations are depicted in gray. Barplots on the top indicate the number of significant gene sets for a specific tumor type, while barplots on the right represent the number of tumor types in which a gene set has been identified as significant.

Results were ranked according to combined p-value, empirical p-value and association rate and then four different filters were applied to consider a result as significant (combined p-value≤10^-3, empirical p-value≤0.3, AS=1 and rank≤10) (Figure 12). RBP gene set resulted as significantly altered in 13 out of the remaining 15 tumors. Moreover, looking at rankings, the RBP gene set was the second most deregulated pathway after DNA replication, which was altered in 14 cancer types.

We next assess the changes of expression within the RBP gene set. In particular, we analyzed the corresponding EC maps, which compares the fraction of RBPs in each expression class between the tumor and normal condition (Figure 13). We found that the number of tumors expressing splicing factors at the highest levels (i.e. HE4 class) was depleted in all cancer types (lower vertex of the triangles). Conversely, tumors that express RBPs at intermediate levels (i.e. from ME to H3 classes) were generally enriched across all cancer types. Major changes (bigger triangles) were found in higher classes of expression, whereas the other classes presented more heterogeneous patterns across tumors. These changes and their extent in each cancer type were summarized by the GSECA association score. In particular, the AS was the highest for Colon Adenocarcinoma (COAD) and lung cancers, namely Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). Changes in expression of RBPs have been identified in both lung and colon cancers, as well as chances in splicing patterns, suggesting splicing as a possible therapeutic target (Coomer et al. 2019; Y. Chen et al. 2021).

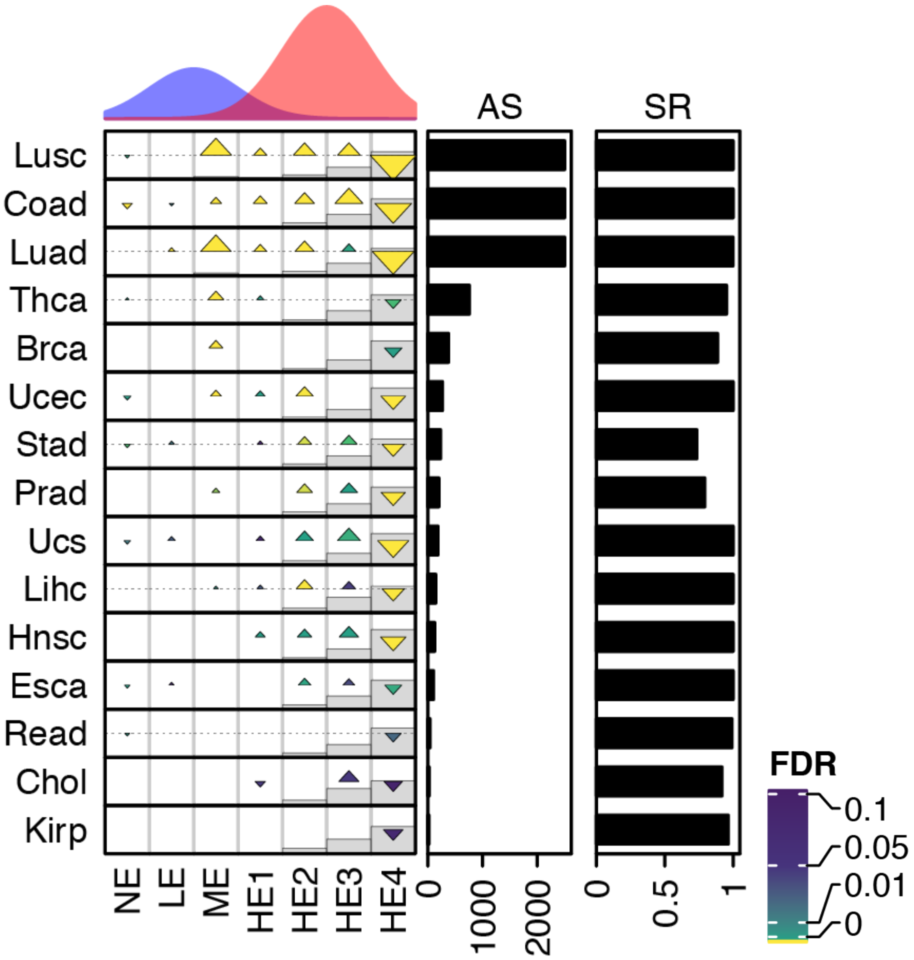


Figure 13. EC map displaying the difference of the cumulative proportion of the genes of RBP gene set in the seven ECs across 15 cancer types. The EC map displays the difference of the cumulative proportion of the genes of the RBP gene set in the seven ECs between the two cohorts as triangles, whose sizes are proportional to such difference. The upper and the lower vertex of the triangles represent enrichment and depletion in cohort A as compared to B, respectively. EC maps depict the proportion *N* of genes in the gene set in each EC as gray bars. Barplots on the right represent the association score (AS) and the success rate (SR) values.

Overall, our results showed that RBP expression is deregulated in almost all cancer types. Even if RBP expression has been frequently found as altered in cancer (Coomer et al. 2019; Koedoot et al. 2019), a comprehensive and pan-cancer analysis of RBP expression deregulation was still missing. The evaluation of deregulation seemed independent from the number of patients available for a certain cancer type.

3. Evaluation of transcription factors' action on RBP expression at a cancer specific level

The soul is the same in all living creatures, although the body of each is different.

Hippocrates

To evaluate whether some transcription factor is the major responsible for the expression of RBPs two approaches were used, exploiting both direct binding level and the expression one.

Identification of active TF binding sites

To establish the features of transcriptional control of RBPs across cancer types, we integrated data on chromatin immunoprecipitation sequencing (ChIP-seq) experiments with assay for Transposase-Accessible Chromatin sequencing (ATAC-seq). In doing so, we defined TF binding sites in regions that are recurrently actively-transcribed across tumors of the same cancer type as "active" binding sites. We then developed a framework to assess the over-representation of RBPs with active binding sites of distinct TFs within cancer-cell-line-specific chromatin-accessible promoter regions.

Specifically, ChIP-Seq experiment significant peaks (*i.e.* p-value$\leq 10^{-5}$) for 485 TFs in 207 cell lines in 31 tissues were retrieved from Remap 2018 (Chèneby et al. 2018) (read also "Appendix B. Insights into literature", A.). To select regulatory regions that are related to sites of active transcription, accessible DNA elements that were defined as reproducible across Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) experiments for 23 cancer types were retrieved from the Genomic Data Commons (GDC) Portal (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG) (Corces et al. 2018) (read also "Appendix B. Insights into literature", B.). Coordinates of 20,298 protein-coding genes were retrieved from GENCODE GRCh37 version 28 (Frankish et al. 2019) and promoter regions were defined as 2,000 base pairs upstream and downstream the transcription start sites of each gene.

Cancer cell lines were associated with the corresponding cancer type, resulting in 107 cell lines associated with 17 tumor types (Figure 14).

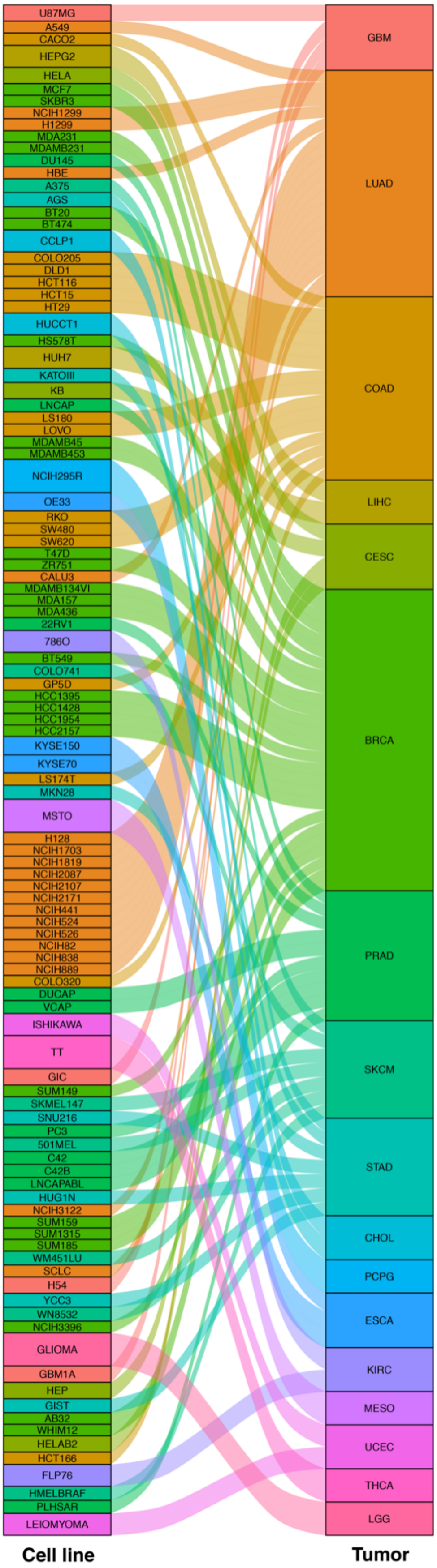


Figure 14. Sankey plot showing associations between cell lines and tumors.

In total, we collected data for 273 TFs. The number of TFs with ChIP-seq data available were greatly variable among cancer types (Figure 15).

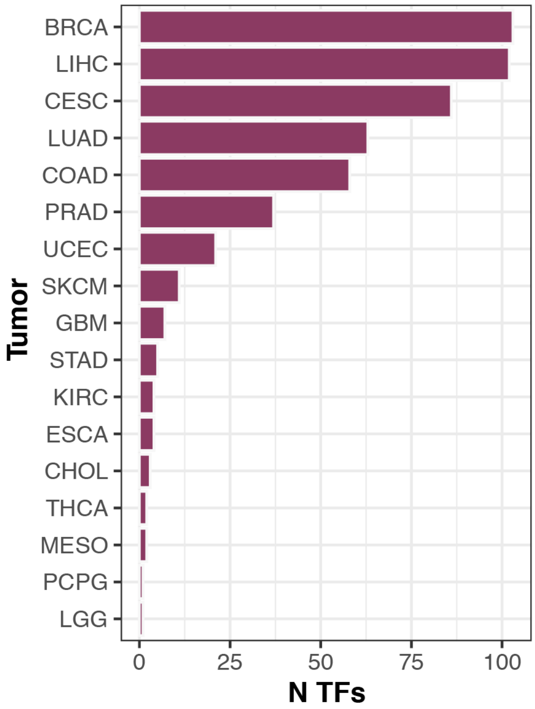


Figure 15. Barplot showing the number of TFs for which we collected ChIP-seq data for each tumor type.

For each cancer type, cell line, and TF, ChIP-Seq peaks of the same TF in a fixed cell line were merged and intersected with regions of corresponding open chromatin to identify the corresponding active bindings. Only overlapping peaks were retained and called, hereafter, TF active binding sites in a specific cell line. In the same manner promoter regions were intersected with accessible elements of the nine tumor types, in order to define cancer-specific transcribed genes. Next, genes with active TF binding sites in the cognate open-chromatin promoter were retained. Finally, to identify TFs that control RBP expression direct transcriptional control, we assessed the over-representation of our set of 148 RBPs with active binding of TFs as compared to the rest of genes with evidence of the same regulation. In particular, an upper-tailed Fisher's Exact test was performed for each TF in each cell line testing the proportion of RBPs with ative bindings as compared to the rest of human genes with active bindings. For each cell line, p-values were corrected for multiple testing using Benjamini-Hochberg method and only those with false discovery rate (FDR)≤0.1 were considered as significant (Figure 16). We focused on TFs that were annotated as cancer drivers according to the Network of Cancer Genes v.6 (Repana et al. 2019) (read also

“Appendix B. Insights into literature”, C.) and cell lines belonging to tumor types for which we had also RNAseqDB data.

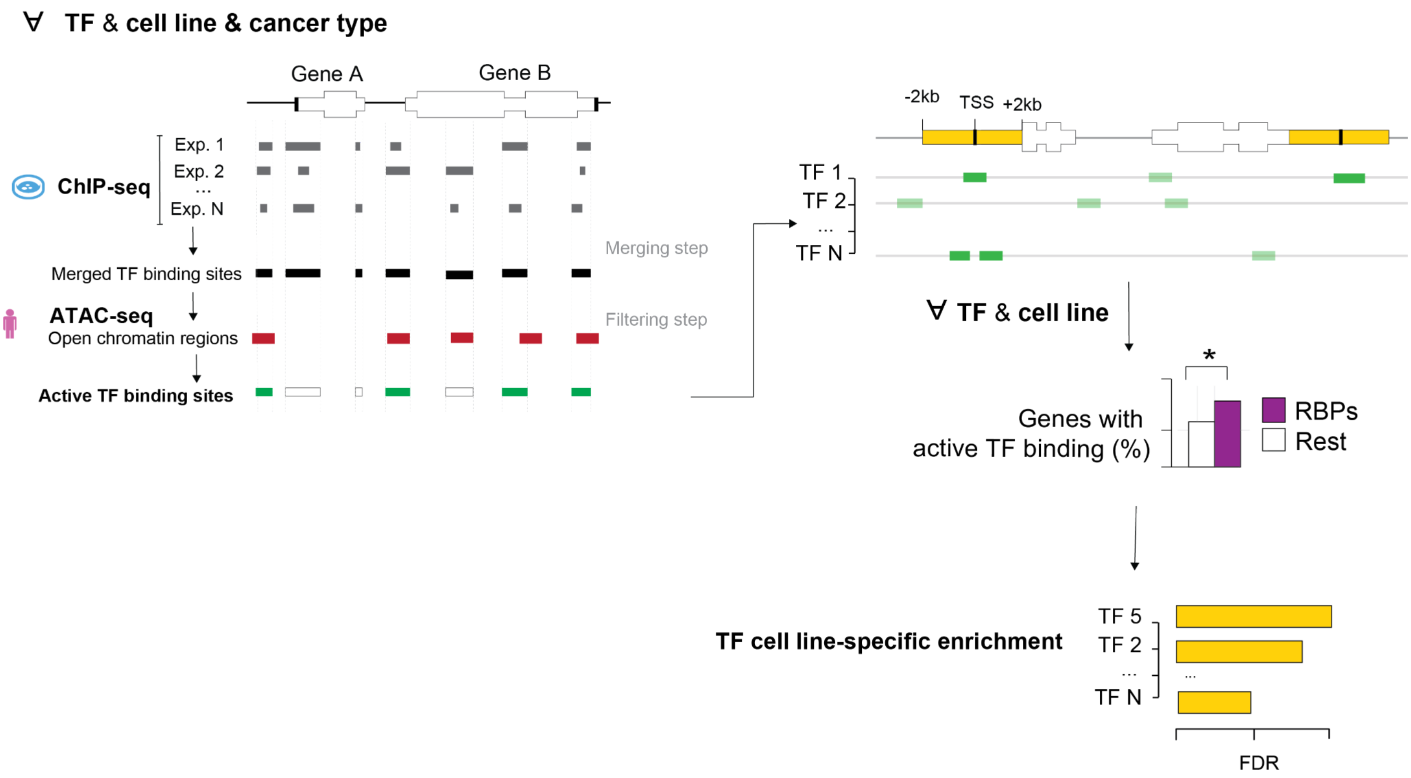


Figure 16. Schematic representation of the pipeline used to assess the overrepresentation of RBPs with TFs active binding sites within cell line-specific chromatin-accessible promoters.

Although each cell line presented its own pattern of regulation, some are the transcription factors that emerged for their enriched association to RBPs (Figure 17).

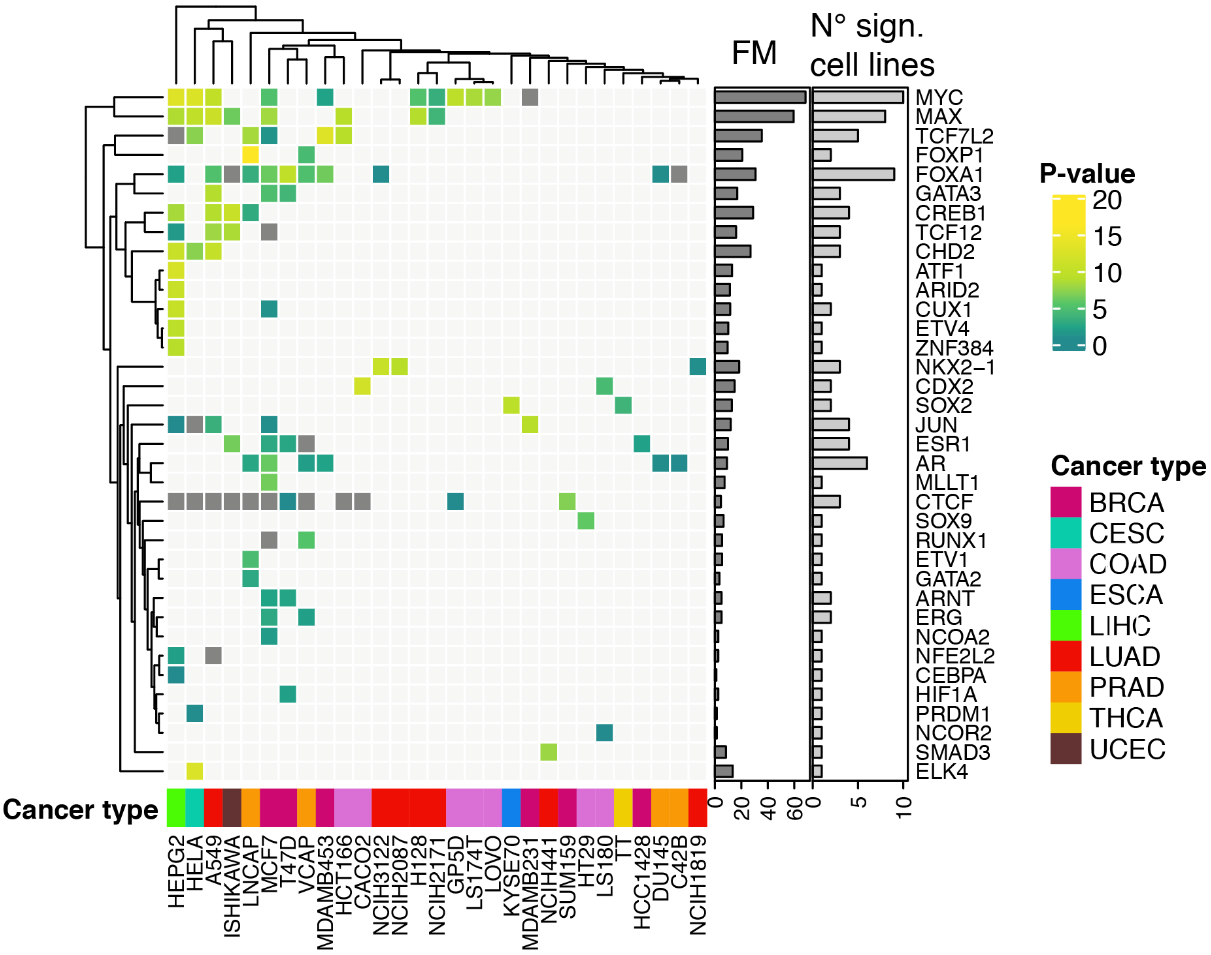


Figure 17. Heatmap showing the $\log_{10}$(adjusted p-value, FDR) for each transcription factor in each cell line. Squares are depicted in gray if FDR was not significant (FDR>0.1). Cream color squares represent the absence of a TF in a specific cell line. Bottom stripe indicates the cell cancer type associated with the cell line. Barplots on the right represent the cancer-specific P-value obtained using the Fisher's Method (log10) and the number of cell lines in which the TF is significant. TFs or cell lines that did not present any enrichment were excluded from the graph.

We found that MYC, MAX and FOXA1 were enriched in many cell lines with generally low FDR. MYC and MAX were enriched in almost the same cell lines, supporting the evidence showing that these TFs works together in a heterodimer (Arsura et al. 1995; Bissonnette et al. 1994; Walhout et al. 1997; Nair and Burley 2003; Grandori et al. 1996). Interestingly, in cell lines where they are both present (coloured squares), MYC and FOXA1 were always significant, suggesting a possible relation between the two.

Identification of cancer-specific TF regulators of RPB expression

To define the cancer type specific RBP major regulators for each cancer type and TF, we combine cell-line specific TF p-values using the following procedure. A cumulative enrichment score, named hereafter as “Score”, was also calculated combining cancer-type specific p-values with the Fisher Method (FM, (Fisher 1992)). The “Score” was then rescaled as follows:

$$Score = log_{10}(Score) * N_{cancer}$$

where N_{cancer} is the number of cancers in which TF resulted as significantly associated with RBPs.

We found that MYC and MAX were the most significant ones, followed by FOXA1, TCF7L2 and CREB1 (Figure 18). Interestingly, MYC, MAX and FOXA1 were not only enriched in several cell lines but also that these cell lines belonged to different tumors.

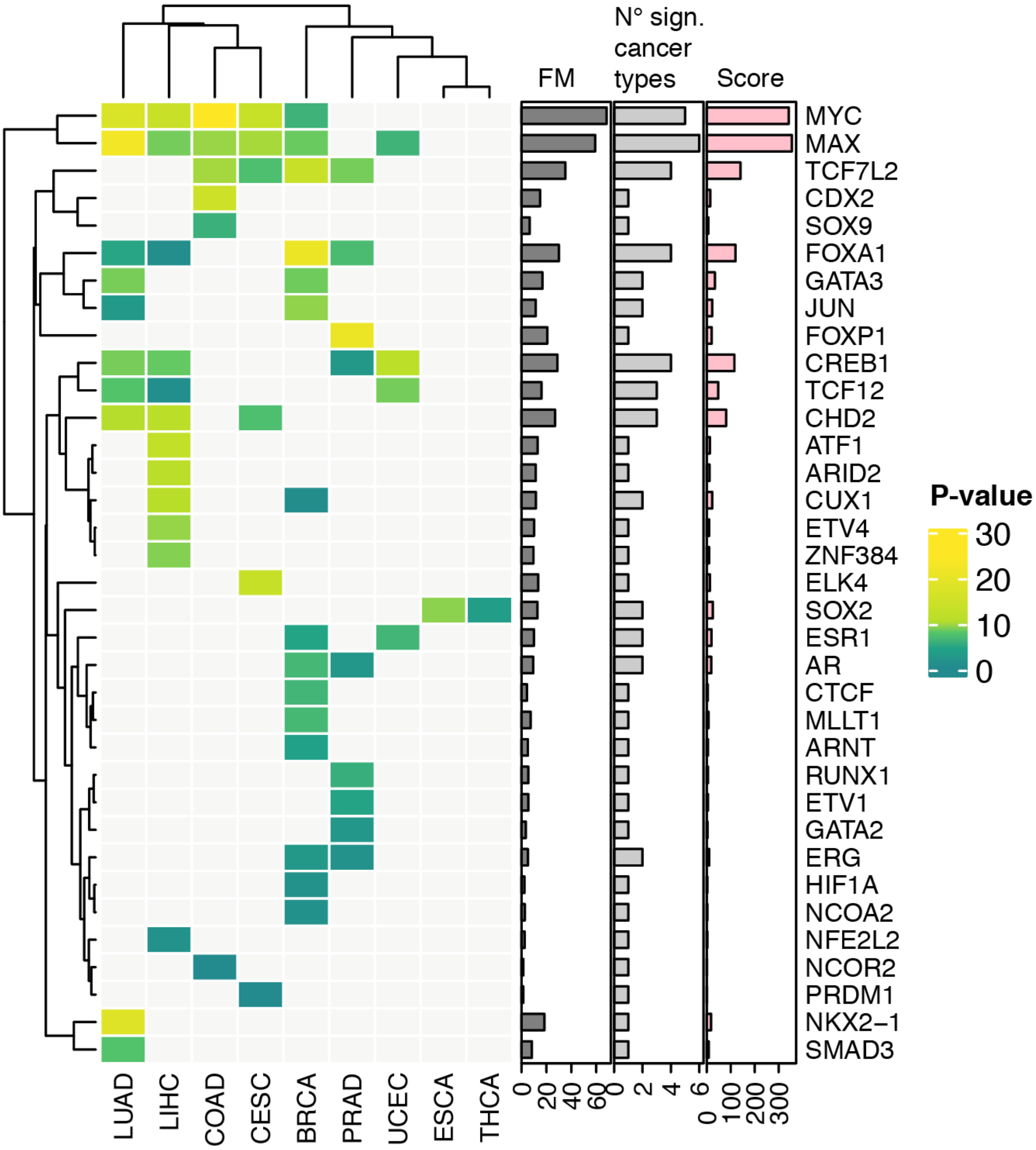


Figure 18. Heatmap showing the log10(combined P-Value (Fisher's Method)) for each transcription factor in each tumor type. Squares are depicted in cream color if the combined P-Value (Fisher's Method, FM) was not significant (FM>0.05) or if the TF is absent in a specific tumor type. Barplots on the right represent the TF-specific P-value obtained using the Fisher's Method, FM, (log10), the number of tumor types in which the TF is significant and the "Score", obtained by multiplying quantities of the two previous barplots.

Modeling RBP expression as a function of TF profiles

We next assessed the putative TF regulation of RBP expression emplying an orthogonal approach. In particular, rather than considering evidence at DNA-level, we exploited RNA-sequencing data looking for proofs of regulation at RNA-level.

Processed RNA-seq expression data were retrieved from RNAseqDB (Q. Wang et al. 2018) for 8,464 samples, divided into 6,142 cancer and 2,322 normal ones. Nineteen tumor types were covered (Figure 9).

To identify transcription factors whose expression has the highest association with RBP expression a generalized linear model (GLM) approach was used. For each cancer type, GLM was fitted to the cumulative expression of RBPs using TFs as regressors. For each tumor type were selected the cancer-related TFs for which we retrieved ChIP-seq data. To increase the statistical power of prediction, corresponding normal tissues were also included. In particular, there were only selected tumor types and TFs for which the analysis of direct bindings were performed. The relative importance (RI) of each TF in the GLM was calculated using the averaging over ordering method ("Introduction to Bivariate and Multivariate Analysis" n.d.) and a TF was considered as putative regulator of RBP expression if its RI was positive (indication of positive regulation) and the associated p-value significant ($p\text{-value} \leq 0.05$).

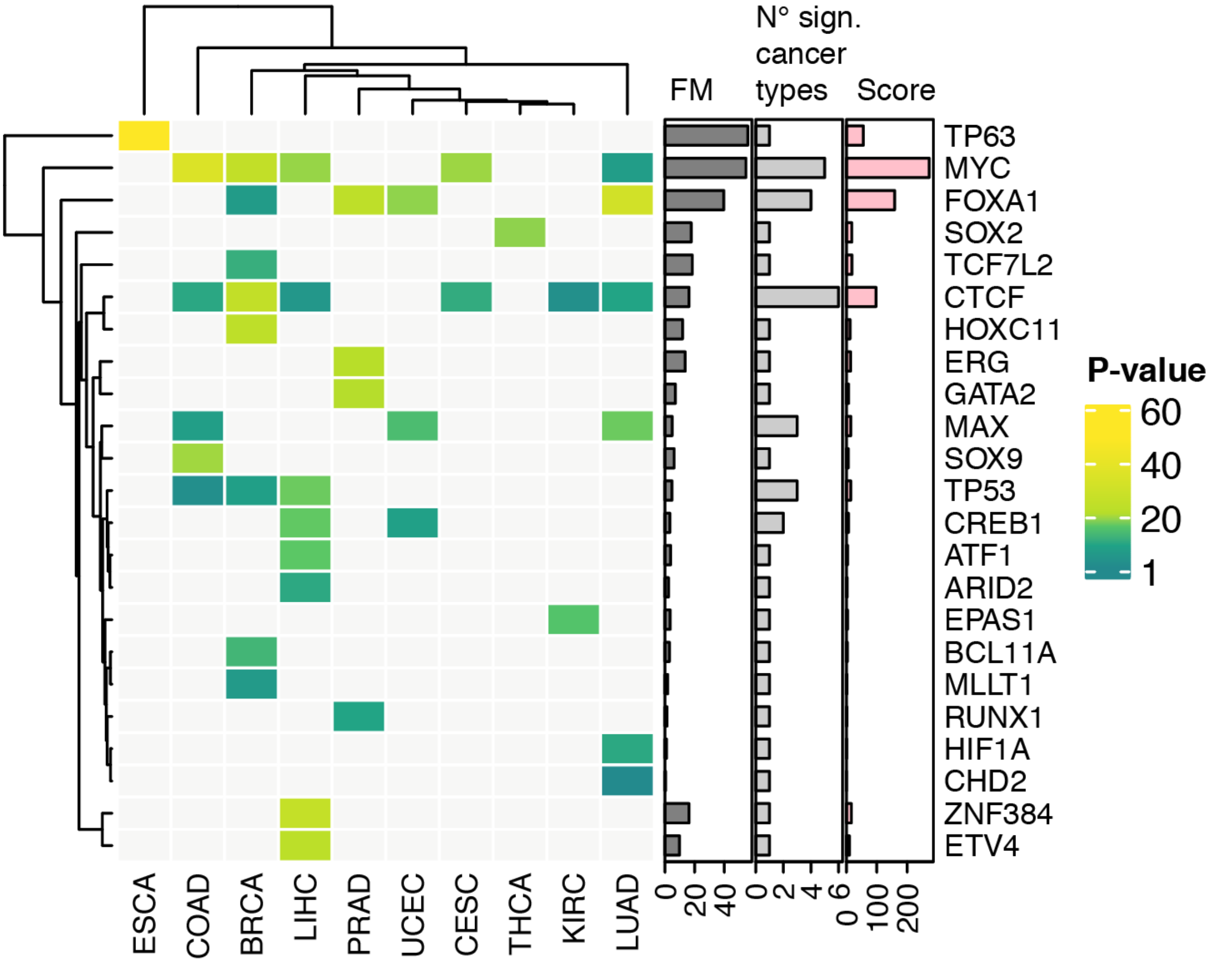


Figure 19. Heatmap showing the log10(P-Value) from the GLM model for each transcription factor in each tumor type. Squares are depicted in cream color if the P-Value was not significant (P-Value>0.05 or negative coefficient). Barplots on the right represent the TF-specific P-value obtained using the Fisher's Method, FM, (log10), the number of tumor types in which the TF is significant and the "Score", obtained by multiplying quantities of the two previous barplots.

We found that MYC and FOXA1 were the TF with both high numbers of cancer types with enrichment and generally low p-values (Figure 19). On the contrary, TP63 presented the lowest p-value but was only enriched in esophageal carcinoma, while CTCF presented a wide number of significant cell lines, but relatively high p-values.

Double-hits of regulations at DNA and RNA levels reveal MYC and FOXA1 as master regulators of RBP expression

For each cancer type, we labeled as putative RBP regulators the TFs that were selected both from the active binding and the expression analyses (Figure 20).

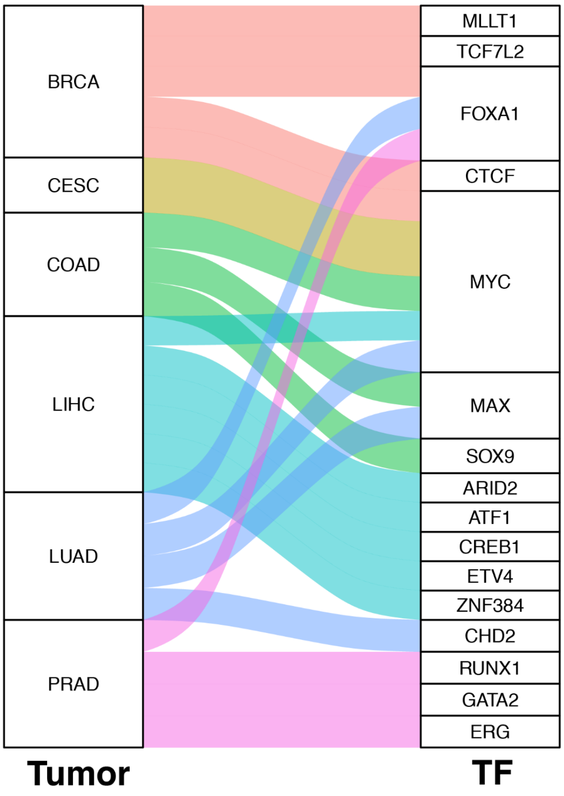


Figure 20. Sankey plot showing candidate TFs for each tumor type.

Although each cancer type presented its own pattern, MYC and FOXA1 were the most powerful candidates for RBP regulation across cancer types showing enrichments at both DNA and RNA levels (Figure 21) .

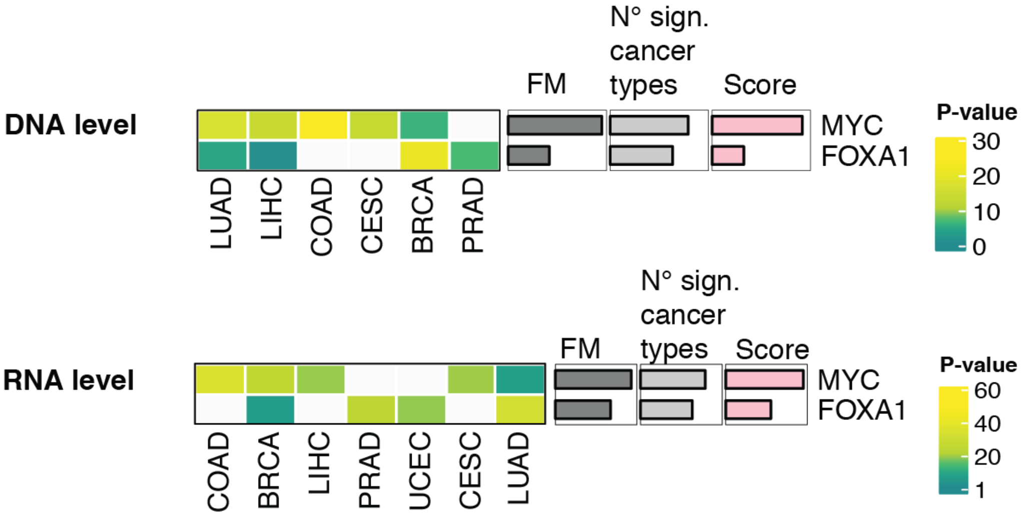


Figure 21. Heatmaps showing the focus on FOXA1 and MYC results at a DNA and RNA level.

Overall, combining ChIP-seq, ATAC-seq and RNA-seq data we showed that some transcription factors harbor both an enrichment of direct bindings on promoters of RBPs and a positive and significant correlation with their expression. In particular, MYC and FOXA1 emerged from both analyses as RBP regulators across cancer types.

4. FOXA1 regulation on RBP expression in prostate cancer

Great empires are not maintained by timidity.

Tacitus

The pioneer transcription factor FOXA1

Since the relation between MYC and splicing in cancer has been widely studied, as we explained in Chapter 1, we focused on FOXA1 as a regulator of RBP expression.

FOXA1 is a pioneer transcription factor and epigenetic modifier (Lupien et al. 2008), primarily involved in developmental processes (Bernardo and Keri 2012; Gasser et al. 2016). In cancer, FOXA1 has been widely studied for its recurrent somatic mutations and variants of its cis-regulatory elements (Parolia et al. 2019; N. Shah and Brown 2019; Teng et al. 2021). It is particularly known for its tumorigenic effect in hormone-related cancers, like endometrium (M. Qiu et al. 2014), breast (Fu et al. 2019; Seachrist, Anstine, and Keri 2021), and prostate cancer (Robinson and Carroll 2012; Teng et al. 2021; Gerhardt et al. 2012; Rhie et al. 2019). It has been shown that FOXA1 binds on AR in promoting AR target transcription (Robinson et al. 2014; Sahu et al. 2011; Jin et al. 2014; Teng et al. 2021; Jones et al. 2015). Nevertheless, in neuroendocrine prostate cancer, where AR is not expressed, FOXA1 promotes proliferation (Baca et al. 2021). The link between FOXA1 expression and cancer proliferation has often been reaffirmed (Imamura et al. 2012; M. Qiu et al. 2014). Moreover, overexpression of the pioneer transcription factor FOXA1 is a frequent event in primary and metastatic prostate cancer (Parolia et al. 2019; Jin et al. 2013) (Figure 22).

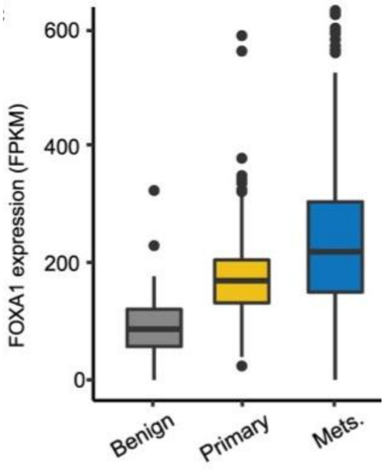


Figure 22. From (Parolia et al. 2019), Boxplot showing FOXA1 expression in normal prostate, pri-PC tissue and metastatic prostate cancer.

As FOXA1 is a known oncogene of prostate cancer, we sougth to investigate its role of RBP regulator in this cancer type, where aberrant AS has recently been related to its progression.

AS dysregulation leads to prostate cancer aggressiveness

Recently, it has been shown that the dysregulation of expression of splicing-related genes is related to prostate cancer aggressiveness and to the increased resistance towards drugs (Jiménez-Vacas et al. 2020). Also, the overexpression of some RBPs has been related to worst survival (Jiménez-Vacas et al. 2020) (Figure 23).

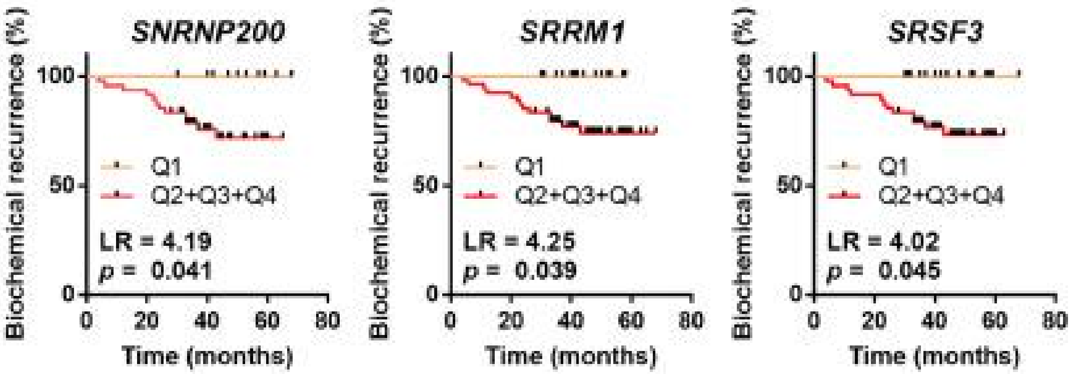


Figure 23. Association between *SNRNP200, SRRM1* and *SRSF3* expression levels and biochemical PCa recurrence in 67 samples from FFPE cohort (samples from patients who underwent adjuvant radiotherapy were not included), calculated by Log Rank analysis (LR). Extracted from (Jiménez-Vacas et al. 2020)

Moreover, dysregulation in intron retention patterns has been proposed as an hallmark of cancer (D. Zhang et al. 2020).

Evidences, both in literature and from our analyses, show that:

- FOXA1 is a major regulator of RBPs;
- FOXA1 expression is often deregulated in prostate cancer, particularly in metastatic one;
- AS is deregulated in almost all cancer types;
- RBP expression has been related to survival in prostate cancer and altered AS has been proposed as an hallmark;

Given that, we explored the splicing landscape in prostate cancer, with a particular focus on FOXA1 transcriptional regulation over RBP expression.

Transcriptional regulation of RBPs in primary and metastatic prostate cancer by FOXA1, ERG, GATA2, and RUNX1

To assess the influence individually exerted by FOXA1, ERG, GATA2 and RUNX1 (Figure 20) to the dysregulation of AS in prostate cancer (PC), we measured the contribution of their expression to the overall transcription of 148 RBPs. To do so, we used available RNA sequencing (RNA-seq) repositories of 409 primary (Network, Cancer, 2015) and 118 metastatic castration-resistant (Robinson et al., 2015). Data were obtained from The Cancer Genome Atlas (TCGA) Data Matrix portal (Level 3, https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) and cBioPortal (Cerami et al. 2012; Y. Chen et al. 2013) websites for 409 primary PC (pri-PC) and 118 metastatic castration-resistance PC (mCRPC) samples, respectively. The number of transcripts per million reads (TPM) was measured starting from the scaled estimate expression values provided for 20,531 genes (Cereda et al. 2016). For the mCRPC dataset, reads per kilobase of transcript per million mapped reads (RPKM) values were converted into TPM. For our quantitative analysis, we fitted RBP cumulative expression as a function of transcription factor expression levels using a generalized linear regression and measured their relative contribution in the model as described above.

We found that, out of the four TFs, FOXA1 was the strongest positive predictor of RBP cumulative expression in primary and metastatic PC (Figure 24).

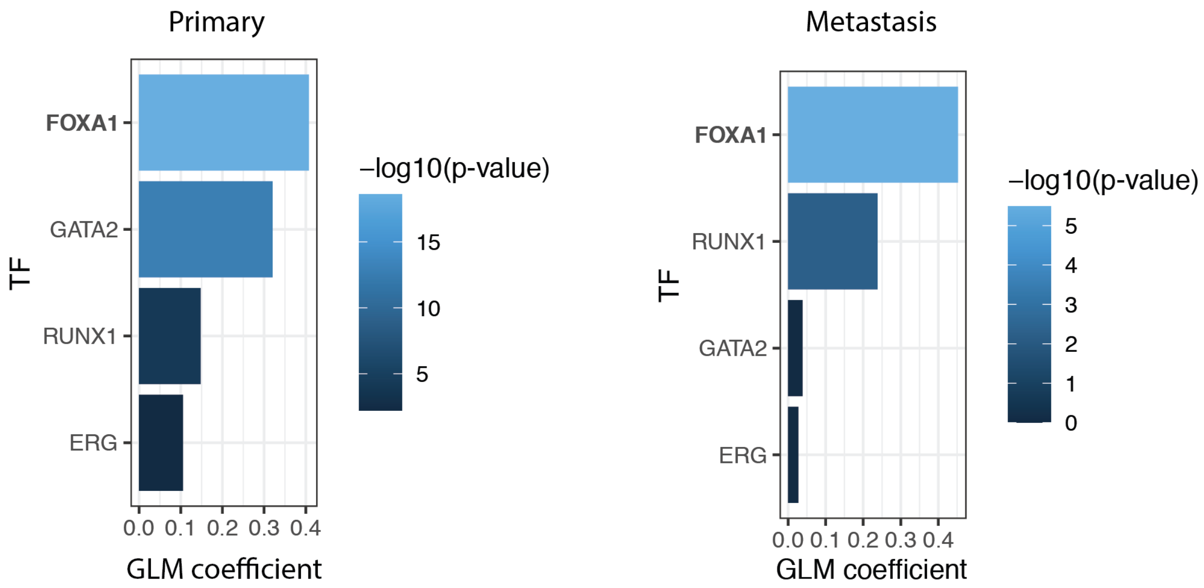


Figure 24. GLM coefficient of each TF by the GLM for pri-PC (left) and mCRPC (right) samples.

FOXA1-regulated RBPs in primary and metastatic PC

We next sought to identify RBPs that were regulated by FOXA1. Since RBPs are highly expressed in the cell (de la Grange et al., 2010; Sebestyén et al., 2016), canonical parametric methods for differential expression analysis can fail to detect statistically significant changes in presence of large gene counts and subtle differences between cohorts (Li and Tibshirani, 2013). Recent studies have shown that nonparametric differential expression analysis approaches are more robust than parametric models to handle this scenario (Shi et al., 2015; Zhu et al., 2019). In this view, we combined parametric and non-parametric analysis to identify FOXA1-regulated RBPs in PC. To do so, we compare the transcriptional targets of the 25% of tumors with the highest expression (HE) of FOXA1 against the remaining ones (REST, Figure 25). In particular, differentially expressed RBPs were identified by comparing their TPM distributions between FOXA1 HE and REST samples with a two-tailed Kolmogorov-Smirnov test. P-values were corrected for multiple tests using the false discovery rate (FDR) by Benjamini–Hochberg method. To estimate the empirical p-value of each comparison, a Monte Carlo procedure was implemented. For 10,000 iterations, FOXA1 HE and REST samples were randomly selected and, for each RBP, the TPM distributions were compared using a two-tailed Kolmogorov-Smirnov test. For each RBP, the empirical p-value was measured as the proportion of tests with p-value smaller than the corresponding observed one over the total number of iterations. Concomitantly, canonical parametric differential expression was performed between FOXA1 HE and REST samples using the R packages 'DESeq2' and 'EdgeR' (Love, Huber, and Anders 2014; Robinson, McCarthy, and Smyth 2010) for the TCGA dataset, for which raw sequencing counts were available. Genes with read count equal to zero across all samples were removed. RBPs with FDR≤0.01, empirical p-value≤0.01, DESeq2 or EdgeR absolute $\log_2$ Fold Change (FC) ≥0.2 and adjusted p-value≤0.01 were considered as differentially expressed in FOXA1 HE samples as compared to REST. For the metastatic dataset, RBPs with FDR≤0.01, emp-pv≤0.01, and an absolute $\log_2$(FC) of median TPM ≥0.2 were considered as altered.

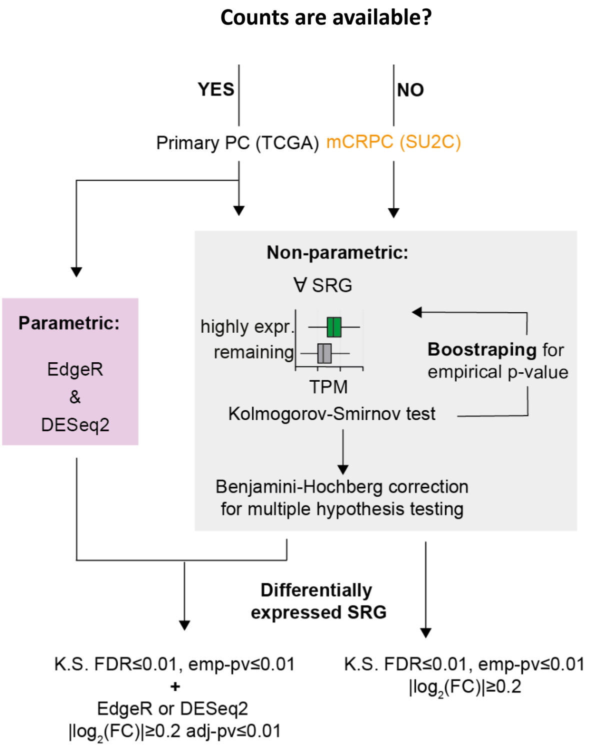

Counts are available?
YES
NO
Primary PC (TCGA)
mCRPC (SU2C)
Parametric:
EdgeR
&
DESeq2
Non-parametric:
∀ SRG
highly expr.
remaining
TPM
Boostraping for empirical p-value
Kolmogorov-Smirnov test
Benjamini-Hochberg correction for multiple hypothesis testing
Differentially expressed SRG
K.S. FDR≤0.01, emp-pv≤0.01
+
EdgeR or DESeq2
|log₂(FC)|≥0.2 adj-pv≤0.01
K.S. FDR≤0.01, emp-pv≤0.01
|log₂(FC)|≥0.2

Figure 25. Schematic representation of the pipeline used to detect RBPs that are differentially expressed in primary and metastatic (mCRPC) PC upon FOXA1 high expression.

Overall, gene expression fold-changes (FCs) between FOXA1 highly expressing and remaining tumors were significantly concordant between the parametric and non-parametric approaches (average Pearson’s correlation coefficient = 0.95, p-value<10-16).

We identified 71 RBPs that were significantly regulated by FOXA1 in either primary or metastatic tumors (Figure 26).

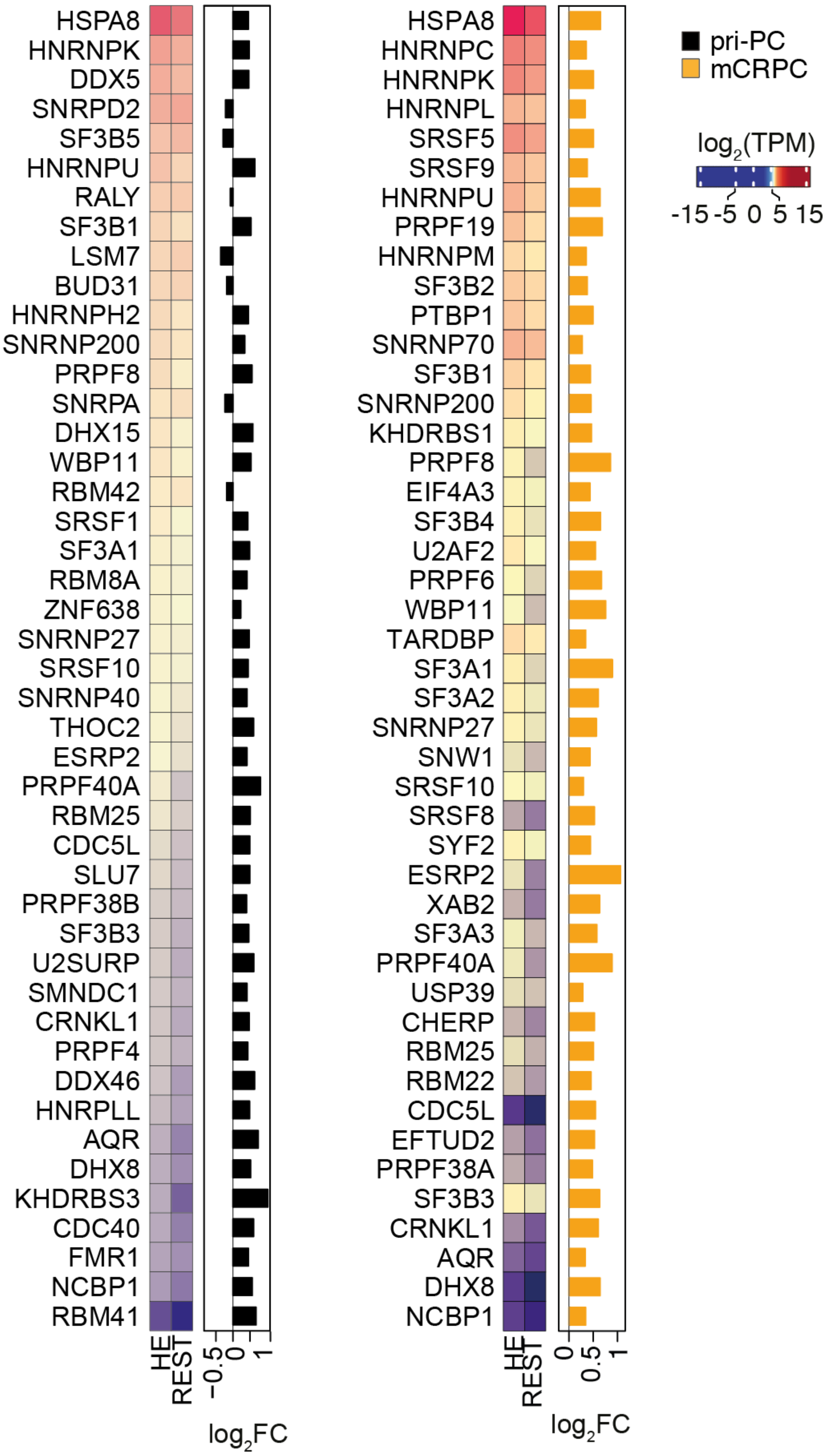


Figure 26. Heatmaps of normalized gene expression levels (TPM) of differentially expressed RBPs between FOXA1 HE and REST samples in pri-PC (left) and mCRPC (right). Barplots on the right indicate the corresponding FC in TPMs between FOXA1 HE and REST samples.

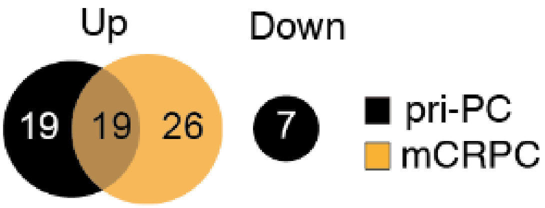


Figure 27. Differentially expressed RBPs between samples with HE of FOXA1 and REST in pri-PC and mCRPC samples.

Of these FOXA1-regulated RBPs, 19 were significantly altered in both primary PC and mCRPC (Figures 27,28).

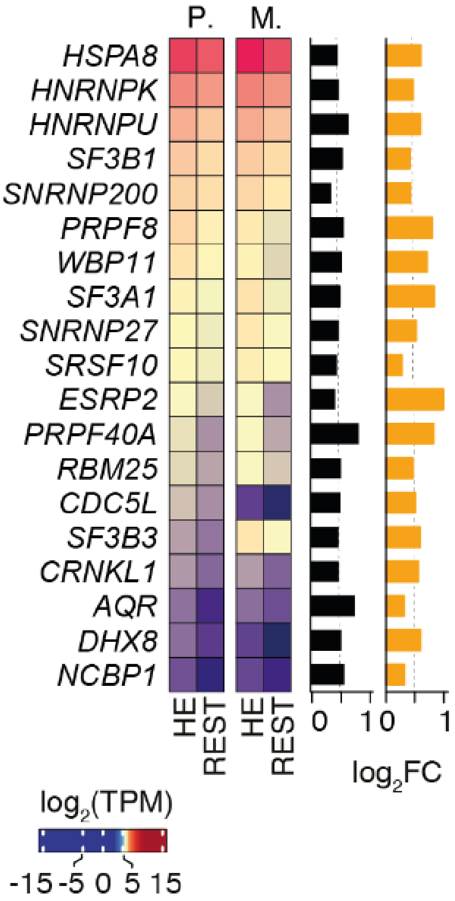


Figure 28. Heatmap of normalized gene expression levels (TPMs) of differentially expressed RBPs between FOXA1 HE and REST samples in pri-PC and mCRPC. Bar plots on the right indicate the corresponding expression fold change (FC) for pri-PC (black) and mCRPC (orange) samples.

To better explore FOXA1 regulation of RBPs in both presence and absence of AR, transient silencing of FOXA1 was performed in AR-dependent (AR^+) VCaP and AR-independent (AR^-) PC3 prostate cancer cell lines.

In vitro validation of FOXA1 calibration of RBP expression

To validate the impact of FOXA1 on RBP expression, we performed deep RNA-seq (~100 million reads) on siRNA-treated (*i.e.* siFOXA1) and non-silenced control (*i.e.* NSI) samples from AR$^+$ VCaP and AR$^-$ PC3 cells, and compared NSI and siFOXA1 data to match the contribution of FOXA1 overexpression in clinical PC (Figure 29, A and B).

VCaP (CRL-2876, ATCC) and PC3 (CRL-1435, ATCC) cells were obtained from ATCC. They were incubated at 37ºC, 5% CO_2 in a humidified incubator and maintained at sub-confluency in RPMI-1640 medium (21875-034, Gibco) or DMEM (41966-029, Gibco) containing 2 mM L-glutamine, supplemented with 10% fetal calf serum (FCS) (Gibco), 100 units/ml penicillin and 100 µg/ml streptomycin (15140-122, Gibco) and regularly tested for the presence of mycoplasma. Transfections with plasmid DNA and siRNA duplexes were carried out as detailed using ViaFect (E4981, Promega) and RNAiMax (13778-075, Thermo Fisher Scientific), respectively, following manufacturers' instructions. TriReagent (AM9738, Invitrogen) was employed to lysate cells and RNA extracted by phase separation using 1-bromo-3-chloropropane. Total RNA was treated with DNAse to exclude genomic contamination and cleared with RNA Clean and Concentration (Zymo Research). Quantity and quality of the material were determined by Qubit Fluorometric Quantitation (Thermo Fisher Scientific) and using the RNA 6000 Nano kit on a Bioanalyzer (Agilent Technologies), respectively. Only samples with an RNA integrity number >7 were selected for library preparation. RNA-seq libraries were generated from 1 µg of RNA using TruSeq total RNA and TruSeq stranded mRNA (Illumina), respectively, following manufacturer's indications. VCaP libraries were sequenced in paired-end modality on Illumina NextSeq500 machine with a read length of 75 nucleotides (nt), while PC3 libraries were sequenced on Illumina NovaSeq6000 in 100nt-long paired-end read modality.

The alignment of raw sequencing reads was performed using STAR (Dobin et al. 2013) on human genome reference GENCODE GRCh37 (Frankish et al. 2019) and gene expression was estimated using featureCounts (Liao, Smyth, and Shi 2014). Hierarchical clustering and principal component analyses of gene expression normalized data showed that samples were appropriately separated upon silencing conditions (Figure 29, C and D).

Next, we performed canonical parametric differential expression analysis between NSI and siFOXA1 samples to match the contribution of FOXA1 overexpression in clinical PC. On average, 76% of the differentially-expressed RBPs in primary PC and/or mCRPC were concordantly regulated by FOXA1 in the two cell lines (Figure 29, E).

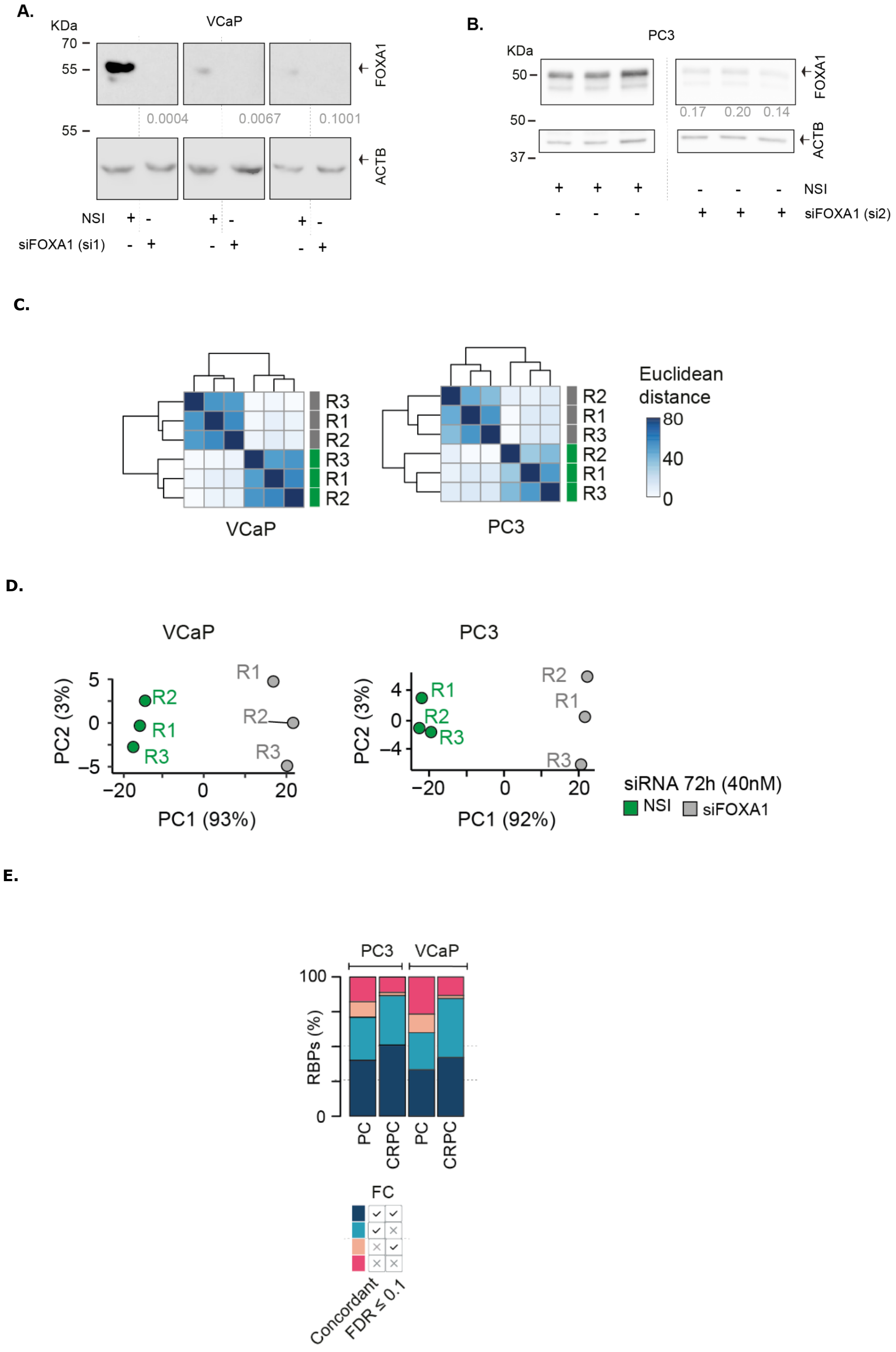

A.
KDa
VCaP
70
55
55
FOXA1
ACTB
0.0004
0.0067
0.1001
NSI + - + - + -
siFOXA1 (si1) - + - + - +
B.
PC3
KDa
50
50
37
FOXA1
ACTB
0.17 0.20 0.14
+ + + - - - NSI
- - - + + + siFOXA1 (si2)
C.
R3 R1 R2 R3 R1 R2
VCaP
R2 R1 R3 R2 R1 R3
PC3
Euclidean distance
80
40
0
D.
VCaP
PC2 (3%)
PC1 (93%)
R1 R2 R3
R2 R1 R3
5 0 −5
−20 0 20
PC3
PC2 (3%)
PC1 (92%)
R1 R2 R3
R2 R1 R3
4 0 −4
−20 0 20
siRNA 72h (40nM)
NSI
siFOXA1
E.
PC3
VCaP
RBPs (%)
100
0
PC CRPC PC CRPC
FC
Concordant
FDR ≤ 0.1
✓ ✓
✓ ×
× ✓
× ×

Figure 29. **(A,B).** Western blotting images of FOXA1 depletion in VCaP (A) and PC3 (B) cells used for RNA-seq analysis by transfection with one siRNA duplex (si1 or si2, 40nM for 72 hours). Normalized (to ACTB) protein expression compared to control (NSI) , calculated by densitometric band quantitation, is shown below FOXA1 blot images. **C.** Hierarchically clustered heatmaps of Euclidean distance between expression values for VCaP and PC3. **D.** Scatter plot of the first two components of a principal component analysis for VCaP and PC3 RNA-seq datasets with the percentage of variance explained by each component reported on each axis. **E.** Bar plots showing the fractions of FOXA1-regulated RBPs that changed expression levels between NSI and siFOXA1 samples in VCaP and PC3 cells with respect to their expression change in pri-PC and mCRPC.

Out of the 19 FOXA1-regulated RBPs that were altered in both primary PC and mCRPC, HNRNPK, SNRNP200, WBP11, ESRP2, SF3B3 and NCBP1 were significantly differentially expressed by FOXA1 in both cell lines (Figure 30).

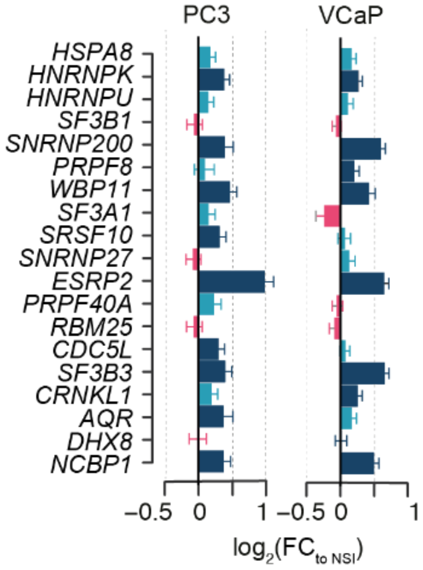


Figure 30. Bar plots showing $\log_2$ FC in the expression between NSI and siFOXA1 from VCaP and PC3 RNA-seq datasets for RBPs differentially expressed in pri-PC and mCRPC. Colors indicate whether the expression change is significant and concordant with that observed in pri-PC and mCRPC analyses.

Overall, these results clearly demonstrate that FOXA1 directly drives the expression of a subset of splicing factors.

Systematic analysis of FOXA1 binding sites in prostate cancer cell lines

We next sought to systematically investigate the three-dimensional architectural features of transcriptional control by FOXA1 in PC. FOXA1 is known for binding mainly on enhancer regions. Moreover, 3D structure of VCaP and LNCaP prostate cancer cell lines has been recently studied (Ramanand et al. 2020), so a pipeline for integrating 3D structure information and PC-specific open chromatin bindings was developed.

To define the prostate cancer-specific landscape of FOXA1 regulation, we combined 3D structure information of two cell lines, VCaP and LNCaP, with FOXA1 ChIP-seq data and ATAC-seq data from PC TCGA patients. RNA Pol II–associated enhancer-gene interactions defined by Chromatin Interaction Analysis by Paired-End Tag sequencing (ChIA-PET) in VCaP and LNCaP cell lines were obtained from Ramanand et al. (Ramanand et al. 2020). Promoter regions were defined as explained in chapter 3. To select regulatory regions that are related to sites of active transcription in PC, promoter and enhancer regions were intersected with PC-specific accessible elements and only overlapping regions were retained. Candidate enhancer-gene interactions were retained if associated with the related promoter and enhancer-gene associations in which the enhancer overlapped with the promoter of the same gene were discarded. Interaction of maximum one million base pairs were retained for further analyses (Figure 31).

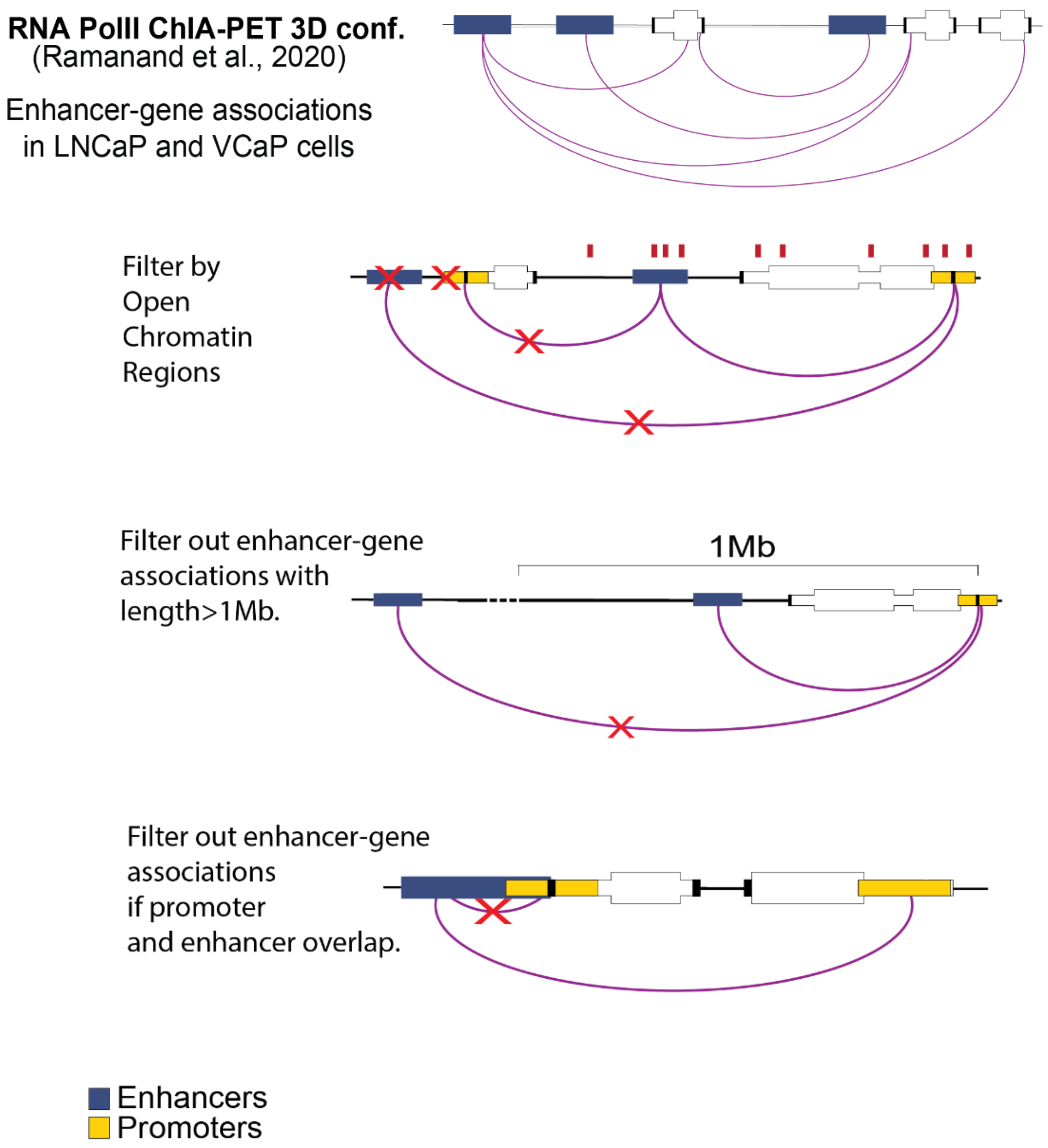

RNA PolII ChIA-PET 3D conf.
(Ramanand et al., 2020)
Enhancer-gene associations in LNCaP and VCaP cells
Filter by Open Chromatin Regions
Filter out enhancer-gene associations with length>1Mb.
1Mb
Filter out enhancer-gene associations if promoter and enhancer overlap.
Enhancers
Promoters

Figure 31. Schematic representation of the filters applied on enhancer-gene associations.

To identify FOXA1 binding regions in LNCaP and VCaP cells, significant peak calls (*i.e.* p-value≤10^{-5}) of four and five FOXA1 ChIP-Seq experiments, respectively, were obtained from ChIP-Atlas (Oki et al. 2018) (read also “Appendix B. Insights into literature”, A.). The choice fell on this database as experiments can be chosen manually to retain only wild type ones (*e.g.* discarding those in which cells are treated with drugs or transiently silenced for other TFs). For each cell line, peaks were positionally sorted and merged and FOXA1 binding regions were intersected with PC-specific accessible elements as explained before. Only overlapping regions were retained and considered as FOXA1 active binding sites (Figure 32).

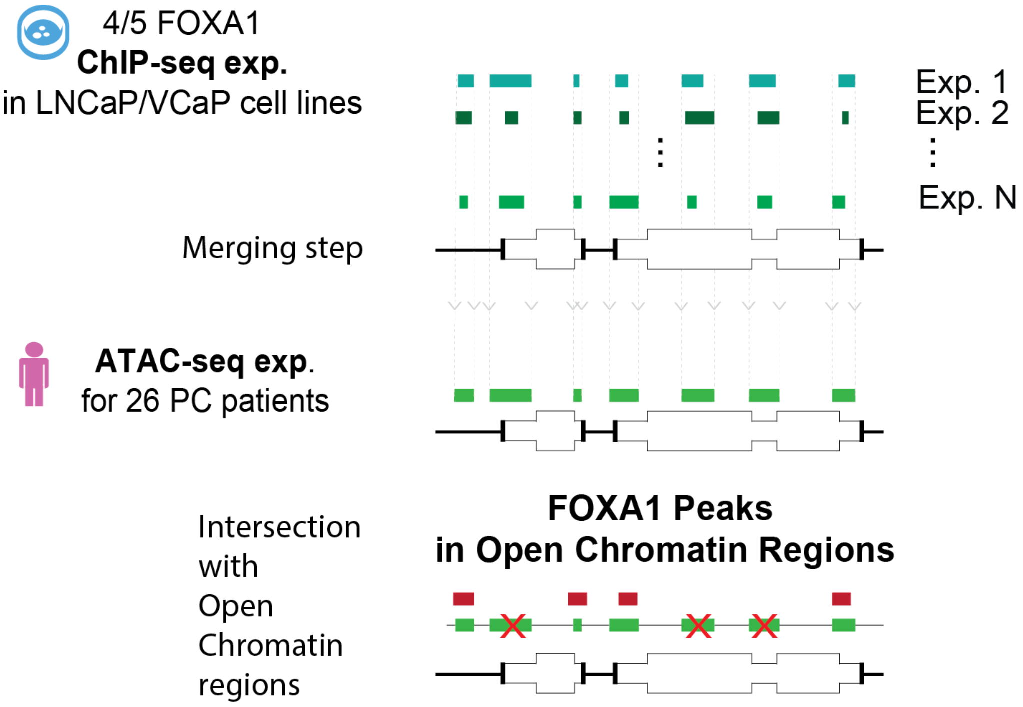


Figure 32. Schematic representation of the pipeline used to identify FOXA1 active binding sites.

To identify genes putatively regulated by FOXA1, active binding sites were intersected with promoter and enhancer regions.

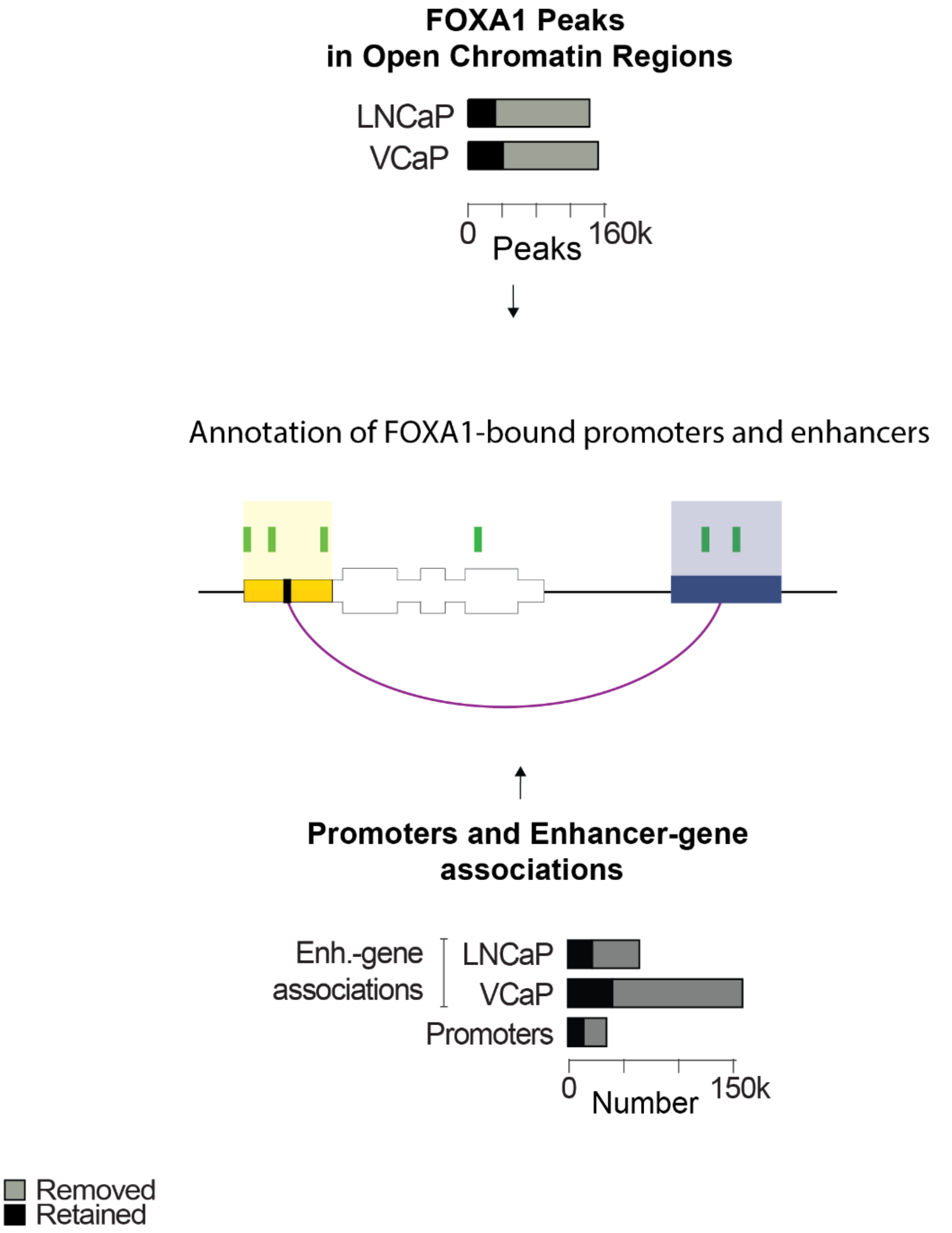

FOXA1 Peaks
in Open Chromatin Regions
LNCaP
VCaP
0
Peaks
160k
Annotation of FOXA1-bound promoters and enhancers
Promoters and Enhancer-gene associations
Enh.-gene associations
LNCaP
VCaP
Promoters
0
Number
150k
Removed
Retained

Figure 33. Schematic representation of the pipeline used to assess the overrepresented genes involved in biologically-relevant processes with FOXA1 active binding sites within PC-specific chromatin-accessible promoters and associated enhancers.

Finally, for each regulatory element (*i.e.* promoters and enhancers), putatively FOXA1-regulated genes were selected and over-representation analyses of genes in the GIP gene sets were performed. Enrichment tests with false discovery rate (FDR)≤0.1 were considered as significant (Figure 33).

Among the 16 GIP gene sets, RBPs showed the highest enrichment of genes with FOXA1 active binding sites in their regulatory regions both in VCaP- and LNCaP-based datasets (Figure 34).

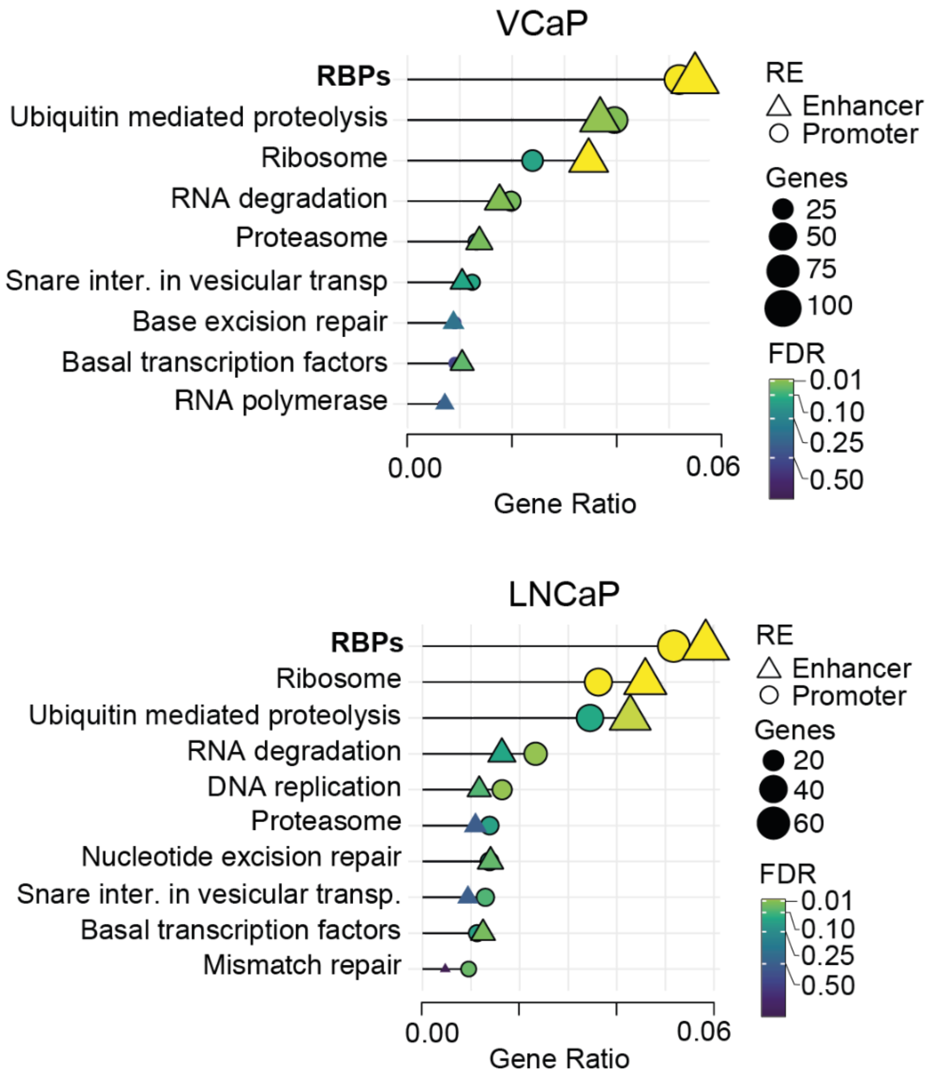


Figure 34. Over-representation analysis (ORA) results on genes with FOXA1 binding sites at promoters (circles) and enhancers (triangles) in (**Upper panel**) VCaP and (**Bottom panel**) LNCaP cells.

Next, differentially expressed RBPs according to differential expression analysis on primary and metastatic patients that harbor FOXA1 active binding sites on promoters and/or enhancers were identified. Out of the 71 RBPs, respectively 85% and 65% harbored FOXA1 active binding sites at their interacting regulatory regions in VCaP and LNCaP cells (Figure 35).

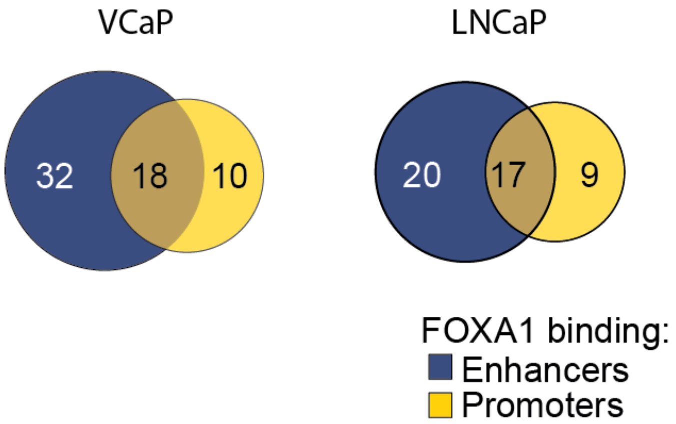


Figure 35. FOXA1-regulated RBPs in pri-PCs or mCRPCs with at least one FOXA1 binding site on gene promoter (yellow) or enhancer (blue) regions in (**Left panel**) VCaP data or (**Right panel**) LNCaP data.

Through a two-tailed Test of Equal Proportions was shown that FOXA1 preferentially binds enhancer over promoter regions of RBPs (p-value=9x10^{-8} and 6x10^{-4} for VCaP- and LNCaP-based dataset, respectively), nevertheless seven FOXA1-regulated RBPs harbored active binding sites in both promoter and enhancer regions in the two datasets (Figure 36).

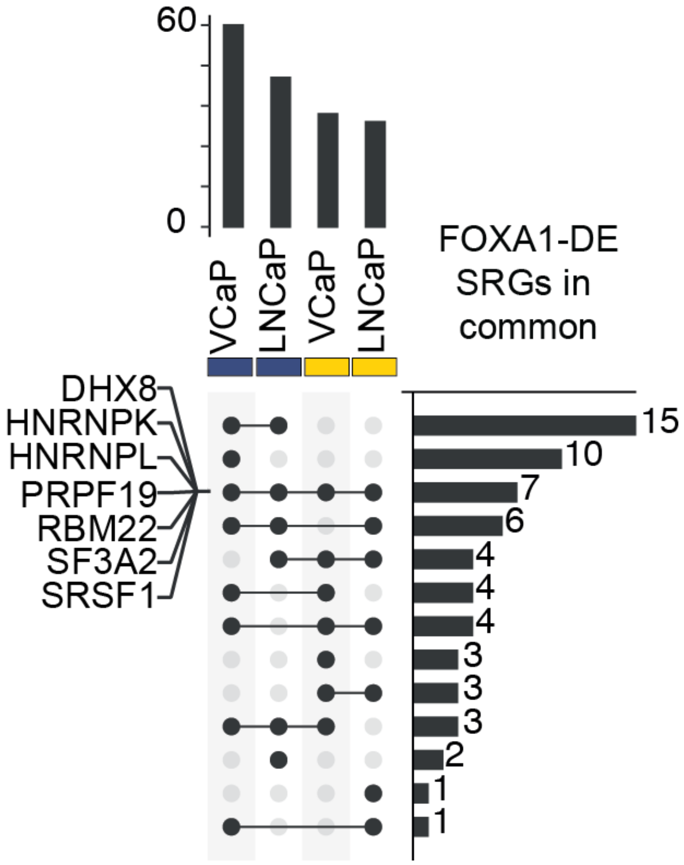


Figure 36. UpSetR plot showing the intersection between FOXA1-regulated RBPs with FOXA1-bound promoters (yellow) and enhancers (blue) in VCaP and LNCaP cells. The seven RBPs with FOXA1 binding sites in both promoters and enhancer regions in both VCaP and LNCaP cells are indicated as text.

An example can be seen in the two forthcoming figures (Figures 37,38). Inspecting the transcriptional architecture of HNRNPK in the VCaP-based dataset, FOXA1 bindings are found on the promoter and six associated active enhancers. Similarly, in the LNCaP-based dataset, FOXA1 bound the HNRNPK promoter and a single interacting active enhancer at ~6.4kb, which was shared between the datasets.

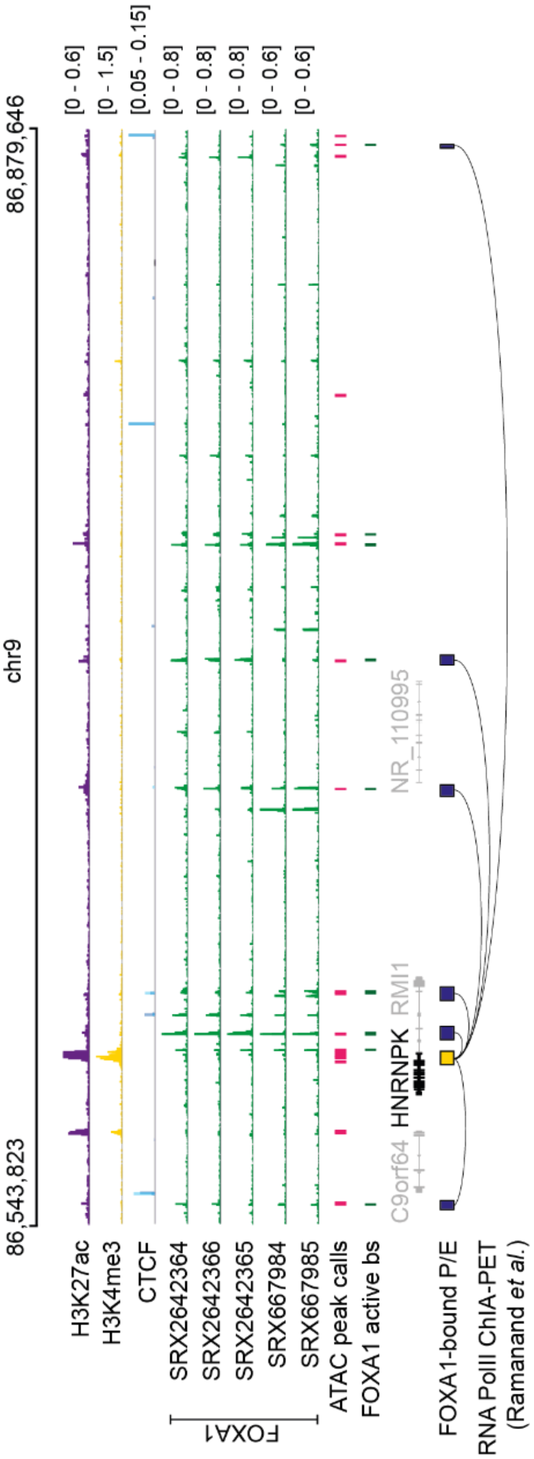


Figure 37. Genome view representing *HNRNPK* promoter and enhancer regions that are bound by FOXA1 in VCaP cells. ChIP-seq density read tracks for H3K27ac, H3K4me3, CTCF (2 overlayed experiments) and FOXA1 in VCaP cells are shown together with pri-PC-reproducible ATAC-seq peak calls, FOXA1 active binding sites (bs) and RNA PolII ChIA-PET-derived FOXA1 bound enhancer-HNRNPK associations. The position of the considered HNRNPK regulatory regions along chromosome 9 is shown at the top of the panel.

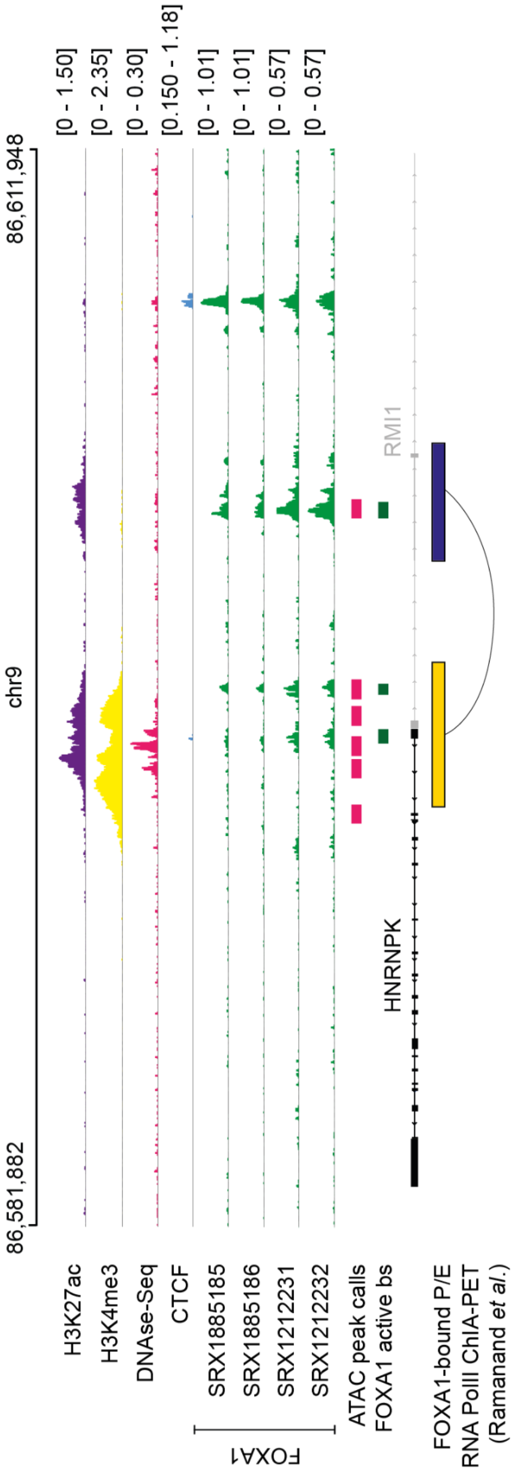


Figure 38. Genome view representing *HNRNPK* promoter and enhancer regions that are bound by FOXA1 in LNaP cells. ChIP-seq density read tracks for H3K27ac, H3K4me3, DNAse-Seq, CTCF and FOXA1 in LNCaP cells are shown together with pri-PC-reproducible ATAC-seq peak calls, FOXA1 active binding sites (bs) and RNA PolII ChIA-PET-derived FOXA1 bound enhancer-HNRNPK associations. The position of the considered HNRNPK regulatory regions along chromosome 9 is shown at the top of the panel.

Overall, these results clearly demonstrate that FOXA1 directly drives RBP expression, by preferentially binding cognate chromatin-accessible active enhancers. The direct transcriptional control of FOXA1 primarily impacts on splicing factors as compared to other fundamental biological processes.

5. FOXA1-dependent AS in prostate cancer

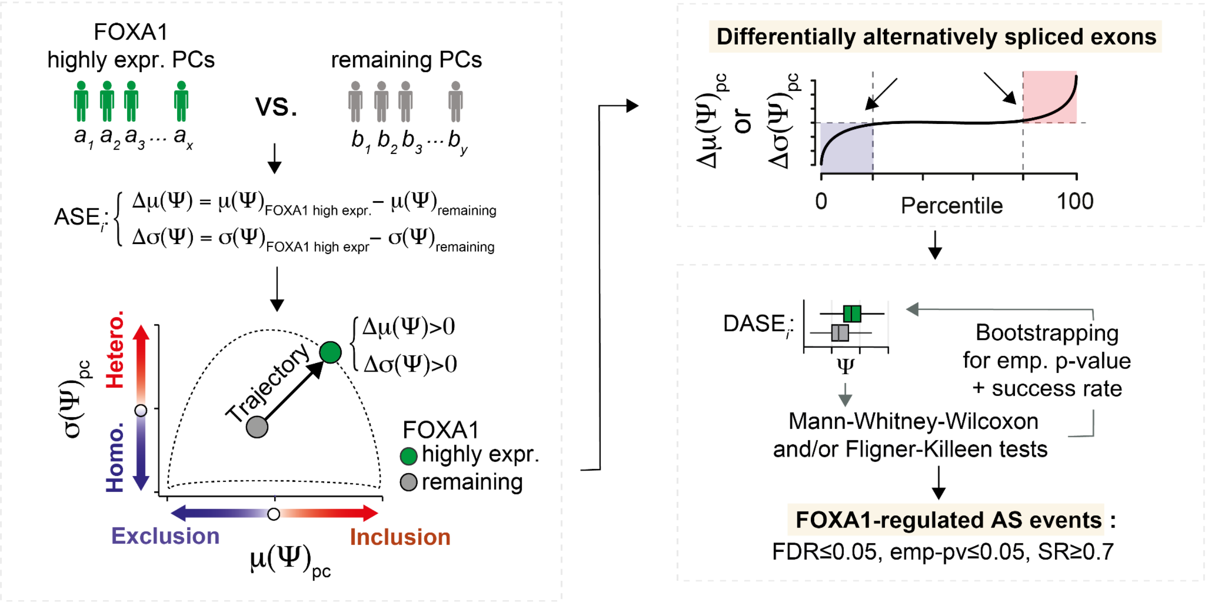
One good way to understand a complex system is to disturb it and see what happens.

Michael Pollan

Identification FOXA1-regulated alternative splicing events

As we found that FOXA1 primarily controls splicing factors, we next sought to determine its impact on the alternative splicing landscape of PC. To do this, we explored the level of inclusion of 60,699 AS events in their corresponding transcripts across 384 primary tumors (Kahles et al. 2018). The catalog of alternative splicing events was obtained from The GDC portal (https://gdc.cancer.gov/about-data/publications/PanCanAtlas-Splicing-2018). This atlas included five categories of AS events: Cassette Exon (CE), Alternative 3’ (A3) and 5’ (A5), Intron Retention (IR) and Mutually Exclusive exons (MEX). The percent of spliced in (Ψ) value was used as a measure of splicing event inclusion in the mature mRNA (Schafer et al. 2015). AS events with (i) available information in more than 75% of the samples (Schafer et al. 2015), (ii) mean (μ) Ψ ranging from 0.01 and 0.99 (*i.e.* not constitutively excluded or included, respectively), and (iii) in genes with less than 500 events were retained for further analysis. For each selected AS event, missing values were replaced by the mean of the corresponding Ψ distribution across samples (Li et al. 2017). For each AS event the μ and standard deviation (σ) of Ψ levels in FOXA1 highly expressing and remaining samples were calculated. The difference of μ and σ of Ψ levels (*i.e.* Δμ(Ψ) and Δσ(Ψ)) between the two cohorts was then measured. To identify AS events associated with FOXA1 high expression, events with negligible changes in Δμ(Ψ) and Δσ(Ψ) were discarded. In particular, variable AS events were defined based on the quantile distributions of Δμ(Ψ) and Δσ(Ψ). An AS event was defined as variable either (i) if Δμ(Ψ) was lower or greater than the 15th or the 85th percentile of Δμ(Ψ) distribution, respectively, or (ii) if Δσ(Ψ) was lower or greater than the 20th or the 80th percentile of Δσ(Ψ) distribution (Attig et al. 2018; Agirre et al. 2021), respectively. In total, we obtained 30,439 variable AS events.

Next, to select variable AS events that were significantly differentially included between FOXA1 HE and REST samples, two non-parametric statistical tests were performed. For each variable AS event, Δμ(Ψ)s between FOXA1 HE and REST tumors were tested using a two-tailed Wilcoxon Rank Sum test, whereas Δσ(Ψ)s were compared using a two-tailed Fligner-Killeen test (Saraiva-Agostinho and Barbosa-Morais 2019). P-values were corrected for multiple testing using the Benjamini–Hochberg procedure. To calculate the emp-pv of each comparison, sample labels were shuffled for 1,000 times and at each iteration the two tests were performed. The empirical p-values of each test were calculated as the number of times p-values were smaller than the observed ones. To account for the sample size difference between FOXA1 HE and REST cohorts, the latter was randomly down-sampled to reach the size of the former for 1,000 times. At each iteration, tests were performed. The success rate (SR) was then computed as the proportion of significant results (p-value<0.05) over the total number of comparisons. Variable AS events with an FDR<0.05, emp-pv<0.05 and SR>0.7 of at least one test were considered as significantly differentially included between FOXA1 HE samples and REST and named as FOXA1-regulated AS events (Figure 39).



FOXA1
highly expr. PCs
remaining PCs
$a_1\, a_2\, a_3 \cdots a_x$
vs.
$b_1\, b_2\, b_3 \cdots b_y$
ASE$_i$: $\Delta\mu(\Psi) = \mu(\Psi)_{\text{FOXA1 high expr.}} - \mu(\Psi)_{\text{remaining}}$
$\Delta\sigma(\Psi) = \sigma(\Psi)_{\text{FOXA1 high expr}} - \sigma(\Psi)_{\text{remaining}}$
$\sigma(\Psi)_{pc}$
Hetero.
Homo.
Trajectory
$\Delta\mu(\Psi)>0$
$\Delta\sigma(\Psi)>0$
FOXA1
● highly expr.
● remaining
Exclusion
Inclusion
$\mu(\Psi)_{pc}$
Differentially alternatively spliced exons
$\Delta\mu(\Psi)_{pc}$ or $\Delta\sigma(\Psi)_{pc}$
0
Percentile
100
DASE$_i$:
Ψ
Bootstrapping
for emp. p-value
+ success rate
Mann-Whitney-Wilcoxon
and/or Fligner-Killeen tests
FOXA1-regulated AS events :
FDR≤0.05, emp-pv≤0.05, SR≥0.7

Figure 39. Schematic representation of the pipeline used to identify FOXA1-regulated events in primary PC. Trajectories (represented as arrows) are defined by mean (μ) and standard deviation (σ) inclusion (Ψ) changes (Δ) of each AS event towards high FOXA1 expression.

We identified 7,121 FOXA1-regulated AS events showing significant μ(Ψ) and σ(Ψ) changes between the two groups (1,181 CEs, 903 A3, 680 A5, 2,991 IR and 1,366 MEX) (Figure 40).

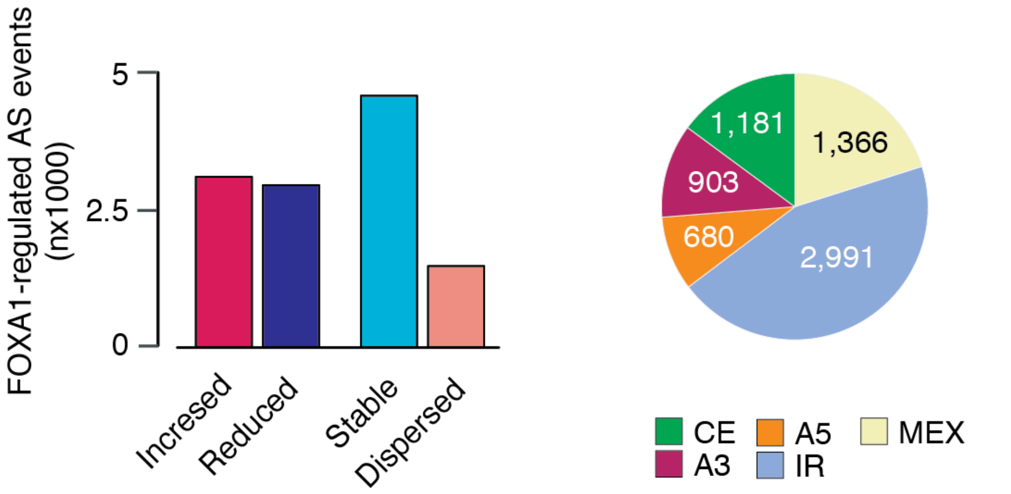


Figure 40. **Left panel.** Barplot showing the number of FOXA1-regulated AS events selected for mean (μ), increased or reduced, and standard deviation (σ), stable or dispersed. **Right panel.** Pie chart showing numbers of FOXA1-regulated AS events by event type.

To characterize the impact of FOXA1-mediated AS regulation on fundamental biological processes, we next performed over-representation analysis of genes harboring FOXA1-regulated AS events in primary tumors.

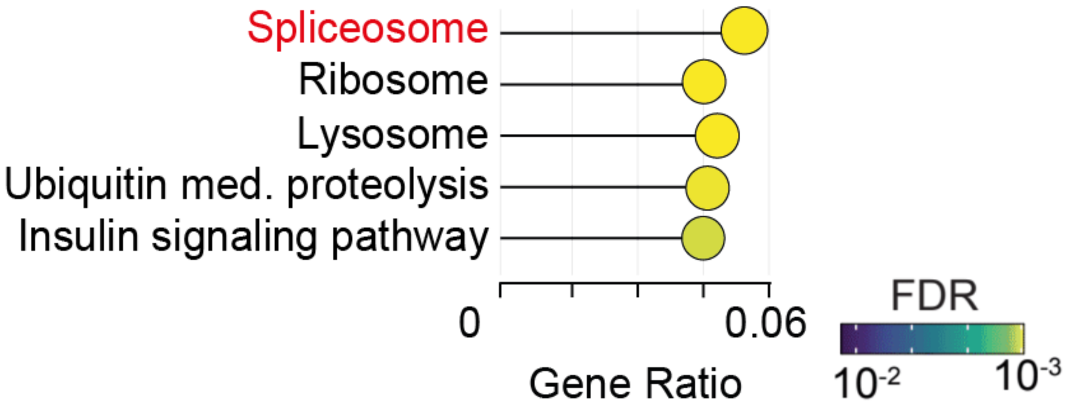


Figure 41. Over representation analysis performed on genes harboring FOXA1-regulated AS events.

Out of 186 canonical KEGG pathways, we found that the spliceosome gene set was the top-ranked affected process (Figure 41). This result was also confirmed when we stratified AS events into categories corresponding to the major types of AS patterns. These results implicate FOXA1 in a regulatory feedback loop involving splicing factors.

Identification of *cis*-acting motifs that controls AS of FOXA1-regulated cassette exons

RBPs regulate splicing according to position-dependent principles, which can be exploited for analysis of regulatory motifs. To identify the principle governing AS of FOXA1-regulated exons we then focused on CEs. To do so, we employed RNAmotifs (Cereda et al. 2014), a method that evaluates the sequence around differentially regulated alternative exons to identify clusters of short and degenerate sequences, referred to as multivalent RNA motifs (Figure 42). It has been shown that diverse RBPs share basic positional principles, but differ in their propensity to enhance or repress exon inclusion.

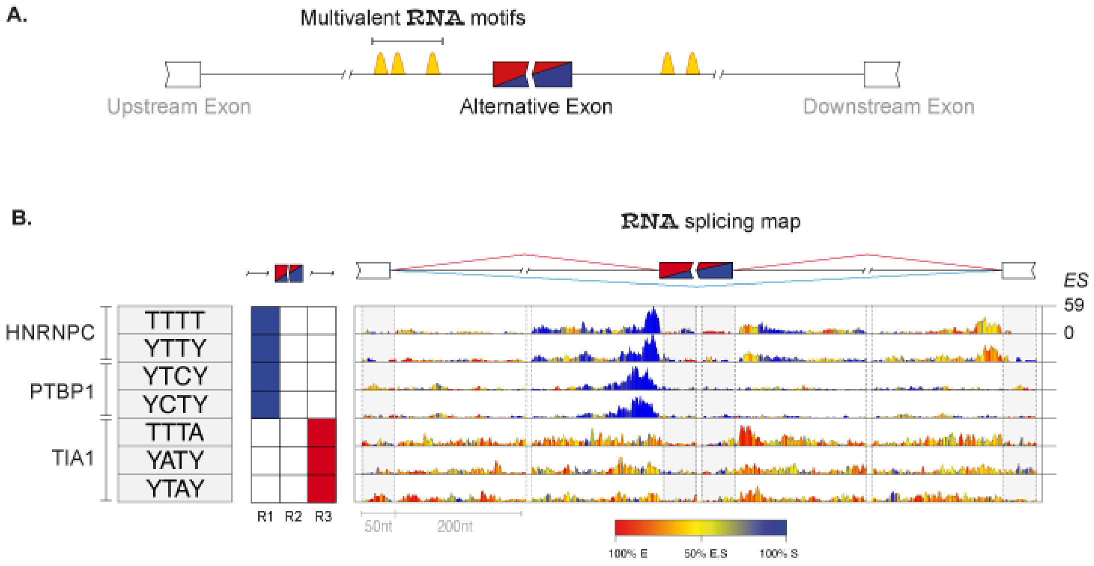


Figure 42. **A.** Schematic representation of multivalent motifs (*i.e.* clusters of short degenerate motifs) surrounding the alternative spliced exon. **B.** RNA splicing maps of multivalent RNA motifs enriched in the 'mixed' set of exons regulated by hnRNP C, PTBP1 and TIA. Sequences of the enriched tetramers are shown on the left, followed by a color-coded panel showing the regions where tetramer enrichment reached the defined threshold around silenced (blue) or enhanced (red) exons. The right panel depicts the nucleotide-resolution RNA splicing map of each motif at the enhanced or silenced exons, and their flanking exons. The color key indicates whether the position-specific contribution originates from enhanced (red),silenced (blue),or both (yellow) sets. The maximum enrichment score (ES) value of the top tetramer is reported on the right.

RNAmotifs was used to identify the 512 degenerate and not-degerate multivalent RNA motifs of 4nt length (*i.e.* tetramers) that occurred in a specific region more often in the 1,181 FOXA1-regulated CEs of interest compared with 2,929 constitutive exons (*i.e.* $\mu_{PC}(\Psi)>0.999$

and $|\Delta\mu(\Psi)|<2.5x10^{-5}$) defined as controls. The tool was run considering three enrichment regions: (i) R_1 [-205:-5] nucleotides of intronic sequence upstream of the 3′ splice site (ss); (ii) R_2 corresponding to the entire exonic sequence (or up to 200 nt from both ss in case of exon was longer than 400 nt); and (iii) R_3 [10:210] nucleotides of intronic sequence downstream of the 5′ ss. RNAmotifs empirical p-values were calculated using 10,000 bootstrap iterations. Tetramers with RNAmotifs Fisher's p-values and empirical p-values smaller than 0.05 (or the 1st percentile of the p-value distribution in case of highly significant results) and 0.0005, respectively, were retained for further analysis. To visualize the exact positions in the pre-mRNA where clusters of RNA motifs are enriched, RNAmotifs was run performing a position-specific enrichment analysis at exon/intron junctions of alternative CEs and flanking exons extending 1,000 and 50 nucleotides into introns and exons, respectively.

We obtained 23 significantly enriched tetramers within FOXA1-regulated CEs and their flanking introns, of which the majority (n=19) were associated with more-excluded exons, corroborating the greater impact of FOXA1 on exon silencing (Figure 43). Clusters of T-rich motifs were enriched at the 3' splice sites (ss) of the upstream introns, matching the location of the polypyrimidine (Py)-tract that is required for exon definition. Conversely, GAC-rich tetramers characterized the 5'ss of FOXA1-regulated exons, supporting the involvement of auxiliary splicing regulators in ss recognition.

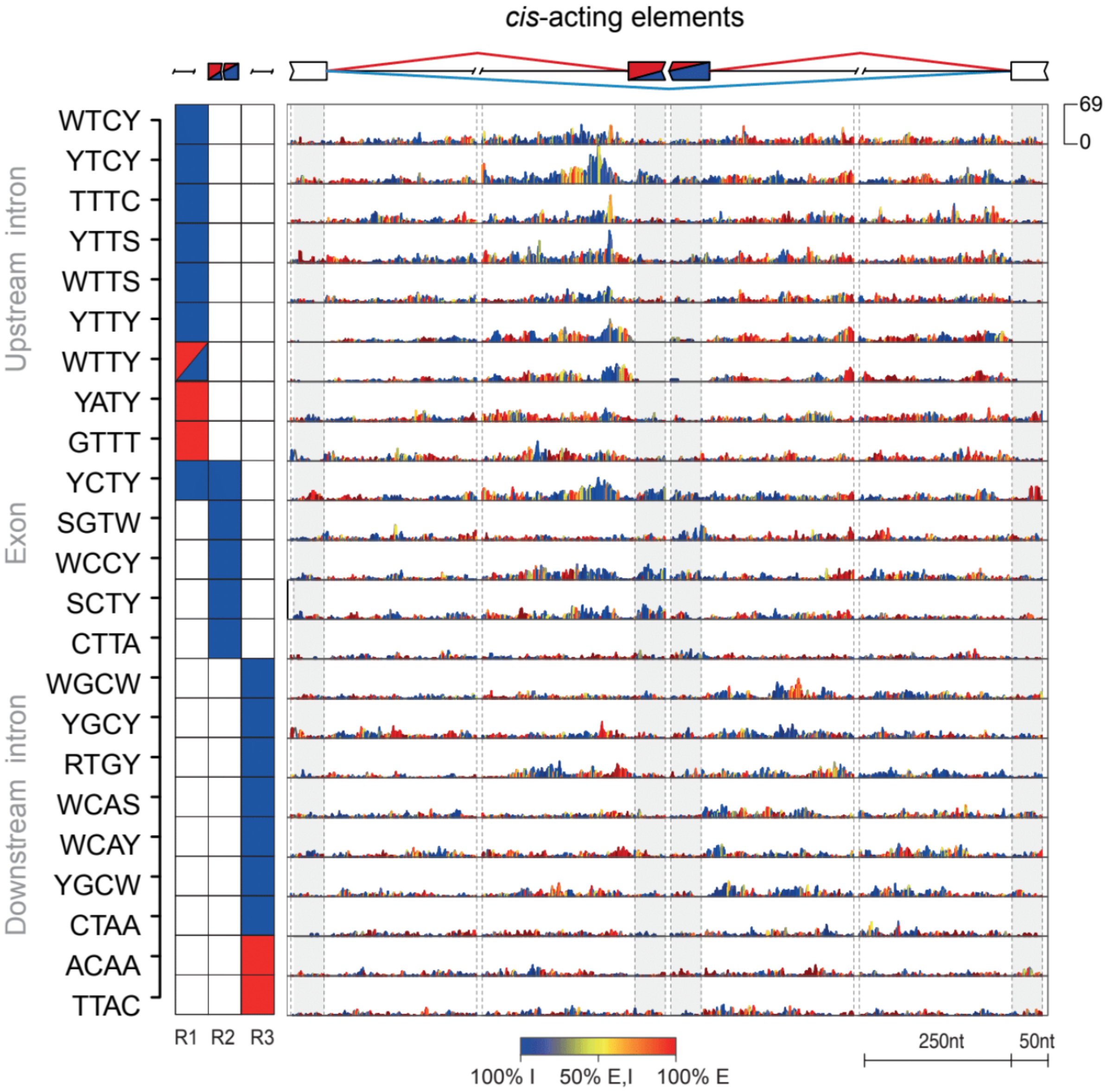


Figure 43. RNA splicing map of multivalent RNA motifs enriched at FOXA1-regulated exons relative to constitutive ones (*i.e.* RNAmotifs Fisher's test p-values cutoffs = 1.5x10^{-3}, 5x10^{-2} and 1.2x10^{-2} for R1, R2 and R3, respectively). Sequences of the enriched tetramers are shown on the left, followed by a color-coded panel showing the regions where tetramer enrichment reached the defined threshold around inhibited (blue) or enhanced (red) exons. The right panel depicts the nucleotide-resolution RNA splicing map of each motif at the FOXA1-enhanced or FOXA1-inhibited exons, and their flanking exons. The color key indicates whether the position-specific contribution originates from enhanced (E; red), inhibited (I; blue), or both (yellow) sets. The maximum RNAmotifs enrichment score value of the top tetramer, which is used to plot all tetramers, is reported on the right. Nt, nucleotides.

Association of *cis*-acting elements to *trans*-acting factors

To assess the trans-acting regulation of CE inclusion, we associated cis-acting sequences to cognate trans-acting factors. In particular, in light of the reproducibility of splicing factor binding positions across cell types (Van Nostrand, Pratt, et al. 2020), we searched for RBP cross-linking sites from eCLIP experiments in HepG2 cells at FOXA1-regulated exons with tetramer instances. Then, we associated tetramers to cognate RBPs on similarity of (1) their sequence with canonical RBP consensus motifs, and (2) position-dependent representation of their occurences (*i.e.* splicing maps) with those of RBP cross-linking sites at exon-intron junctions.

To do so, a list of 466 11-nt long position weight matrices (PWMs) derived from HepG2 eCLIP data for 62 RBPs was collected from the mCross database (Feng et al. 2019). For each enriched tetramer, a PWM was computed on tetramer occurrences at regulated exons extending both tetramer sides of two nucleotides. Similarities between tetramer and mCross PWMs were calculated using the MACRO-APE tool (Vorontsov, Kulakovskiy, and Makeev 2013) with parameters --position J,direct with J=-3,-2,-1,0 to allow up to four different alignments to the most informative seven core positions of the mCross PWM (Feng et al. 2019). For each tetramer and RBP pair, the highest similarity amongst the four alignments was retained. In case of multiple mCross PWMs for the same RBP, the different similarity values were averaged. Hence, the similarity value was measured as follows:

$$\forall\ RBP\ and\ tetramer{:}\ \Omega\ = \frac{\sum_{i=1}^{N_{PWM}} \omega_i}{N_{PWM}}$$

where ω_i is the similarity value between the tetramer and the i^{th} mCross PWM or N_{PWM} is the total number of mCross PWMs of a RBP. Ω was named “sequence similarity score”.

Next, the similarity between profiles of the RNAmotifs maps of each enriched tetramer and those of eCLIP-based RNA splicing maps of the 62 RBPs was assessed. Firstly, cross-linking sites, as iCounts peak instances, from eCLIP experiments in HepG2 cells for each RBP were collected (Curk et al. (2019) iCount: protein-RNA interaction iCLIP data analysis (in preparation)). Then, for each tetramer, eCLIP-based splicing maps of all RBPs were generated around exons with tetramer instances (*i.e.* extending 1,000 and 50 nucleotides into introns and exons). At each position, and for each RBP, a cross-linking enrichment score

(CES) was computed by performing a Fisher’s exact test comparing the proportion of FOXA1-regulated and constitutive exons having at least one iCounts peak:

$$CES\ =\ -\ 2log(p)$$

where p is the p-value of the Fisher’s exact test.

The similarity between the RNAmotifs and eCLIP-based RNA splicing map was then evaluated by calculating the Battacharyya coefficient (BC) (Rizzo et al. 2019) as follows:

$$BC(q,t)\ =\sum_{i=1}^{n}\sqrt{q_i t_i}$$

where q_i is the RNAmotifs ES of the tetramer at position i on the map and t_i is the CES of the RBP at the same position i, and n is the length of the maps.

Finally, for each tetramer and RBP a global Matching Score (MS) was computed as the product of the sequence similarity score Ω and the map similarity given by the Battacharyya coefficient:

$$\forall\ RBP\ and\ tetramer\text{:}\ Matching\ Score\ =\ \Omega\cdot BC$$

RBPs with MS ≥ 75th percentile of its distribution were considered as significantly associated with the corresponding tetramer.

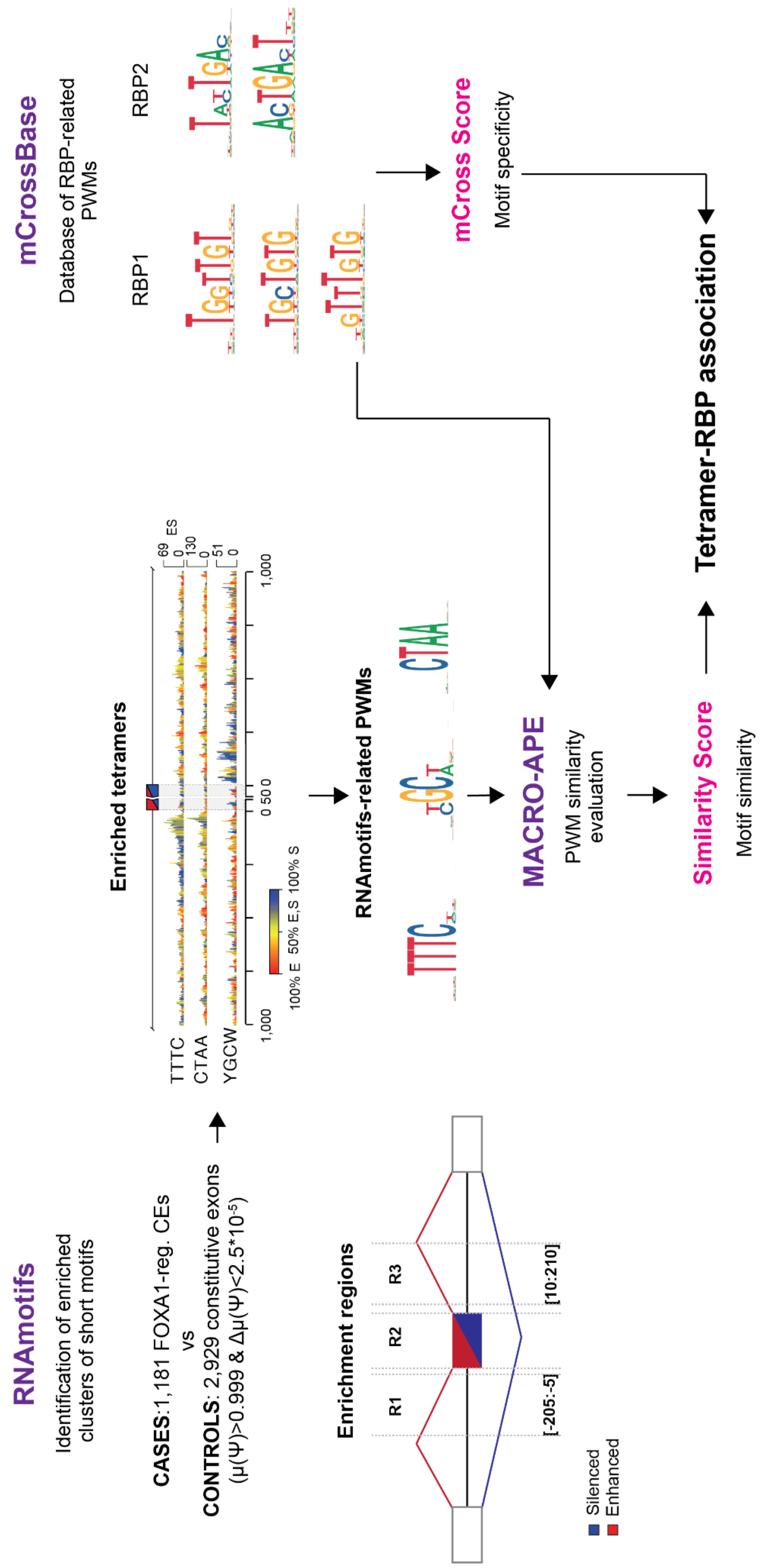


Figure 44. Schematic overview of the integrated analysis employed to associate *cis*-acting elements of FOXA1-regulated CEs to *trans*-acting factors.

The outcomes of this analysis recapitulated many known RBP binding principles (Figure 45), including HNRNPC, PTBP1 and U2AF2 at 3'ss (Sutandy et al. 2018; Zarnack et al. 2013), SF3 complex components (SF3B4 and SF3A3) at the branch point site (Lardelli et al. 2010), PRPF8 at 5'ss (Wickramasinghe et al. 2015), and HNRNP family (HNRNPK, HNRNPM, and HNRNPU) at upstream and downstream intronic regions (Van Nostrand, Freese, et al. 2020; Van Nostrand, Pratt, et al. 2020).

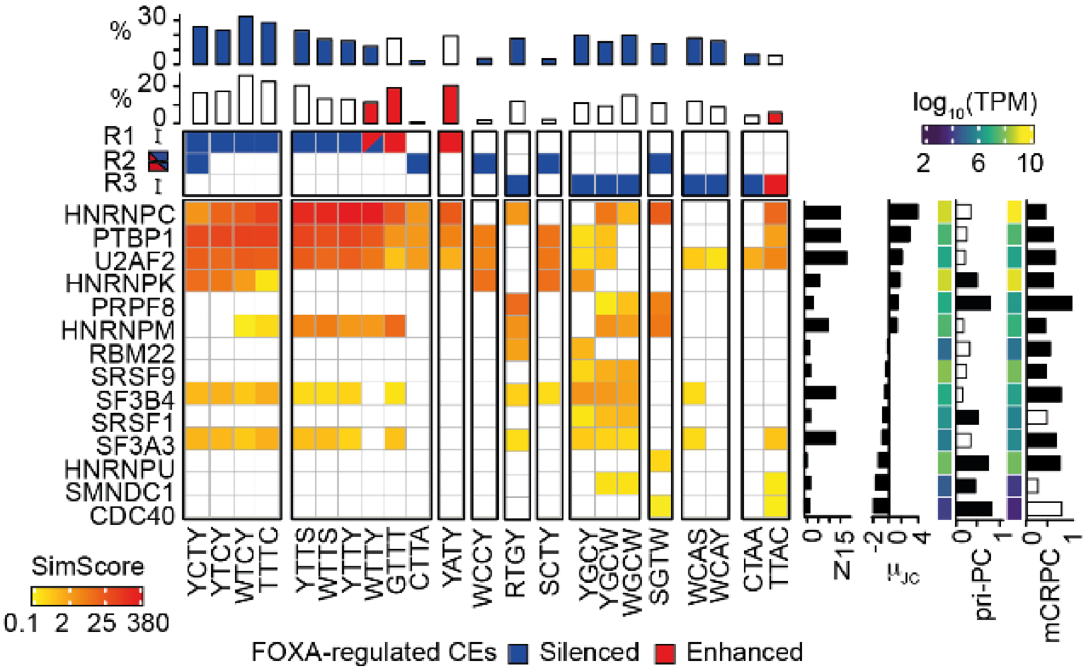


Figure 45. Heatmap of similarity scores (SimScore) between significantly enriched multivalent RNA motifs identified amongst FOXA1-regulated CEs relative to constitutive exons (RNAmotifs two-tailed Fisher's test p-values cutoffs = $1.5\text{x}10^{-3}$, $5\text{x}10^{-2}$ and $1.2\text{x}10^{-2}$ for R1, R2 and R3, respectively) and mCross position weight matrices (PWMs) of RBPs differentially expressed in pri-PC or mCRPC. Only SimScore ≥ 75th percentile of their distribution across all motif-RBP pairs are shown. Top annotation heatmap shows the RNAmotifs regions (R1, R2, R3) where tetramer enrichment reached the defined threshold around more-excluded (blue) or more-included (red) CEs. Top bar plots depict the percentage of FOXA1-regulated CEs with tetramer occurences. Motifs are grouped based on clusters identified by RNAmotifs. Right annotation shows the number of significant associations with a tetramer for each RBP (N), the $\log_2$ average similarity across significant associations (μ_{JC}), and the average expression level of each RBP in pri-PC and mCRPC with the corresponding $\log_2$FC between FOXA1 HE and REST samples.

Together, these findings define the FOXA1-mediated splicing code where trans-acting splicing factors, particularly U2AF2 and HNRNPK, controls exon inclusion.

Genomic features depict an exon definition model for FOXA1-regulated CEs

To gain insights into FOXA1-mediated calibration of AS, we initially sought to delineate the features of exon definition. Compared to constitutive exons, cassette exons have weaker splice sites (ss), are strongly conserved during evolution, and are usually shorter with longer flanking introns (Keren, Lev-Maor, and Ast 2010; Mazin et al. 2021). By performing conventional splice strength analysis, we did not find any significant difference in 3'ss and 5'ss scores between FOXA1-regulated and FOXA1-unrelated exons of primary tumors (data not shown).

First, basewise conservation (PhyloP) data across 100 species were retrieved from UCSC (http://genome.ucsc.edu) (Kent et al. 2002). We then compared the single base distributions of PhyloP score between FOXA1-regulated and control CEs through a Wilcoxon Rank Sum Test and correct p-values using the Benjamini Hochberg method.

FOXA1-regulated CEs were significantly more conserved across 100 species than controls, with strongest enrichment of sequence conservation within 100nt of the exon/intron junctions (Figure 46).

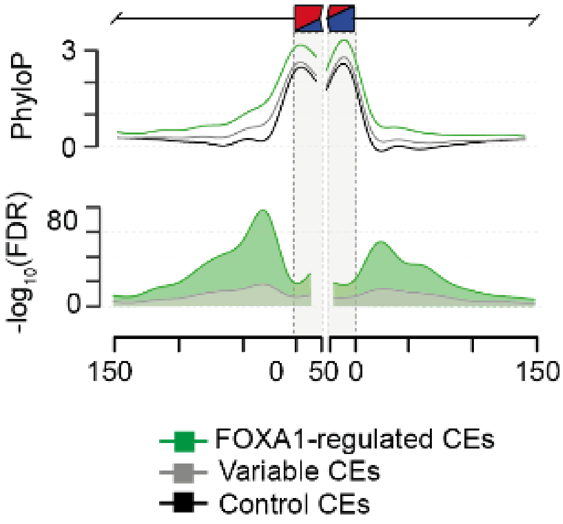


Figure 46. Upper panel shows the position-specific smoothed PhyloP conservation score of FOXA1-regulated (green), variable (gray) and remaining (controls, black) more-excluded and more-included CEs in exonic and intronic regions extending 50 and 150 nt from the splice sites, respectively. Bottom panel reports the

position-specific -$\log_{10}$(FDR) of Wilcoxon Rank Sum test comparing the PhyloP score distributions between FOXA1-regulated and control CEs (green), and FOXA1-regulated and variable CEs (gray).

Furthermore, compared to controls, FOXA1-regulated exons were significantly shorter, whereas their flanking introns were significantly longer (Figure 47). These features were more evident for FOXA1-regulated CEs in RBPs.

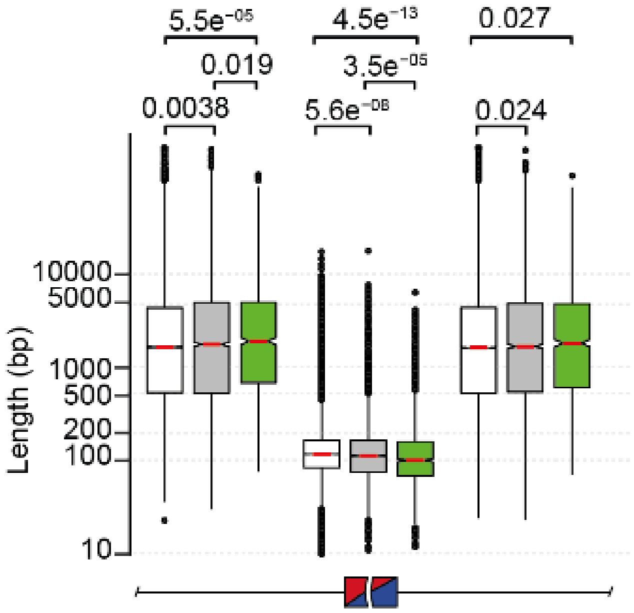


Figure 47. Box plots showing length distributions of the FOXA1-regulated (green), variable (gray) and control (white) CEs and their flanking introns.

Together these findings evidenced an exon definition model for FOXA1-regulated cassette exons identified by short exons with long introns that are under selective evolutionary pressure.

6. FOXA1 controls inclusion levels of nonsense-mediated decay cassette exons

So are all truths, once they are discovered; the point is in being able to discover them.

Galileo Galilei

Splicing factors self-tune the production of their isoforms

Splicing factors can auto- and cross-regulate their mRNAs by controlling the inclusion of exons that result in the introduction and prevention of a premature termination codon (PTC) in the reading frame. Introduction of a PTC triggers nonsense mediated decay (NMD) impacting isoform production (Pervouchine et al. 2019; Kurosaki, Popp, and Maquat 2019; Hug, Longman, and Cáceres 2016). From here on we will call them PTC-introducing and PTC-preventing exons (Figure 48).

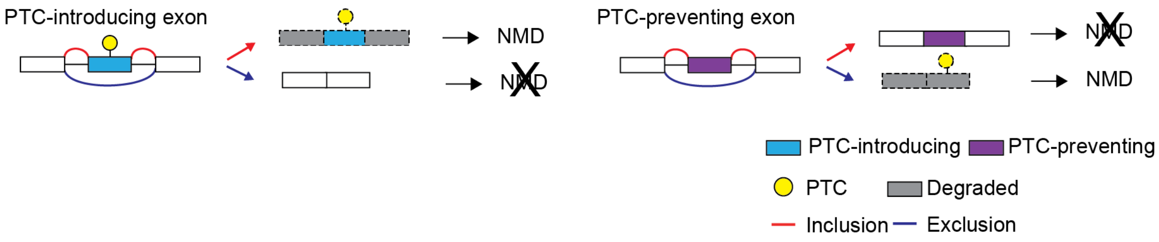


Figure 48. Schematic overview of the mechanisms of AS-controlled nonsense mediated decay (NMD) through the inclusion of PTC-introducing (left) or skipping of PTC-preventing (right) exons. PTC, premature termination codon.

FOXA1 modulates NMD-determinant exon inclusion levels

Since FOXA1 predominantly regulates exons in splicing-related genes, we sought to assess the regulation of NMD-determinant exons by the pioneer transcription factor.

Using a list of 15,518 NMD-determinant CEs (Pervouchine et al. 2019), we identified PTC-introducing and PTC-preventing CE events that are regulated by FOXA1 in primary PC. The UCSC custom track of annotated PTC-introducing and -preventing events available from this resource was obtained for comparative analysis. Genomic coordinates of these events were intersected with those of TCGA variable CE events using intersectBed command from BEDTools toolset (Quinlan and Hall 2010) and CEs were annotated accordingly.

We found a significant enrichment of PTC-introducing and PTC-preventing exons amongst FOXA1-regulated CEs relative to remaining CEs, both variable and constitutive (Figure 49).

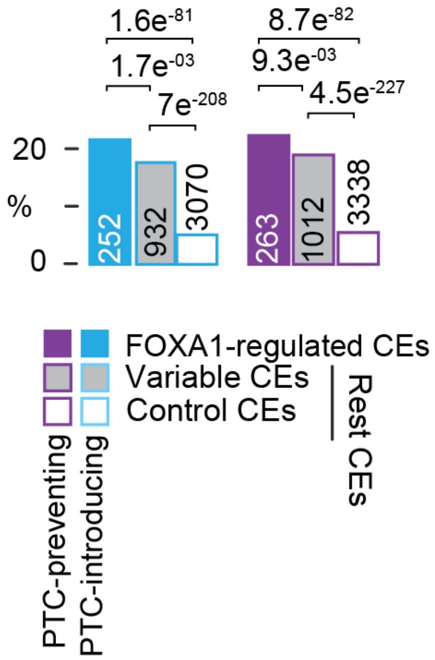


Figure 49. Bar plots representing the proportion of PTC-introducing (cyan) and PTC-preventing (purple) exons among FOXA1-regulated, variable and control CEs. Number above/within each bar gives the total number of CEs in each group.

Interestingly, FOXA1 more-excluded exons were significantly enriched for PTC-introducing exons and depleted of PTC-preventing exons relative to variable CEs that exhibited non-significant changes in inclusion levels in primary PC (Figure 50). Furthermore, more-included CEs were enriched for PTC-preventing exons relative to variable CEs (Figure 50).

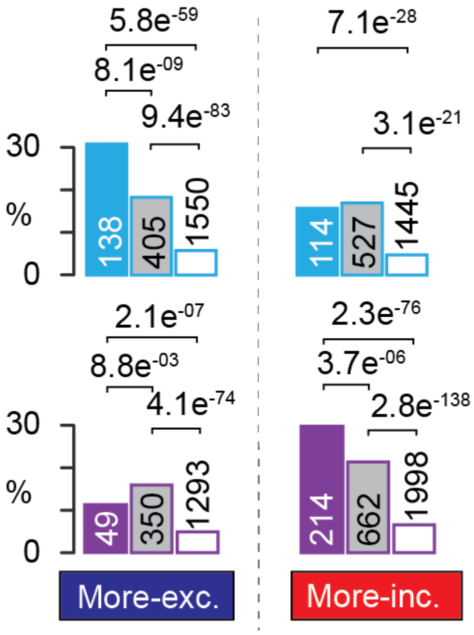


Figure 50. Bar plots representing the proportion of PTC-introducing (cyan) and PTC-preventing (purple) exons among FOXA1-regulated, variable and control CEs. Number above/within each bar gives the total number of CEs in each group. The proportion among more-excluded and more-included CEs is shown.

These results highlight a major effect of FOXA1 high expression in silencing PTC-introducing and enhancing PTC-preventing exon inclusion. By inspecting the Δμ(Ψ) distribution, we confirmed that FOXA1-regulated PTC-introducing and -preventing CEs were significantly more-excluded and more-included than variable CEs, respectively (Figure 51). Nevertheless, a small fraction of PTC-introducing CEs were more-included by FOXA1 high expression.

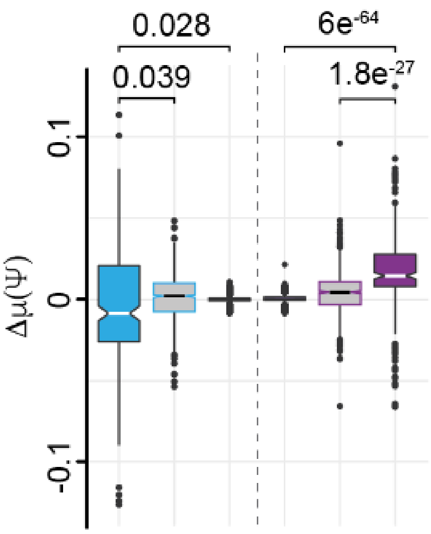


Figure 51. Boxplots showing the Δμ(Ψ) distribution of PTC-introducing and PTC-preventing FOXA1-regulated, variable and control CEs.

Consistent with previous reports (Pervouchine et al. 2019; Kurosaki, Popp, and Maquat 2019; Hug, Longman, and Cáceres 2016), over-representation analysis of genes harboring NMD-determinant CEs in primary PC revealed RBPs as the only significantly over-represented process amongst FOXA1-regulated PTC-introducing and -preventing exons (Figure 52), to a greater extent relative to the other CE groups.

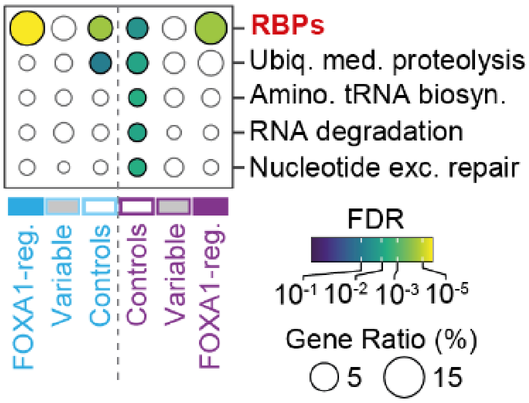


Figure 52. ORA results performed on genes harboring FOXA1-regulated, variable and control PTC-introducing and PTC-preventing CE events annotated.

Next, in light of recent evidence implicating PTC-introducing exons in lung cancer disease-free survival (DFS) (Thomas et al. 2020), we evaluated the clinical impact of the level of inclusion of PTC-introducing and PTC-preventing exons.

FOXA1 regulation of NMD-determinant exons impacts on prostate cancer patient survival

In light of recent evidence implicating PTC-introducing exons in lung cancer disease-free survival (DFS) (Thomas et al. 2020), we sought to investigate whether FOXA1 regulation of NMD-determinant exons could impact on primary PC recurrence.

Clinical data for 332 primary PC patients were obtained from the TCGA Data Matrix portal (Level 3, https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm). Disease-free survival (DFS) was defined as the length of time between primary treatment and the diagnosis of disease progression, as defined by biochemical or clinical recurrence, or the end of follow-up. PTC-introducing and -preventing FOXA1-regulated CEs were divided into silenced and enhanced events according to their $\Delta\mu(\Psi)$ sign upon FOXA1 HE, resulting into four groups (*i.e.* 141 silenced PTC-introducing, 125 enhanced PTC-introducing, 52 silenced PTC-preventing and 225 enhanced PTC-preventing FOXA1-regulated CEs). As previously proposed (Thomas et al. 2020), for each group and each patient, the following S statistic was computed:

$$S = n_{25} n_{75}$$

where n_{25} and n_{75} are the number of events with $\Psi \leq 25^{th}$ and $\geq 75^{th}$ percentiles, respectively, of their inclusion distribution across patients.

For each group of FOXA1-regulated exons, patients were stratified into high and low expressors based on the 25thand 75th percentile of the S statistics distribution, respectively. Exploiting this stratification, DFS analysis was performed by fitting a univariate Cox proportional hazards (PH) model with log-rank test ("Modeling Survival Data: Extending the Cox Model" n.d.).

Cox PH model revealed that a low cumulative inclusion of FOXA1-regulated more-excluded PTC-introducing exons was significantly associated with a longer patient DFS relative to high inclusion (Figure 53). Similarly, a high cumulative inclusion of more-included PTC-preventing CEs was significantly associated with a longer DFS than low cumulative inclusion (Figure 53).

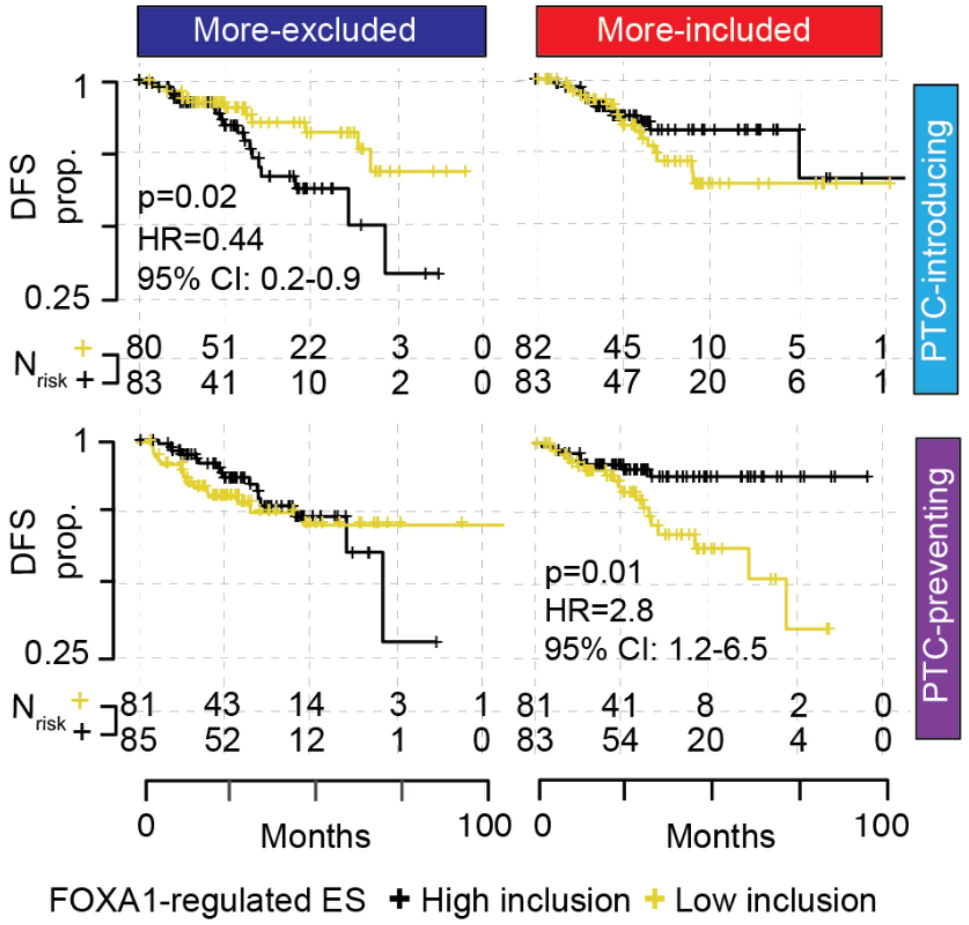


Figure 53. Disease-free survival of pri-PC patients, stratified according to the 25th and 75th percentile of the cumulative inclusion levels of PTC introducing or PTC-preventing exons that are more-excluded or more-included by FOXA1 HE. Univariate hazard ratios (HR) with 95% confidence intervals (CI) and two-tailed log-rank test p-values are shown when statistically significant (p-value≤0.05).

To assess the contribution of each PTC-introducing and -preventing FOXA1-regulated CE on DFS, patients were stratified according to the 25th and 75th percentiles of the Ψ level distribution of each event and DFS analysis was performed as described above. Log-rank test p-values were corrected for multiple testing with the Benjamini–Hochberg procedure. FOXA1-regulated CEs with FDR<0.05 were considered as significantly associated with DFS.

Out of 515 events, we found 85 NMD-determinant CEs with a significant DFS association (*i.e.* two-tailed log-rank test p-value<0.05). Out of the 39 exons with HR>1, 62% were more-excluded PTC-introducing events (Figure 54, top quadrants). Conversely, out of the 46 exons with HR<1, 60.9% were FOXA1-regulated “enhanced PTC-preventing” events (Figure 54, bottom quadrants). Eight FOXA1-regulated NMD-determinant CEs exhibited the strongest association with patient DFS (*i.e.* FDR<0.05), with the inclusion of exon 30 of the cancer gene FLNA identified as the top ranked event (FDR=0.002, HR=5.6, 95% CI: 2.4-12.7).

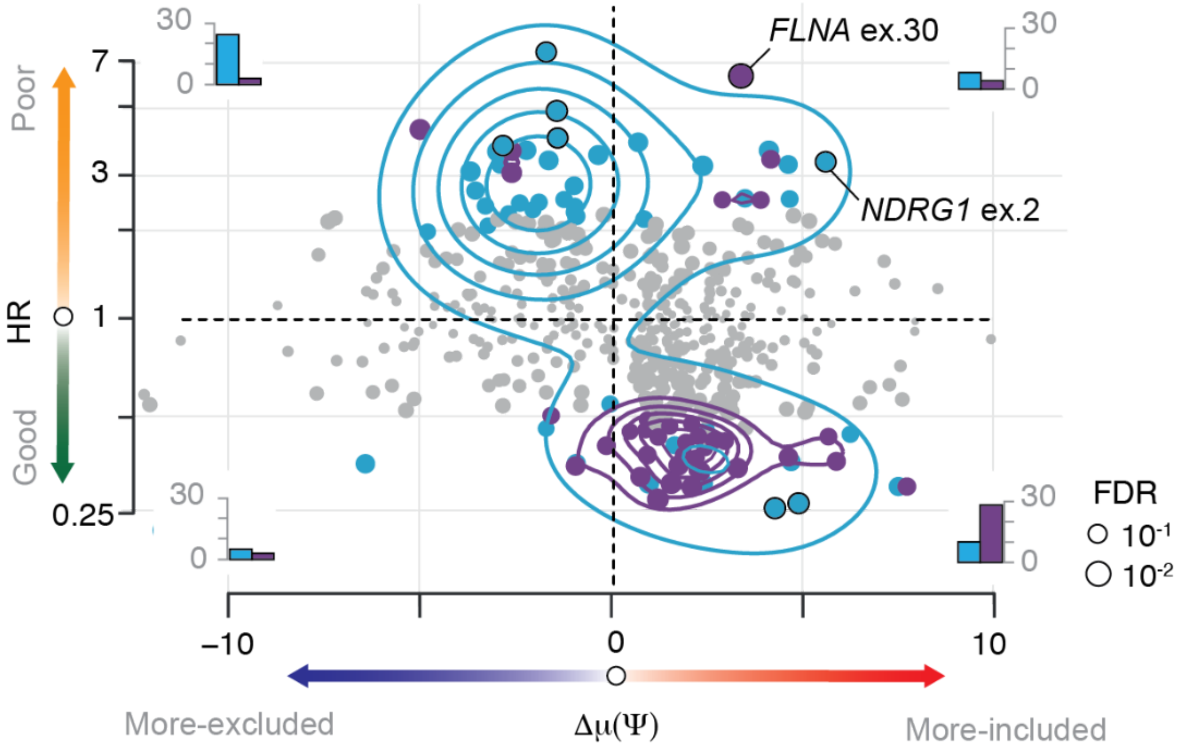


Figure 54. Scatter plot of univariate HRs of FOXA1-regulated PTC-introducing and PTC-preventing CE inclusion levels with respect to their Δμ(Ψ) upon FOXA1 HE. The size of dots represents the log-rank test FDR. Events with log-rank test p-value>0.05 are shown in gray. Contour lines indicate the density of points for PTC-introducing and PTC-preventing exons. Bar plots show the number of PTC-introducing and PTC-preventing exons with log-rank test p-value<0.05 in each quadrant of the scatter plot. CEs with log-rank test FDR<0.05 are indicated as text.

We then focused on the two survival-associated events whose inclusion, promoted by FOXA1, was deleterious for survival, FLNA ex. 30 and NDRG1 ex. 2 and we exploited their association with the FOXA1 expression using a linear model. Ψ levels of the eight survival-associated events and FOXA1 expression were normalized using near-zero variance filter, Yeo-Johnson transformation, centering around their mean, and scaling by their standard deviation implemented in the preProcess function in the R ‘caret’ package, with parameters method = c("center", "scale", "YeoJohnson", "nzv"). A GLM was fitted to FOXA1 expression based on the events Ψ levels. Relative importance of each event in the GLM was calculated using the averaging over ordering method (“Introduction to Bivariate and Multivariate

Analysis" n.d.). Confidence intervals of the regressor contributions were measured using a bootstrap procedure. For 1,000 iterations the full observation vectors were resampled and the regressor contributions were calculated.

FLNA exon 30 inclusion level was the strongest predictor of FOXA1 expression (Figure 55).

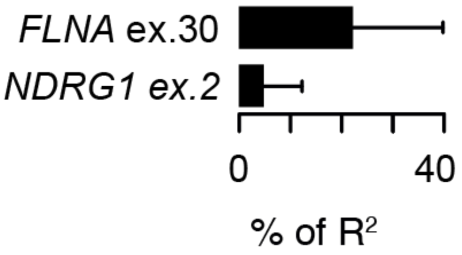


Figure 55. Relative importance of each survival-associated CEs to the R^2 measured on the GLM.

Overall, results of our cumulative and individual exon analyses reveal that FOXA1 predominantly silences PTC-introducing exons (75%) that are associated with a poor patient prognosis and enhances PTC-preventing exons (90%) that are associated with favorable outcomes. However, FOXA1 also enhances the inclusion of the essential FLNA exon 30 that is associated with shorter DFS, thereby underlying a more aggressive cancer phenotype. As so, we went deeper into the regulation of FLNA exon 30.

FOXA1-regulated exon 30 in FLNA promotes prostate cancer growth

To determine the impact of FLNA exon 30 on PC cell phenotypes, we transfected AR^- PC3 cells with ectopic expression vectors with and without exon 30 (*i.e.* FLNA+ex30 and FLNAΔex30, respectively), and confirmed exon 30 inclusion levels by in vitro splicing assays (Figure 56).

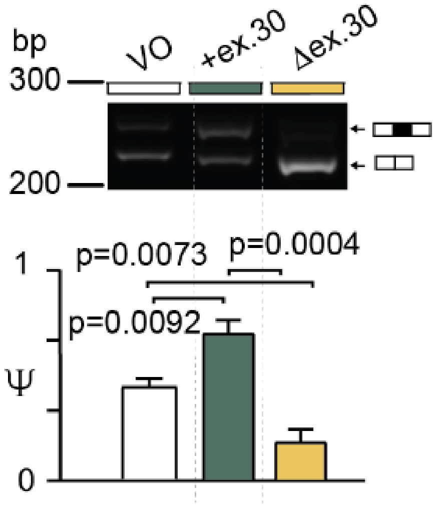


Figure 56. Ψ for *FLNA* exon 30 in PC3 cells was measured by splicing assays upon ectopic expression of *FLNA* with or without exon 30 (*i.e.* FLNA+ex30 or FLNAΔex30, respectively, or vector only (VO)).

Using a cell viability MTT assay, we observed a statistically significant increase in growth of cells overexpressing FLNA+ex30 than FLNAΔex30 (Figure 57). Furthermore, in a cell growth clonogenic assay, FLNA+ex30, but not FLNAΔex30, resulted in a statistically significant increase in colony number and staining intensity (Figure 57).

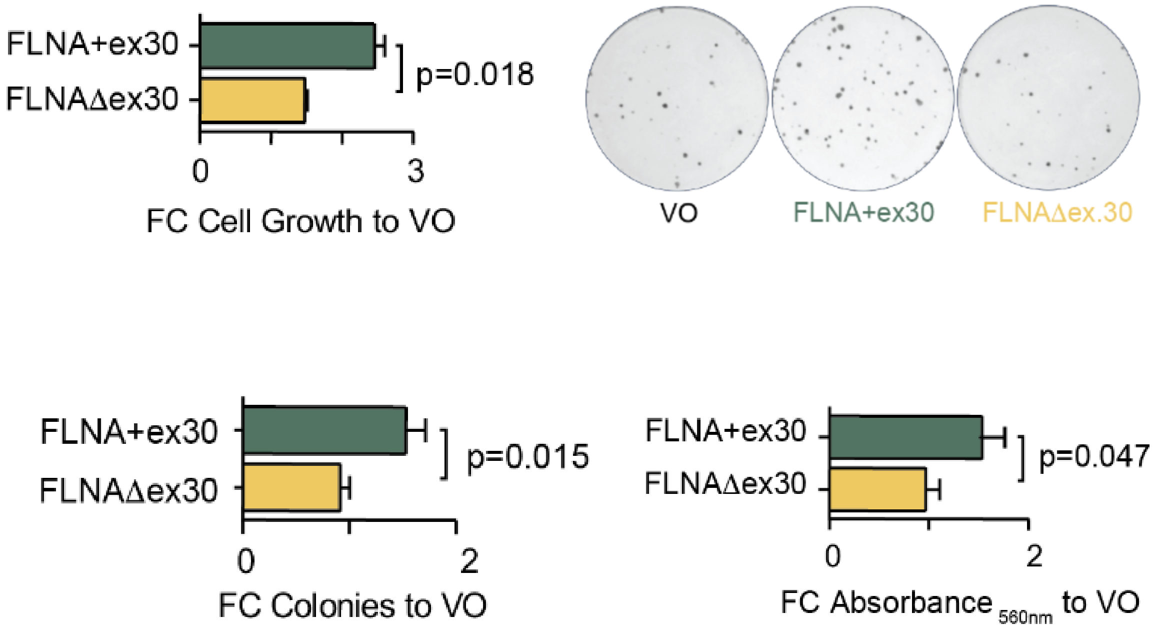


Figure 57. **Upper panel, Left.** PC3 cell growth was measured by MTT assay following ectopic expression of FLNA vectors. **Upper panel, Right.**PC3 clonogenic potential was measured by crystal violet assays following ectopic expression of FLNA+ex30 or FLNAΔex30. **Bottom panel, Left**. Colony number. **Bottom panel, Right.** Staining intensity.

Consistent with our DFS analysis, these functional data demonstrate that FLNA exon 30 inclusion provides a growth advantage to PC cells.

FLNA exon 30 inclusion is regulated by SRSF1 and HNRNPK

To determine putative regulators of FLNA exon 30 inclusion modulated by FOXA1, we employed a GLM to fit exon inclusion as a function of the expression levels of ten RBPs that were significantly and concordantly regulated by FOXA1 across primary PC. Using the averaging over ordering method on the GLM, we found that SRSF1 was the most important positive predictor of FLNA exon 30 inclusion (Figure 58), followed by HNRNPK.

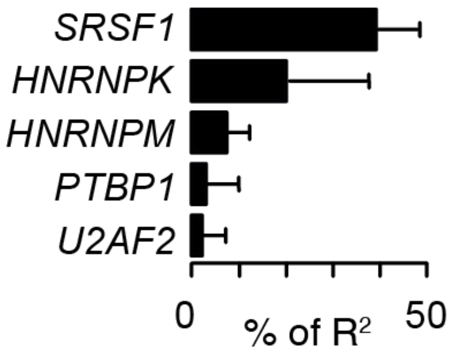


Figure 58. Relative importance of each RBP to the R^2 measured by the GLM fitting *FLNA exon 30* inclusion.

To validate the contribution of SRSF1 to FLNA exon 30 inclusion in the context of FOXA1, we stratified primary PC samples according to high and low expression of the two genes (*i.e.* 75th and 25th percentiles of normalized expression distributions, respectively). We found a statistically significant higher inclusion of the event in samples with high expression of both genes compared to other groups of samples (Figure 59, left panel). Similarly, stratification by HNRNPK and FOXA1 expression levels revealed a significantly greater inclusion of FLNA exon 30 in samples with high expression of both genes (Figure 59, right panel).

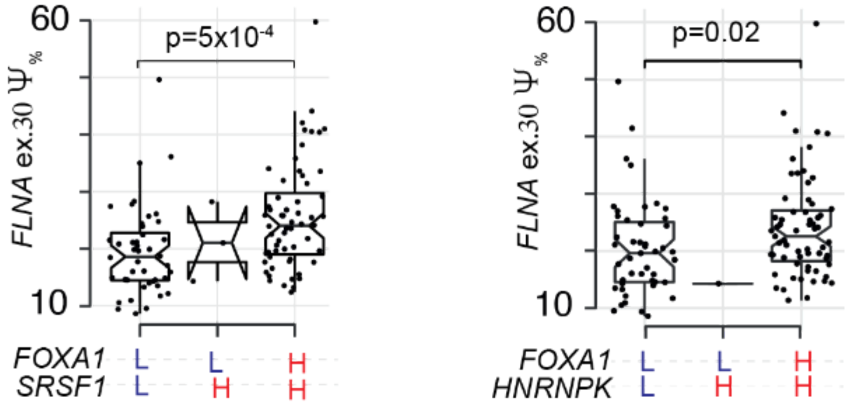


Figure 59. Boxplots displaying *FLNA exon 30* inclusion level in sets of patients stratified according to the high or low expression (above the 75th and below the 25th percentile, respectively) of *FOXA1* and *SRSF1* (**Left panel**) or *HNRNPK* (**Right panel**).

Using results from our integrated analysis of FOXA1 binding to chromatin-accessible regions in PC, we confirmed that FOXA1 binds to the promoter and associated active enhancers of SRSF1 (Figures 60,61), revealing a direct regulation of SRSF1 by FOXA1 similarly to HNRNPK (Figures 37,38).

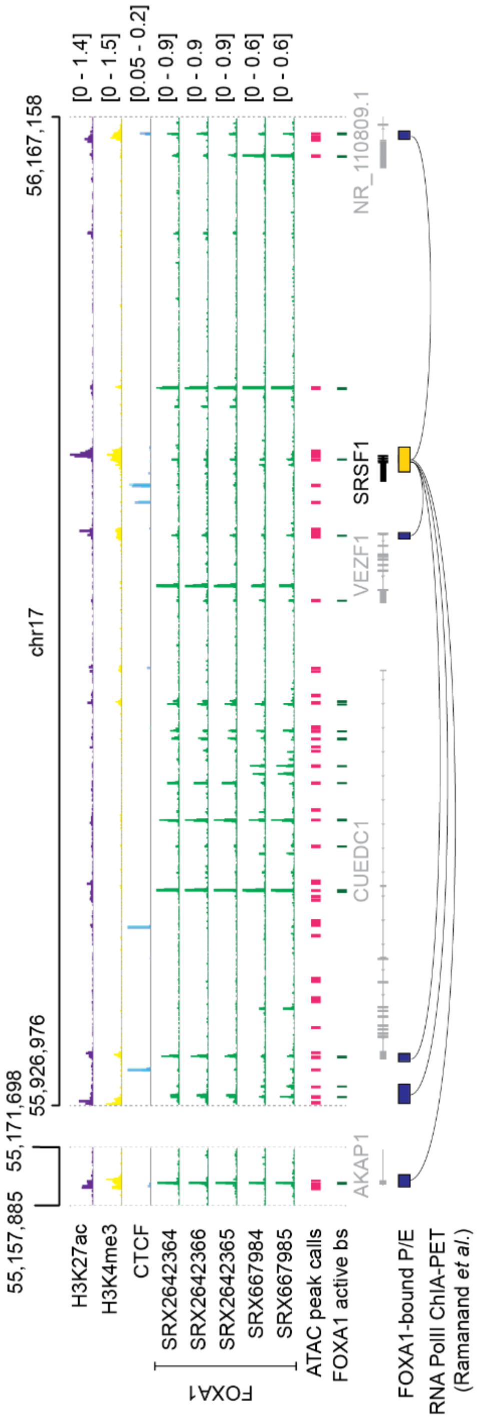


Figure 60. Genome view representing *SRSF1* promoter and enhancer regions that are bound by FOXA1 in VCaP cells. ChIP-seq density read tracks for H3K27ac, H3K4me3, CTCF (2 overlayed experiments) and FOXA1 in VCaP cells are shown together with pri-PC-reproducible ATAC-seq peak calls, FOXA1 active binding sites (bs) and RNA PolII ChIA-PET-derived FOXA1 bound enhancer-SRSF1 associations. The position of the considered SRSF1 regulatory regions along chromosome 17 is shown at the top of the panel.

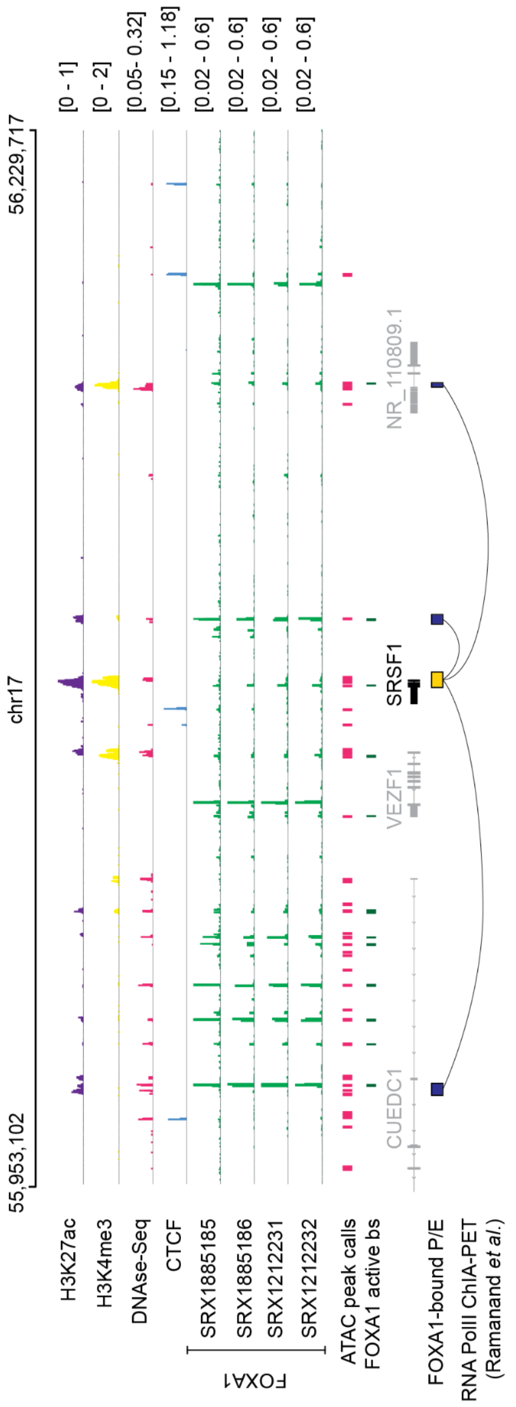


Figure 61. Genome view representing *SRSF1* promoter and enhancer regions that are bound by FOXA1 in LNaP cells. ChIP-seq density read tracks for H3K27ac, H3K4me3, DNAse-Seq, CTCF and FOXA1 in LNCaP cells are shown together with pri-PC-reproducible ATAC-seq peak calls, FOXA1 active binding sites (bs) and RNA PolII ChIA-PET-derived FOXA1 bound enhancer-SRSF1 associations. The position of the considered SRSF1 regulatory regions along chromosome 17 is shown at the top of the panel.

We further assessed SRSF1 and HNRNPK binding to FLNA exon 30 AS region using eCLIP data for the two proteins in HepG2 cells (Van Nostrand, Pratt, et al. 2020; Van Nostrand, Freese, et al. 2020). We observed clusters of eCLIP reads, and significantly cross-linked sites as detected by iCounts (Curk et al. (2019) iCount: protein-RNA interaction iCLIP data analysis (in preparation)) , within FLNA exon 30 and around 600nt in the downstream intron for SRSF1 and HNRNPK, respectively (Figure 62).

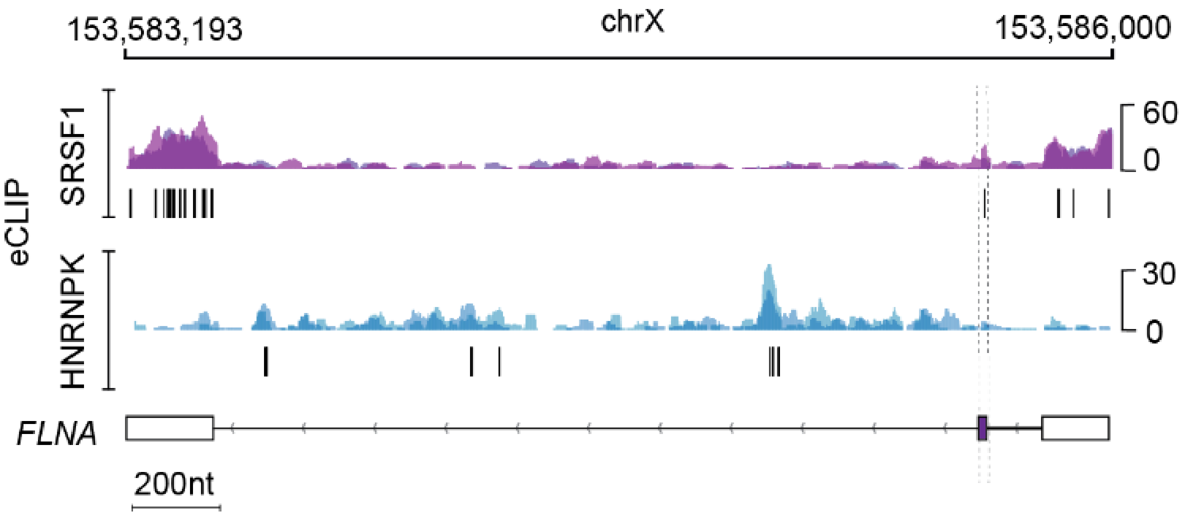


Figure 62. SRSF1 and HNRNPK eCLIP density read distribution in HepG2 cells in the AS region of *FLNA* exon 30. Significant cross-linked sites detected by iCounts of the two proteins are shown in black.

Importantly, this result supports the role of distal intronic binding of HNRNPK in exon inclusion (Figure 45).

To confirm the regulation of FLNA exon 30 by SRSF1 and HNRNPK, we performed siRNA-mediated depletion of SRSF1 and HNRNPK individually and in combination in PC3 cells (Figure 63).

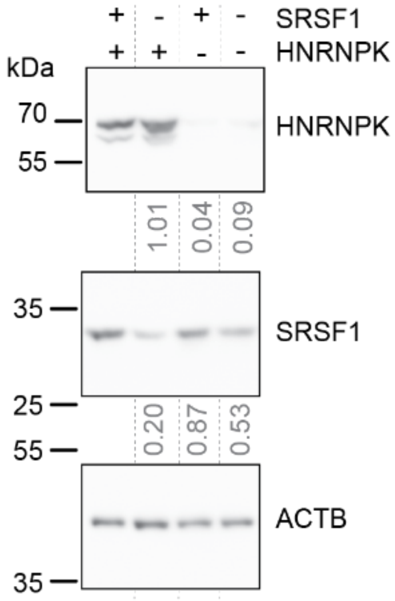


Figure 63. Representative Western blotting images of SRSF1 and HNRNPK depletion in PC3 cells by transfection with one siRNA duplex sequence per gene alone or in combination (40nM for 72 hours). Normalised (to ACTB) protein expression compared to control (NSI), calculated by densitometric band quantitation, is shown below SRSF1 and HNRNPK blot images.

Using splicing assays, we demonstrated a statistically significant decrease in FLNA exon 30 inclusion in individual and combined siRNA conditions compared with NSI controls, with SRSF1 depletion providing the strongest contribution (Figure 64).

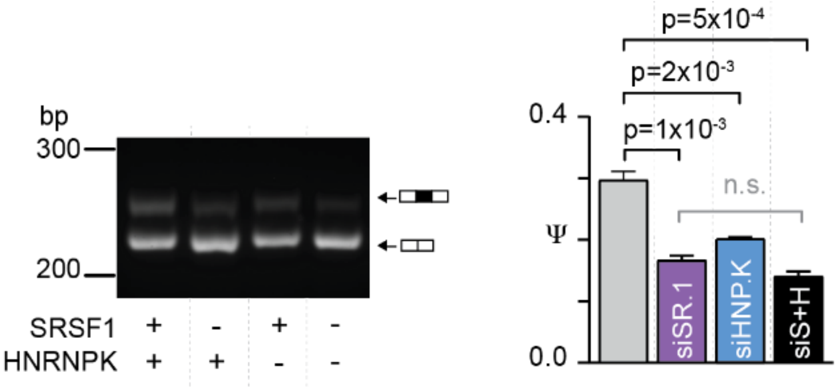


Figure 64. Ψ changes for *FLNA* exon 30 in PC3 cells were measured by splicing assays upon depletion of SRSF1 or HNRNPK or both. Two-tailed T-test was used to compare conditions.

Taken together, these findings clearly demonstrate that FLNA exon 30 inclusion is controlled by SRSF1 and HNRNPK, which are directly regulated by FOXA1 (Figures 37,38,60,61). The increased inclusion of FLNA exon 30 confers a growth advantage to PC cells and is associated with a shorter patient DFS.

7. Discussion and Conclusion

Never discourage anyone...who continually makes progress, no matter how slow. Plato

The expression of splicing factors is altered in the majority of cancer types

In this study, we aimed at disentangling AS deregulation across cancer types focusing specifically on the role of transcription factors in modulating the expression of splicing factors. To do so, we first developed a new algorithm to account for the heterogeneity proper of cancer transcriptomes, particularly in presence of high-volume datasets, and extract the relevant biological information. Heterogeneity is a fundamental characteristic of information associated with complex traits, which can arise from subtle deregulation of distinct genes in different patients rather than of a single gene (Cereda et al. 2016; A. Subramanian et al. 2005; Gambardella et al. 2017). In diseases such as cancer, heterogeneity strongly impacts on its progression and drug response (Cereda et al. 2016; A. Subramanian et al. 2005; Gambardella et al. 2017). Therefore, dissecting the contribution of heterogeneity becomes crucial to detect defective biological processes and to the therapy management of patients. Thus, we addressed this challenge and introduced the novel concept of discretization of gene expression levels, which we derived from probabilistic modeling and shaped upon knowledge of RNA biology. Our Gene Set Enrichment Class Analysis (GSECA) algorithm exploits the bimodal behavior of RNA-sequencing gene expression profiles to identify altered gene sets in heterogeneous patient cohorts. We showed that GSECA outperformed 'state-of-art' algorithms in handling gene sets characterized by expression changes of groups of genes that are more intensively activated or repressed in a heterogeneous manner across samples. GSECA can detect functionally related altered cell mechanisms in a condition of interest considering more heterogeneous cohorts as compared to other available methods. By boosting signal-to-noise ratio, GSECA can successfully manage the heterogeneity of thousands of samples and provides useful insights on clinical and biological patterns proper of a phenotype. With this algorithm we introduced the paradigm shift of "less is more" in treating large heterogeneous RNA-seq datasets showing that it improves the detection of the altered biological processes in the phenotype of interest.

We therefore applied the “less is more” paradigm to assess the dysregulation of RBP expression in different cancer types. GSECA detected the alteration of the RBP-related pathway at a high confidence in thirteen cancer types, suggesting that not only alternative splicing itself but also the expression levels of RBP is a hallmark of cancers. Specifically, Head and Neck Squamous Carcinoma (HNSC), Breast Cancer (BRCA), Colon Adenocarcinoma (COAD), and Prostate Adenocarcinoma (PRAD) showed the strongest deregulation of the RBP gene set. Notably, the prognostic role of splicing has been recently proposed in HNSC (Ding, Feng, and Yang 2020) and COAD (Y. Chen et al. 2021), while in BRCA aberrant expression of splicing factors leads to the proliferation of somatic cells (Cieśla et al. 2021). Similarly, deregulated expression of RBP expression leads to alternative pre-mRNA in aggressive prostate cancer (Phillips et al. 2020). Conversely, Bladder Urothelial Carcinoma (BLCA), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), Kidney Renal Papillary Cell Carcinoma (KIRP) and Kidney Renal Clear Cell Carcinoma (KIRC) did not show any enrichment. Overall, Cholangiocarcinoma (CHOL) and Kidney Chromophobe (KICH) presented a limited number of enriched pathways, with no enrichment for RBP deregulation. Nevertheless, these results can be imputed to the low number of samples available for these cancer types.

The analysis of discretized expression levels of genes in the RBP gene set across tumors revealed a pattern of somatic rewiring of splicing factors' expression. We found that the number of tumors expressing RBPs at highest levels was significantly reduced compared to normals, whereas those expressing RBPs at the intermediate levels significantly increased. This dichotomous effect may reflect the inter- and intra- sample heterogeneity of tumors, where different tumors may aberrantly activate different splicing factors through specific transcriptional and epigenetic programs. Different cancer types showed a distinct extent of this effect, with COAD, Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC) exhibiting the strongest patterns. In lung cancer, changes in expression of RBPs have been identified, as well as chances in splicing patterns, suggesting splicing as a possible therapeutic target (Coomer et al. 2019).

Our unbiased pan-cancer gene set enrichment analysis identified the disruption of RBP expression as a cancer hallmark of the majority of tumor types. Furthermore, our results revealed an heterogeneous utilization of splicing factors that are generally expressed at intermediate levels of expression across tumors, underlying specific transcriptional programs in different subsets of samples.

A subset of transcription factors controls splicing factors' expression across cancer types

As a systematic pan-cancer analysis of transcription factor regulation on RBPs was still lacking, we developed a pipeline for integrating both DNA and RNA levels, namely active binding sites and expression profiles. Combining different layers of information, provenient from different omic data, has been shown to be fundamental for understanding biological processes (I. Subramanian et al. 2020; Planell et al. 2021; Vasaikar et al. 2018; Hawe, Theis, and Heinig 2019). As a novelty in literature, we filtered cell line ChIP-seq data with ATAC-seq peaks from their corresponding tumor, as our main goal was to exploit mechanisms in patients and, currently, no ChIP-seq in vivo is available. This pan-cancer analysis was possible thanks to the recent years increase of available data, a phenomenon known as Big Data era, that allows the creation of databases whose data are shared with the scientific community (Del Giudice et al. 2021).

RBP were enriched for active binding sites of MYC, MAX, and FOXA1 on their promotorial regions in the greatest number of cell lines. In particular, MYC and MAX resulted as active transcriptional regulators of RBPs in the same cell lines, confirming their cooperative work as heterodimer (Arsura et al. 1995; Bissonnette et al. 1994; Walhout et al. 1997; Nair and Burley 2003; Grandori et al. 1996). Interestingly, in most of these cell lines, FOXA1 also resulted as an active transcriptional regulator of RBPs. These results may underline a possible concomitant involvement of the three proteins in the transcriptional regulation of splicing factors.

To evaluate whether the enrichment of a TF's active binding on RBP promoters over the rest of genes was reflected in expression changes, we modeled the cumulative expression of RBPs through the contribution of transcription factors using a generalized linear model. Regression models have become the most used method for feature selection, with the advent of artificial intelligence in RNA-seq data analyses (Del Giudice et al. 2021). In accordance with the DNA level analysis we found that MYC and FOXA1 were the most significant TFs.

In recent years, the relation between MYC and RBPs has been widely studied. In particular, it has been shown that MYC upregulates components of the spliceosome in promoting lymphomagenesis (Koh et al. 2015) and is often associated with RBP aberrant expression (David et al. 2010). Moreover, MYC has been related to RBP expression in triple negative breast cancer (Cieśla et al. 2021) and prostate cancer and to the inclusion level of more than a thousand cassette exons, many of which are nonsense-mediated decay-determinant exons, also in genes encoding RNA binding proteins (Phillips et al. 2020). As the direct targeting of MYC has been clinically unsuccessful, targeting its downstream effector pathways seems a valid opportunity. Given the strict relation between MYC and splicing in promoting cancer invasiveness, it has been recently proposed that methods targeting RBPs or their downstream splicing targets may be a potential avenue for treatment (Urbanski et al. 2021).

FOXA1 has been recently proposed as a splicing factors' regulator (Foster et al. 2018), nevertheless its role was not entirely defined. FOXA1 exhibits a common pioneer function for AR and MYC-driven PC transcriptional programs, occupying the majority of binding regions shared by AR and MYC (Barfeld et al. 2017). An increased FOXA1 co-occupancy at AR binding sites characterizes MYC-induced suppression of AR transcriptional programs, which in turn accelerates PC progression towards CRPC (X. Qiu et al. 2021). This evidence may suggest a possible relationship between FOXA1 and MYC in regulating RBP expression.

Our multi-omic analysis provided the tissue-specific landscape of transcription factor regulation on RPB expression. Furthermore, we identified MYC and FOXA1 as the major regulators of RBPs across cancer types.

FOXA1 orchestrates alternative splicing dysregulation in prostate cancer

We revealed that the pioneer transcription factor FOXA1 orchestrates alternative splicing regulation in PC, impacting on gene expression control mechanisms control and disease recurrence risk.

Collectively, our results, in primary PC and mCRPC patients, indicate that, among the key PC TFs tested, FOXA1 expression is a predominant hallmark of the transcriptional dysregulation of splicing-related genes. By defining the architectural features of FOXA1 transcriptional control in PC, we showed that this pioneer factor preferentially binds chromatin-accessible active enhancers of RBPs. As a pioneer factor, FOXA1 accesses compact chromatin and opens up nucleosomal domains for DNA binding of distinct transcription factors controlling the expression of nearby essential genes (Fei et al. 2019; Lupien et al. 2008). Therefore, FOXA1 may open multiple channels to transmit transcriptional signals to RBP loci. The preferential active binding of FOXA1 to regulatory regions of splicing factors, rather than genes underpinning other fundamental biological processes, may implicate the recruitment of different transcription factors to achieve a comprehensive alternative splicing regulation.

We assessed the AS changes in primary PC and by the analysis of multivalent RNA motifs and we generated a RBP-mRNA-interaction atlas that defines the position-dependent splicing regulation of cassette exons driven by FOXA1 high expression. Our results recapitulate many known RBP regulatory principles at exon/intron junctions and support the coordinate regulation of CEs by multiple proteins. FOXA1-regulated exons are defined by T- and C-rich motifs at 3’ss and 5’ss, respectively, extending hundreds of nucleotides within introns, which work as intronic splicing enhancers and silencers (Murray et al. 2008). By integrating results of ENCODE experiments, such as assessment of RBP cross-linking sites by eCLIP and splicing changes by RNA-seq upon RBP depletion, we confirmed coordinate splicing control by FOXA1-regulated RBPs, highlighting HNRNPK and U2AF2 as major players. RNA splicing maps validate the canonical binding of these two RBPs proximal to FOXA1-regulated exons (Van Nostrand, Freese, et al. 2020; Briese et al. 2019). Interestingly, HNRNPK exhibits a distinctive footprint hundreds of nucleotides (~600) downstream of FOXA1-regulated exons. The enrichment for HNRNPK cross-linking sites at both more-excluded and more-included exons support its known bidirectional splicing regulatory effect (Venables et al. 2009), promoting (Expert-Bezançon et al. 2004) or inhibiting (Marchand et al. 2011) exon inclusion. By binding downstream the 5’ ss HNRNPK can interact with components of the U1 snRNP and modulate AS (Hegele et al. 2012; Thompson et al. 2018). Similarly, our analyses revealed that U2AF2 can cooperate with HNRNPK and SF3A3 to silence and enhance the inclusion of sets of FOXA1-regulated exons, respectively. It has been shown that auxiliary RBPs can stabilize or clear U2AF2 binding in vivo at 3′ss and control splicing decisions (Sutandy et al. 2018). U2AF2 competes with HNRNPs to bind the Py-tract (Sutandy et al. 2018; Zarnack et al. 2013), whereas it cooperates with SF3A3 in branch point recognition for exon definition (Briese et al. 2019). Taken together, it is tempting to speculate that FOXA1-regulated RBPs orchestrate the binding of core spliceosomal components, such as U2 and U1 snRNPs, by competing or cooperating with them to silence or enhance exon inclusion, respectively.

We showed that exons responding to FOXA1 expression are shorter with longer flanking introns that maintain a greater sequence conservation than the rest of CEs. A smaller exon size and higher intronic sequence conservation have been associated with a higher exon exclusion to inclusion rate, under evolutionary constraints, to control relative isoform frequencies (Baek and Green 2005). By integrating analyses of cis-acting elements and trans-acting factors, we demonstrated that FOXA1 calibrates AS by enlisting splicing factors under its transcriptional control. Amongst all FOXA1-controlled factors, HNRNPK emerges as the major mediator of FOXA1-induced alternative splicing.

FOXA1-mediated calibration of AS preferentially impacts on splicing genes, evidencing a regulatory feedback loop involving splicing factors (Müller-McNicoll et al. 2019). We showed that FOXA1 affects the inclusion of NMD-determinant exons, with the strongest effect on silencing of poison exons, reinforcing the hypothesis of TF-driven auto- and/or cross-regulation of RNA splicing networks in PC (Munkley et al. 2019; Phillips et al. 2020; K. Shah et al. 2020). By performing a survival analysis of FOXA1-regulated NMD-determinant exons in PC patients, we showed that either low or high cumulative inclusion of more-excluded poison or more-included essential CEs, respectively, are associated with favorable patient prognosis. This observation is consistent with the role of NMD-determinant exons as tumor suppressors, mediators of cell viability, and patient outcomes (Thomas et al. 2020). Importantly, FOXA1-mediated splicing regulation leads to an increased inclusion of a subset NMD-determinant exons that are strongly associated with disease recurrence (*e.g.* NDGR1 and FLNA). We demonstrated in vitro that the FOXA1 more-included essential exon 30 in the cancer gene FLNA promotes PC cell growth. Combining statistical modeling of FOXA1-regulated RBPs, integration of eCLIP data, sequence motif analysis, and in vitro validation, we showed that the inclusion of FLNA exon 30 is controlled primarily by the oncogenic SRSF1 and could be influenced by HNRNPK. Similarly to HNRNPK, FOXA1 controls SRSF1 expression by physically interacting with its promoter. Interestingly, we showed that HNRNPK binds at ~600nt downstream the 5'ss of FLNA exon 30, exemplifying the putative role of HNRNPK distal binding in AS regulation. HNRNPK and SRSF1 have recently been shown to have an evolutionary conserved role in intron retention regulation in B-cell development (Ullrich and Guigó 2020). Therefore, together with our results, it is tempting to speculate that FOXA1-mediated splicing regulation of NMD-determinant exons, and more generally of exons, may underlie evolutionary conserved constraints of mRNA isoform production.

In summary, we assessed the dysregulation of RBPs at a pancancer level, proposing transcription factors MYC and FOXA1 as the major regulators of splicing factor expression. In particular, focusing on prostate cancer, we revealed a novel role for the pioneer transcription factor FOXA1 in orchestrating alternative splicing regulation in PC at different stages of gene expression. FOXA1 controls the expression of splicing-related genes, including HNRNPK, by binding to their promoters or interacting enhancer regions. Moreover, a link between FOXA1 and the inclusion of NMD-determinant exons has been found.

Further functional studies are necessary to determine whether FOXA1 cooperates with specific TFs, such as MYC, chromatin modifiers, and RNA Polymerase II, to rewire the alternative splicing landscape of PC, and determine whether targeting splicing is a therapeutic vulnerability for FOXA1-driven tumors.

Appendix A. Gene Set Enrichment Class Analysis

The details are not details. They make the design.

Charles Eames

We developed GSECA (Lauria et al. 2020) as described in chapter 2 Also, GSECA was compared with other previously published Gene set enrichment analysis algorithms (*i.e.* GSVA (Hänzelmann, Castelo, and Guinney 2013), Z-score (E. Lee et al. 2008), PLAGE (Tomfohr, Lu, and Kepler 2005), ssGSEA (Barbie et al. 2009), Globaltest (Goeman et al. 2004), ROAST (Wu et al. 2010) and GSEA (A. Subramanian et al. 2005)), designed to manage microarray data but widely used also on RNAseq data. To test GSECA performance with respect to the other methods we firstly used simulated data. In particular, to measure the type I error rate, we generated read counts for N samples and 1,000 gene sets of equal size P in the condition of no differential expression and we tested the null hypothesis of no difference between the two cohorts. We tested different combinations of parameters: N (60, 150, 300, 500) and P (25, 50, 100, 300), repeating the analysis 10 times to obtain more stable results. GSECA showed the lowest type I error rate. The conservativeness of our method is due to the intrinsic conservativeness of Fisher's Exact Test (FET). As the sample size grows, the ability to detect small variations increases, so GSECA is able to account for false positives better than the other algorithms. Moreover, GSECA specificity was not influenced by sample and gene set sizes (Figure 65).

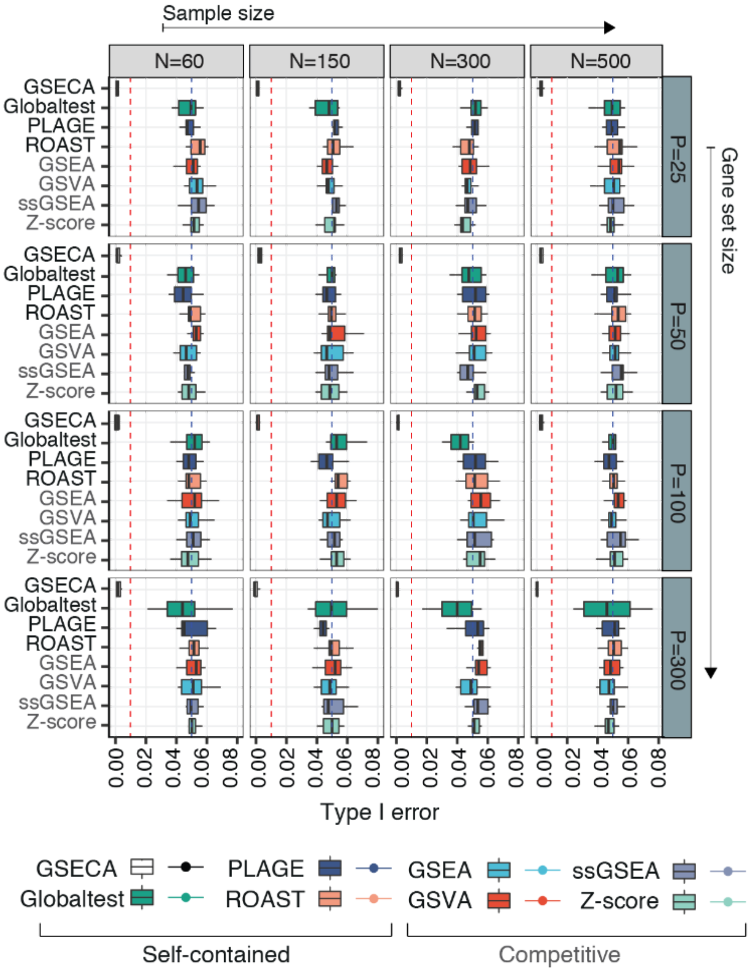


Figure 65. From (Lauria et al. 2020), Boxplots depicting the type I error rates for GSA methods evaluated for different settings of sample and gene set sizes on ten replicates. Red and blue dashed lines show the nominal values of 0.01 and 0.05, respectively.

Then we also tested the statistical power of each algorithm, performing two independent simulations, named "FC" and "dispersion" studies, modeling the contribution of fold and dispersion variations, respectively, in gene expression between the two cohorts. Three different parameters were taken under consideration: the proportion of gene sets that contain differentially expressed (DE) genes β, the percentage of DE genes in each gene set γ and the FC in gene counts between the two cohorts. Also, a scaling factor D was introduced. For both FC and dispersion studies we modeled eight conditions where, out of 1000 gene sets, the fraction of true positives was equal to the 5% and 25% and the percentage of DE genes in each gene set was set at 25% and 50% for gene set sizes of 25 and 100. Moreover, for the FC study, fold changes ranging from 1.5 and 3 were selected, without changes in dispersion value (D=1). For dispersion study, instead, D varied between 1.5 and 10 with no or low values of FC (FC={1,1.1,1.25}, modeling expression changes due to inter-sample heterogeneity. So, we generated expression values for N samples divided in two groups. To take into account specificity of gene set enrichment methods we introduced the calculation of the F1 score, a performance evaluation metric that provides a harmonic mean of the precision and sensitivity in case of an uneven distribution of true and false positives for all simulations. Regardless of gene set size and changes in percentage of the DE genes in the gene sets the statistical power of GSECA increased with sample sizes and FC (Figure 66). For medium and high sample sizes GSECA reached high statistical power, while for little ones other methods outperformed it. Nevertheless, always GSECA showed a better tradeoff between precision and sensitivity (F1 > 0.7), reflecting the high specificity of GSECA in detecting truly altered gene sets.

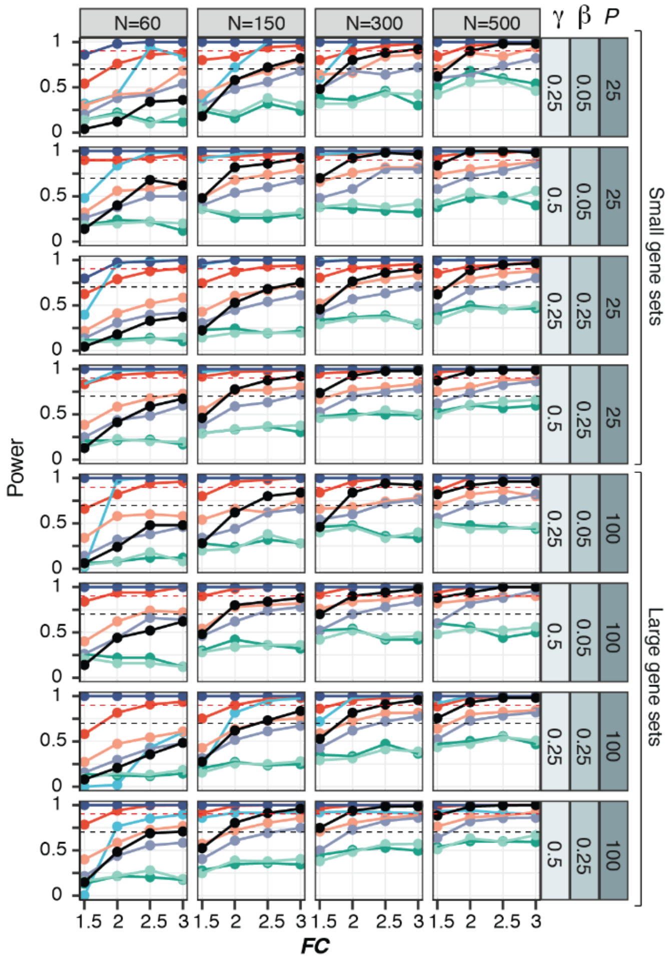


Figure 66. From (Lauria et al. 2020), Scatter plots depicting the statistical power of each GSA algorithm at the increase of FC between cohorts for different settings of sample size N, gene set size P, the proportion of gene sets containing differentially expressed genes β, the percentage of DE genes in each gene set γ.

In case of inter-sample heterogeneity, the statistical power of GSECA grew exponentially with the dispersion parameter, outperforming other methods (Figure 67). Even at a lesser extent, ssGSEA performed similarly to GSECA in handling heterogeneity, thanks to its collapsing of gene expression to common scale that reduces heterogeneity.

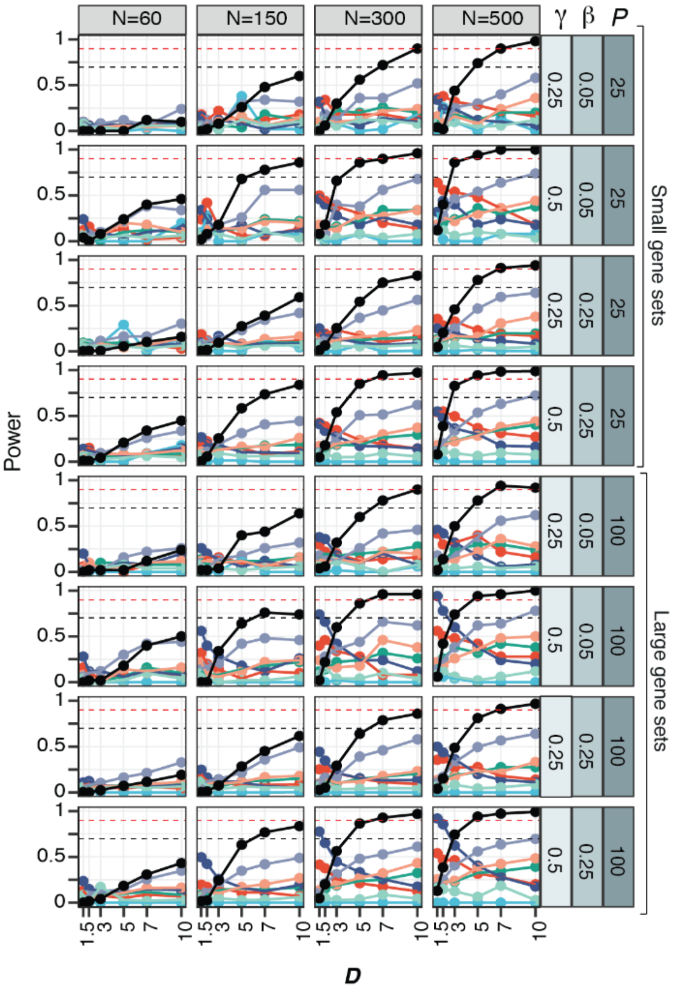


Figure 67. From (Lauria et al. 2020), Scatter plots depicting the statistical power of each GSA algorithm at the increase of dispersion factor D at a fixed FC of 1.1 between cohorts for different settings of N, P, β, and γ.

GSECA achieved the highest F1 scores, showing high sensitivity and specificity in deeply heterogeneous gene expression profiles, regardless of FC and dispersion (Figure 68).

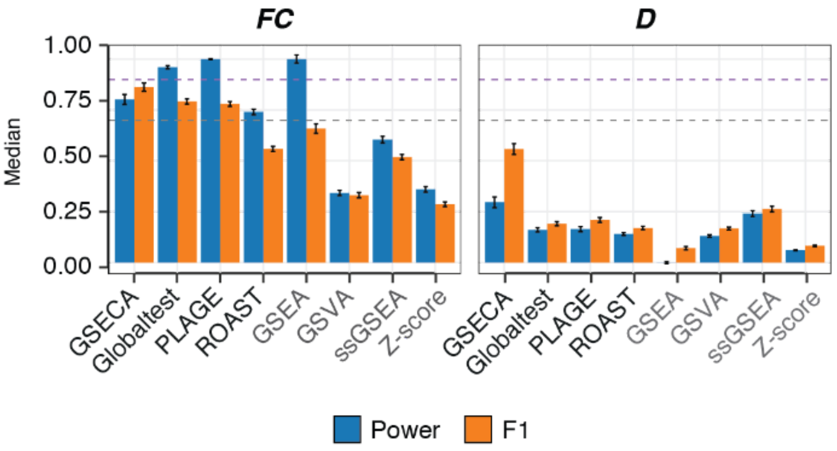


Figure 68. From (Lauria et al. 2020), Bar plots representing the median values of statistical power and F1 score measured for all GSA methods in all simulations for the FC and dispersion studies. Gray and purple dashed lines represent values of 0.7 and 0.9, respectively. Black error bars depict standard errors.

We also tested GSECA on real samples. In doing so, we focused on Prostate Adenocarcinoma samples. Somatic mutation, RNA-seq expression, protein expression and phosphorylation data for 498 samples of prostate adenocarcinoma (PRAD) samples were downloaded from TCGA Data matrix portal (Level 3, https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm). PTEN was considered as somatically lost if undergoing homozygous/heterozygous gene deletions, truncating mutations or damaging mutations. 158 manually curated gene sets from the Kyoto Encyclopedia of Genes and Genomes (KEGG) available from MSigDb35 (version 5, https://software.broadinstitute.org/gsea/msigdb/) were used. From literature we know that the loss of PTEN brings to the alteration of PI3K/AKT signaling pathway that involves nine genes according to KEGG. First, a differential expression analysis was performed between the PTEN-loss and the PTEN-wt cohorts. We tested the significantly lower expression of PTEN in PTEN-loss samples and also the alteration in the expression of the PI3K/AKT pathway genes. Also, we tested the higher phosphorylation level of AKT1 in PTEN-loss tumors. Looking at PTEN-loss and wild type samples, we evaluated the log fold change and the dispersion of data, along with density distributions (Figure 69).

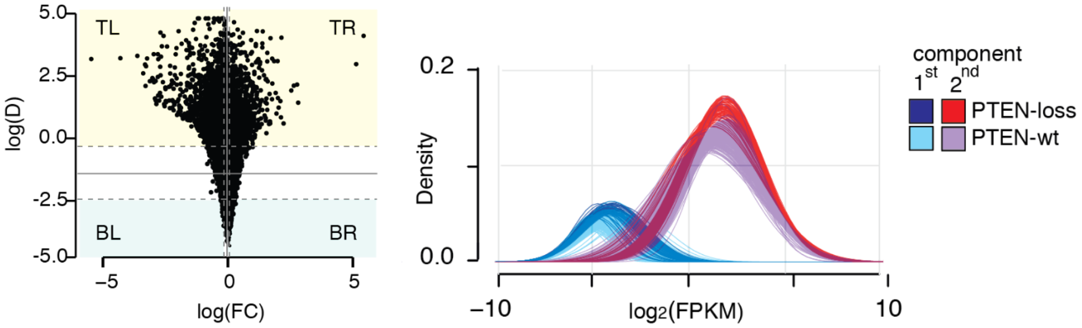


Figure 69. From (Lauria et al. 2020), **Left panel.** Scatter plot showing the log2 fold change (FC) and dispersion (D) values of all genes between PTEN-loss and PTEN-wt samples. Gray lines represent the median values of FC and D. Dashed gray lines show the 25th and 75th percentile of the FC and D distributions and define four regions of expression changes (TL=top-left; TR=top-right; BL=bottom-left; BR=bottom-right). **Right panel.** Kernel density distributions of FPKM values for PTEN-loss and PTEN-wt samples.

Then, we run GSECA between the PTEN-loss and the PTEN-wt cohort. The fraction of genes in the gene sets reflected the landscape of FC and dispersion values of the cohorts. We found 21 significantly altered (AS≤0.01, Pemp≤0.001 and SR≥0.9) signaling pathways. Among them the second top-ranked was the PI3K/AKT signaling pathway, showing a significant increase in the number of samples expressing genes in the LE, ME and HE1 and a significant decrease in the HE2 and HE3 classes (FDR<0.1), supporting the presence of high inter-sample heterogeneity of PI3K/AKT genes at low level of expression (Figure 70).

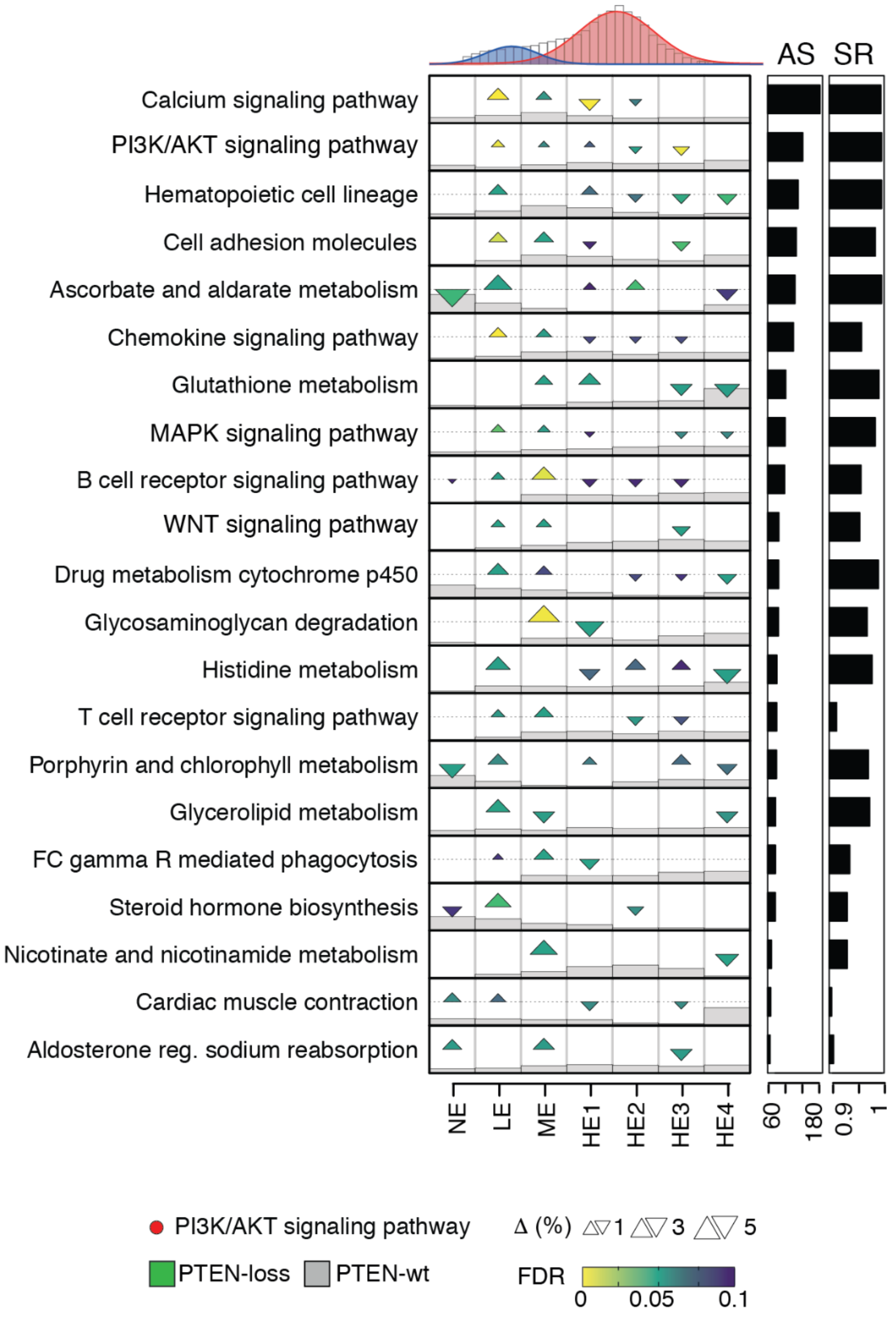


Figure 70. From (Lauria et al. 2020), EC map for the AGSs identified by GSECA in PTEN-loss prostate adenocarcinoma.

We next compared GSECA with other tools on the prostate dataset. Z-Score, PLAGE, ssGSEA identified about 20 enriched gene sets each, comparably with GSECA, while GSVA selected 41 altered sets and ROAST and GSEA only one (adjusted P-value < 0.1, SR >0.9). We computed the Jaccard index (JC) to assess the concordance of results (Figure 71). Similarity between tools resulted generally low (mean JC = 12%).

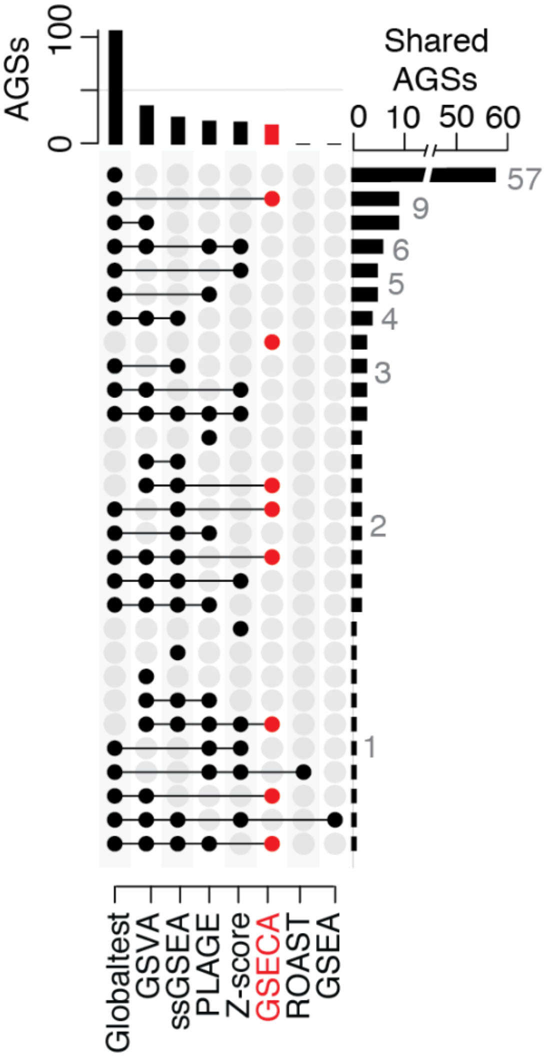
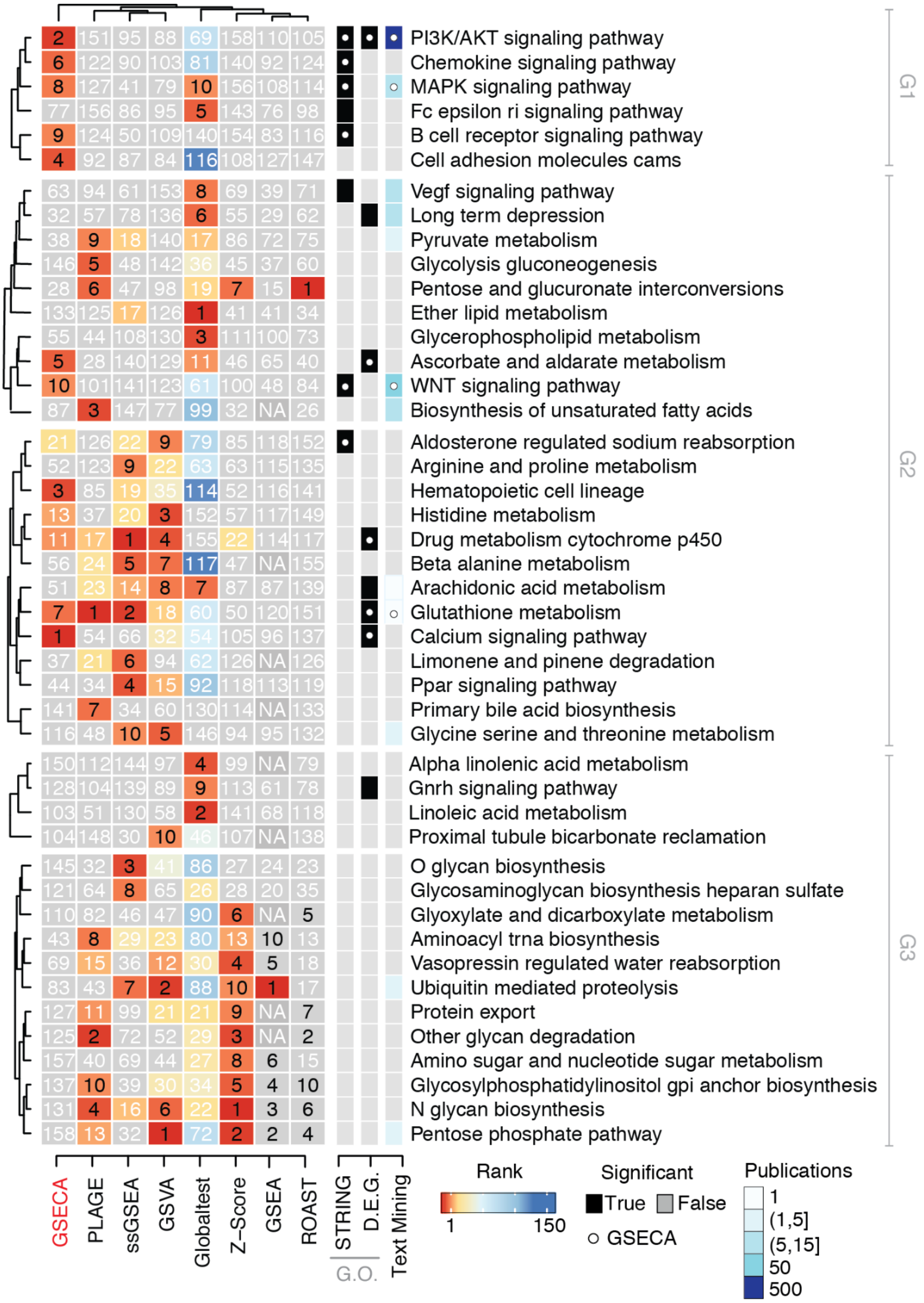


Figure 71. From (Lauria et al. 2020), Overlap of AGSs in PRAD PTEN-loss samples identified by the GSA algorithms. GSECA results are reported in red.

Then, gene set analysis (GSA) methods were also compared to other orthogonal approaches. In particular, the ten top-ranked AGSs for each method were selected. Then, gene ontology (GO) analyses were performed using STRING PPI network and the differentially expressed genes in PTEN-loss as compared to PTEN-wt tumors. Also, a text mining of published articles was carried out. Hierarchical clustering revealed the presence of five groups of gene sets that were: (i) identified only by GSECA and Globaltest (G1); (ii) detected by various methods (G2); enriched from methods out of GSECA (G3) (Figure 72). Importantly, GSECA was the only algorithm that indicated the PI3K/AKT pathway as a top-ranking result (Figure 72).



GSECA	PLAGE	ssGSEA	GSVA	Globaltest	Z–Score	GSEA	ROAST	STRING	D.E.G.	Text Mining	
2	151	95	88	69	158	110	105	■ ○	■ ○	○	PI3K/AKT signaling pathway
6	122	90	103	81	140	92	124	■ ○			Chemokine signaling pathway
8	127	41	79	10	156	108	114	■ ○		○	MAPK signaling pathway
77	156	86	95	5	143	76	98	■			Fc epsilon ri signaling pathway
9	124	50	109	140	154	83	116	■ ○			B cell receptor signaling pathway
4	92	87	84	116	108	127	147				Cell adhesion molecules cams
63	94	61	153	8	69	39	71	■			Vegf signaling pathway
32	57	78	136	6	55	29	62		■		Long term depression
38	9	18	140	17	86	72	75				Pyruvate metabolism
146	5	48	142	36	45	37	60				Glycolysis gluconeogenesis
28	6	47	98	19	7	15	1				Pentose and glucuronate interconversions
133	125	17	126	1	41	41	34				Ether lipid metabolism
55	44	108	130	3	111	100	73				Glycerophospholipid metabolism
5	28	140	129	11	46	65	40		■ ○		Ascorbate and aldarate metabolism
10	101	141	123	61	100	48	84	■ ○		○	WNT signaling pathway
87	3	147	77	99	32	NA	26				Biosynthesis of unsaturated fatty acids
21	126	22	9	79	85	118	152	■ ○			Aldosterone regulated sodium reabsorption
52	123	9	22	63	63	115	135				Arginine and proline metabolism
3	85	19	35	114	52	116	141				Hematopoietic cell lineage
13	37	20	3	152	57	117	149				Histidine metabolism
11	17	1	4	155	22	114	117		■ ○		Drug metabolism cytochrome p450
56	24	5	7	117	47	NA	155				Beta alanine metabolism
51	23	14	8	7	87	87	139		■		Arachidonic acid metabolism
7	1	2	18	60	50	120	151		■ ○	○	Glutathione metabolism
1	54	66	32	54	105	96	137		■ ○		Calcium signaling pathway
37	21	6	94	62	126	NA	126				Limonene and pinene degradation
44	34	4	15	92	118	113	119				Ppar signaling pathway
141	7	34	60	130	114	NA	133				Primary bile acid biosynthesis
116	48	10	5	146	94	95	132				Glycine serine and threonine metabolism
150	112	144	97	4	99	NA	79				Alpha linolenic acid metabolism
128	104	139	89	9	113	61	78		■		Gnrh signaling pathway
103	51	130	58	2	141	68	118				Linoleic acid metabolism
104	148	30	10	46	107	NA	138				Proximal tubule bicarbonate reclamation
145	32	3	41	86	27	24	23				O glycan biosynthesis
121	64	8	65	26	28	20	35				Glycosaminoglycan biosynthesis heparan sulfate
110	82	46	47	90	6	NA	5				Glyoxylate and dicarboxylate metabolism
43	8	29	23	80	13	10	13				Aminoacyl trna biosynthesis
69	15	36	12	30	4	5	18				Vasopressin regulated water reabsorption
83	43	7	2	88	10	1	17				Ubiquitin mediated proteolysis
127	11	99	21	21	9	NA	7				Protein export
125	2	72	52	29	3	NA	2				Other glycan degradation
157	40	69	44	27	8	6	15				Amino sugar and nucleotide sugar metabolism
137	10	39	30	34	5	4	10				Glycosylphosphatidylinositol gpi anchor biosynthesis
131	4	16	6	22	1	3	6				N glycan biosynthesis
158	13	32	1	72	2	2	4				Pentose phosphate pathway

Figure 72. From (Lauria et al. 2020), Hierarchical clustering of the first ten top-ranked AGSs in PRAD PTEN-loss detected by GSA algorithms. Each cell reports the rank of the gene set of a specific method. The ranks of the top ten ranked gene sets are reported in black. Annotation heatmap (right) depicts gene sets identified by GO

analysis performed considering the STRING PPI network and differentially expressed genes (*i.e.* D.E.G.) in black and (*ii*) evidence coming from literature text mining in the color key of blues.

Have to be noticed that group 1 contained 75% of gene sets enriched from the STRING PPI gene ontology, confirming the higher ability of GSECA in detecting that are functionally altered upon the loss of PTEN compared to the other methods. GSECA also identified five out of eight gene sets enriched for significantly up- and down-regulated genes. The GO analysis showed the enrichment of PI3K/AKT only when considering significantly activated and repressed genes.

This consideration suggested the agility of GSECA in identifying processes where genes are significantly altered in both directions and not only activated or repressed.

Four out of the ten GSECA top-ranked gene sets showed, moreover, evidence in literature, including the most cited one, PI3K/AKT, that were found in 499 articles related to PTEN loss (Figure 72).

We finally assessed four parameters (*i.e.* FPKM values, FC, absolute FC and dispersion) measuring range, direction, intensity and inter sample heterogeneity (IH) of gene expression captures as altered by each method (Figure 73). In a few words, we evaluated the mean μ and the standard deviation σ of each parameter across prostate cancer samples and, then the average values across genes. Overall, GSECA outperformed the other tools, thanks to its ability in handling gene sets characterized by expression changes of genes that are more intensively activated or repressed (*i.e.* direction and intensity) at all levels (*i.e.* range) in an heterogeneous manner across samples (*i.e.* IH). Has to be noticed that Globaltest performed similarly to GSECA, nevertheless also showed the highest type I error, character that confers less confidence to the results.

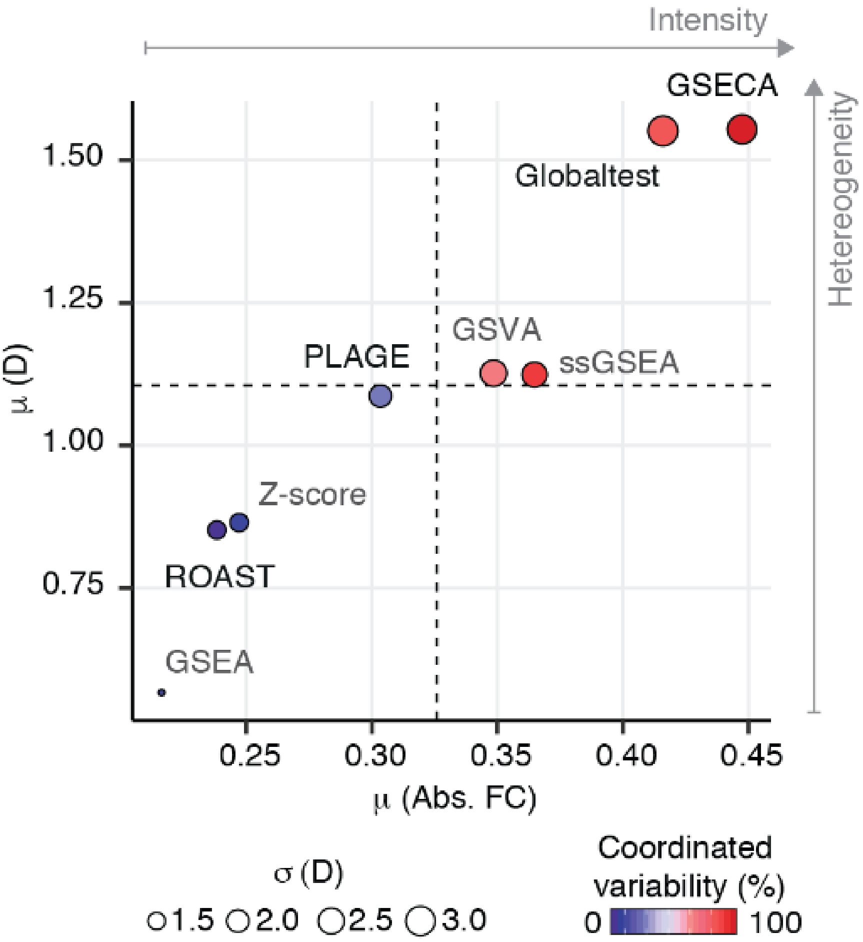


Figure 73. From (Lauria et al. 2020), Scatter plot of the mean absolute FC (Abs. FC), and D averaged on the 20 top-ranked gene sets detected by each method. Dot size represents the average standard deviation (s) of D for the 20 top-ranked gene sets. Color key depicts the percentage of the 20 top-ranked gene sets that contain both activated and repressed gene sets, namely coordinated variability.

As somatic loss of PTEN is frequent across various cancer types, we disentangled its effect at a pancancer level. Somatic alterations and transcriptome profiling data were downloaded from TCGA Data Matrix portal (Level 3, https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) for other 30 cancer types for a total of 9,944 samples and for each PTEN-loss samples were identified as described for the PRAD cohort. Cancer types with at least 30 samples harboring the somatic loss of PTEN were retained for further analysis. In total, 13 cancer types, excluding PRAD, were retained. Running GSECA with 158 KEGG gene sets, we found that 10 out of 13 cancer types showed at least one AGS (AS≤0.05, pemp≤0.05 and SR≥0.7). Significantly, six cancer types showed the alteration of the PI3K/AKT signaling pathway (Figure 74).

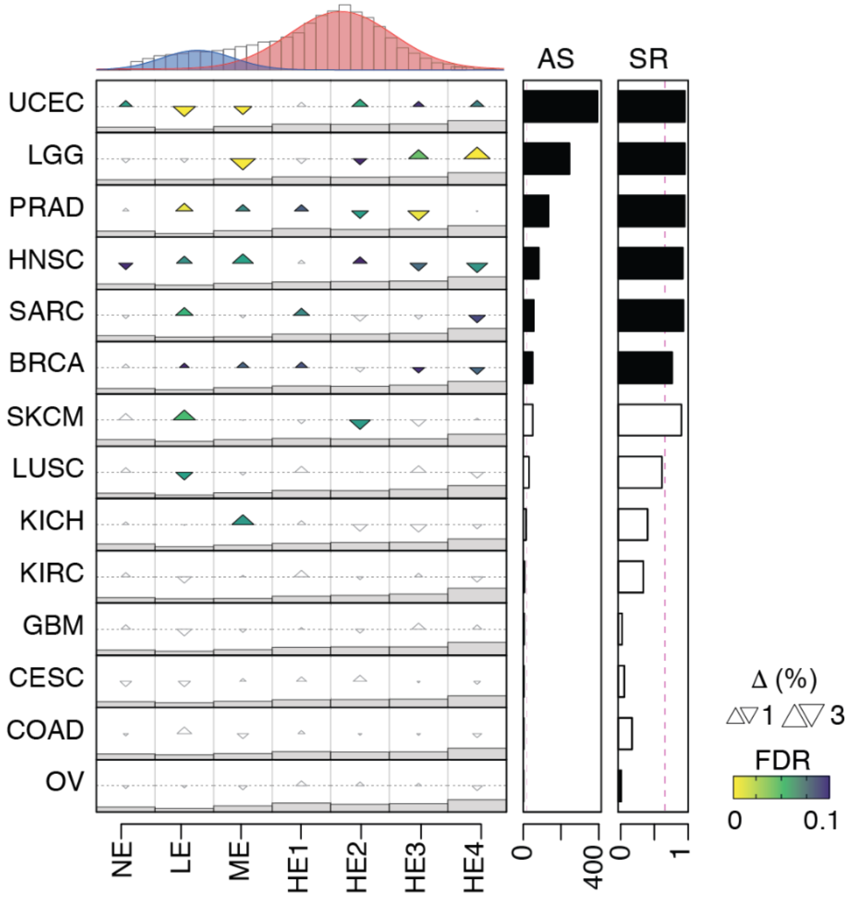


Figure 74. From (Lauria et al. 2020), GSECA EC map showing the pan-cancer alteration of PI3K/AKT signaling pathway as a consequence of the somatic loss of PTEN.

AS was positively correlated with the statistical difference in the cumulative expression of PI3K/AKT genes in PTEN-loss tumors as compared to wild-type ones (Figure 75).

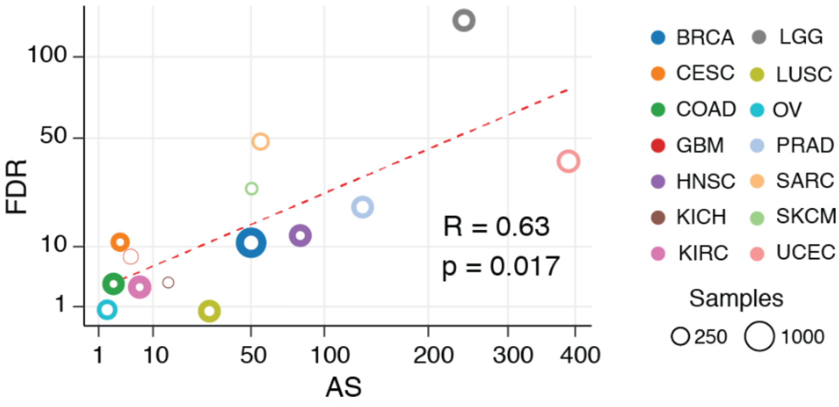


Figure 75. From (Lauria et al. 2020), Scatter plot showing the correlation of GSECA AS and the significant alteration of PI3K/AKT signaling pathway in PTEN-loss as compared to PTEN-wt tumors measured in terms of

the adjusted p-value (*i.e.* FDR) across cancer types. The size of the colored circles shows the number of samples, whereas the inner white circles the number of PTEN loss samples.

We further inspected results, by selecting the 10 top-ranked AGSs for each cancer type. We found that metabolic processes, information-related processes and immune system gene sets were the most frequent (Figure 76). In particular, SARC, KIRCH and SKCM showed the highest number of immune-related AGSs.

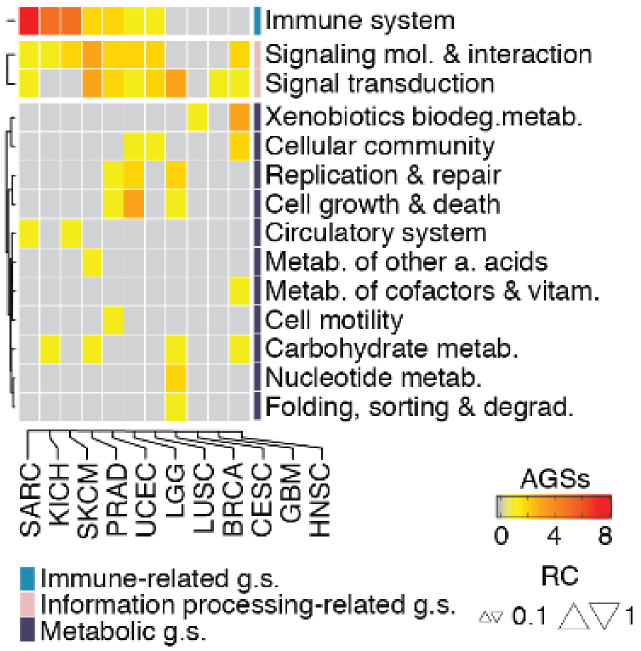


Figure 76. From (Lauria et al. 2020), Heatmap showing the altered classes of gene sets across cancer types. Classes are defined according to the KEGG category. Each cell reports the number of AGSs. The annotation heatmap indicates the KEGG superclass of biological processes.

Focusing on immune-related gene sets, we compared GSECA results with the ones of the other gene set analysis methods and evaluated the changes in the tumor immune microenvironment (TIME) upon the loss of PTEN (Figure 77). In particular, we collected information about the cellular composition of immune infiltrates for 14 TCGA tumor types and statistically measured the differences in the composition of 22 immune cell types. The final degree of alteration of the immune cell population was obtained combining the significance level of each comparison through the Fisher's Method into an unique immune score (IS).

GSECA detected the highest number of tumors with a significant alteration of immune cell fractions and showed the highest positive correlation between the number of immune-related AGSs and IS (Pearson's correlation coefficient R=0.77, P-value=0.003).

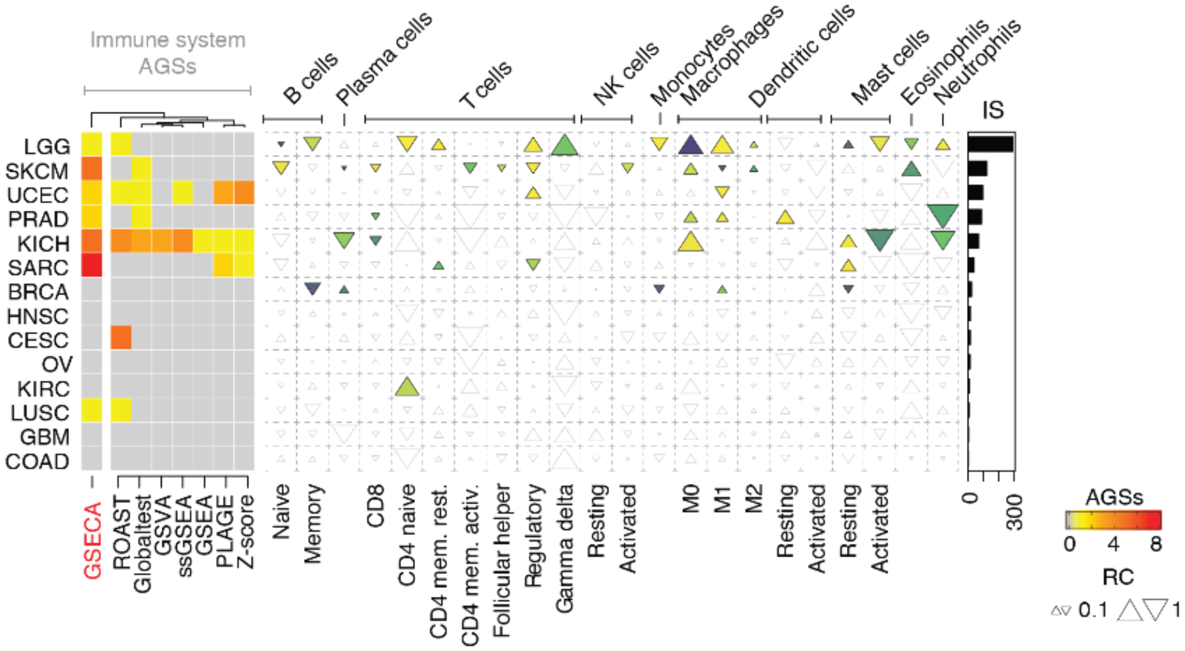


Figure 77. From (Lauria et al. 2020), Heatmap on the left panel shows the number of immune-related gene sets that are altered upon the loss of PTEN across cancer types according to GSECA and the other GSA methods. On the right panel, EC map-like heatmap depicts the statistically significant alteration of the immune cell population across cancer types. The size of triangles, the relative change of the percentage of tumor immune infiltrates between PTEN-loss and wild-type samples. The upper and lower vertexes of the triangles represent the increase or decrease of immune cells in PTEN-loss samples as compared to PTEN-wt tumors. The bar plot reports the IS for each cancer type.

To further investigate the immunosuppressive trait of the loss of PTEN in PRAD, that had been previously suggested (L. Chen and Guo 2017), we ran GSECA on a collection of 102 expression signatures representative of different immune cell activities, states and modes in tumor tissues (Thorsson et al. 2018). The altered pathways were 15 (Figure 78). Six of the top ten AGSs characterized the state and activity of T and B cells, showing a general reduction of gene expression in the high expression classes and reciprocal increase of genes in lower ones.

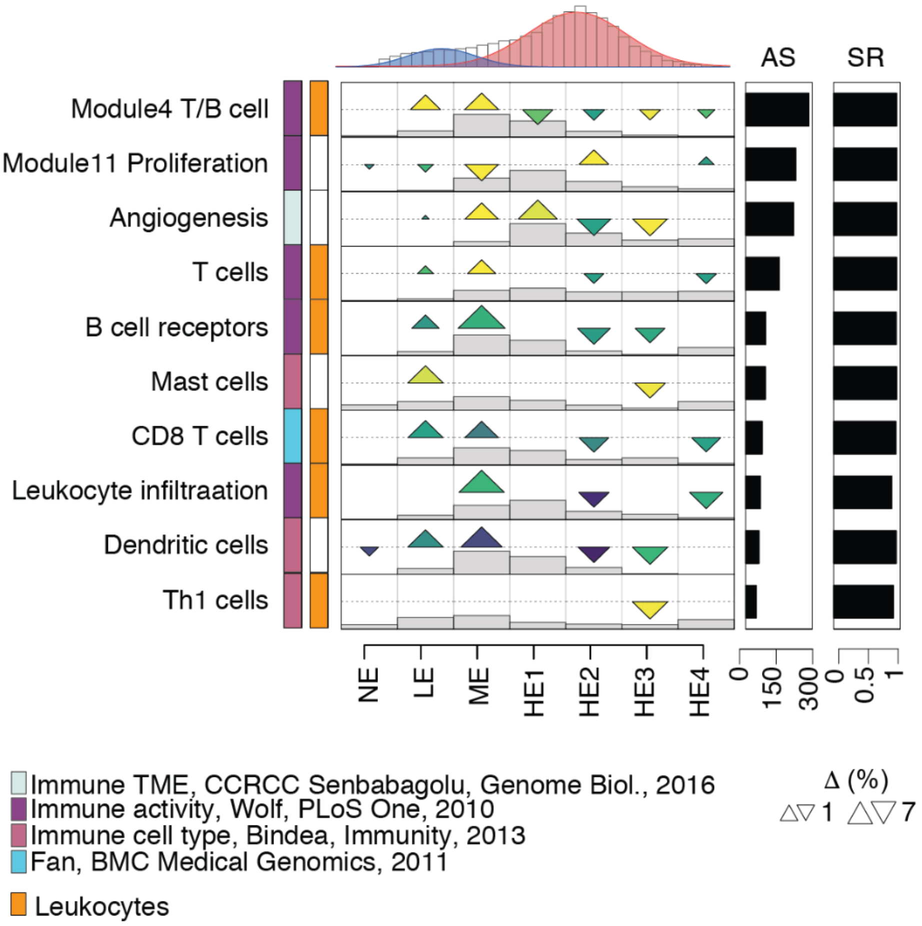


Figure 78. From (Lauria et al. 2020), GSECA EC map showing the altered immune expression signatures as a consequence of the somatic loss of PTEN in PRAD.

The two top-ranked gene sets contained markers of lymphocyte activation and proliferation. Intriguingly, down-regulation of T/B cell modules in combination with the upregulation of proliferation had been related to poor survival in breast cancer (Wolf et al. 2014), so we explored the impact of the loss of PTEN on disease free survival (DFS) of prostate cancer patients.

Clinical data were downloaded from the GDC data portal (https://portal.gdc.cancer.gov) for 355 PRAD patients and the DFS time was defined as the interval between the date of treatment and disease progression, as defined by biochemical or clinical recurrence, or until the end of follow up ("The Molecular Taxonomy of Primary Prostate Cancer" 2015). We observed a statistically significant difference in DFS in the first 24 months from the treatment between patients with PTEN loss and rest (Figure 79), confirming the harmful impact of PTEN loss on cancer phenotype.

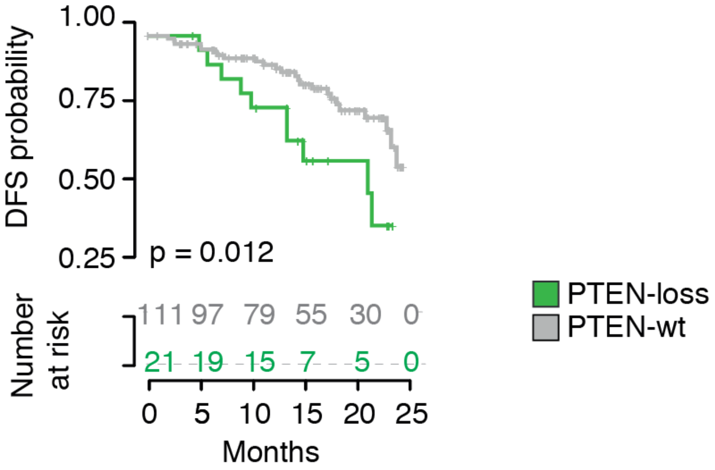


Figure 79. From (Lauria et al. 2020), Disease-free survival (DFS) Kaplan-Mayer curves for PTEN-loss and PTEN-wt patients.

We also tested if the absolute PTEN expression levels were also prognostic of survival. The maximally selected rank statistics approach was used. We found that patients with PTEN expression levels lower than 3.58 TPM had a statistically significantly shorter DFS time in the first two year from the treatment (Figure 80).

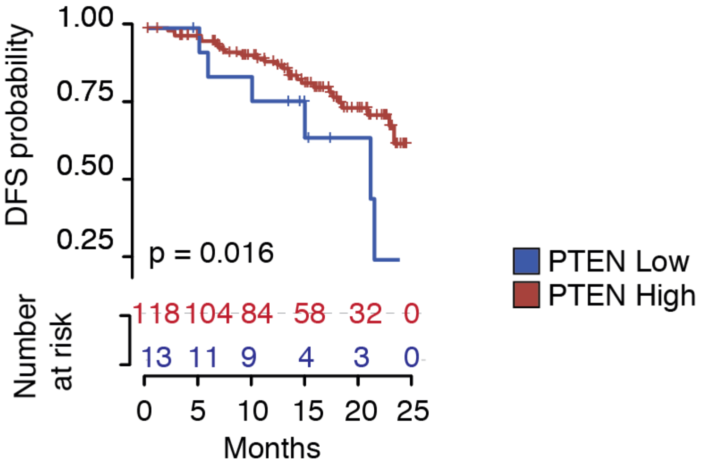


Figure 80. From (Lauria et al. 2020), DSF Kaplan-Mayer curves measured stratifying PRAD patients on the optimal PTEN expression level (*i.e.* TPM=3.56, maximally selected rank statistics = 2.34) within two years from the initial treatment.

Taken together, these results showed that GSECA is able to detect altered biological processes between highly heterogeneous cohorts, keeping in consideration genes both if decreased or increased. In particular, we accurately associated the loss of PTEN to the alteration of PI3K/AKT signaling pathway and to the different regulation of immune-related processes across cancer types. Our results also support the emerging role of PTEN in the immune system (Chakravarthy et al. 2018; Armstrong et al. 2016; Leavy 2015) and therapy resistance (George et al. 2017; Peng et al. 2016; Tilot et al. 2016).

Appendix B. Insights into literature

To learn to read is to light a fire; every syllable that is spelled out is a spark.

Victor Hugo

A. Compendium of ChIP-sequencing experiments: ReMap and Chip-Atlas databases

ReMap (Griffon et al. 2015; Chèneby et al. 2018, 2020; Hammal et al. 2021) is an online database containing ChIP-seq datasets. It was created in 2015 and updated in 2018 (now it is actually in the 2022 version, after another in 2020, but the 2018 version is the one that was available when I started the PhD). Curators analyzed 2,829 quality controlled ChIP-seq experiments from ENCODE and public sources (GEO, ArrayExpress). The public ChIP-seq datasets (n = 1,763), as well as the ENCODE ChIP-seq data (n = 1,066), have been mapped to the GRCh19/hg19 human assembly. They defined a "dataset" as a ChIP-seq experiment in a given series (*e.g.* GSE46237), for a given TF (*e.g.* NR2C2), in a particular biological condition (*i.e.* cell line, tissue type, disease state or experimental conditions; *e.g.* HELA). Datasets were labeled by concatenating these three pieces of information. ENCODE and Public data have been analyzed to propose an unified integration of both data sources, producing a unique atlas of regulatory regions for 485 transcription regulators (TRs). They found 125 TRs common to the two sets, 154 proteins specific to ENCODE and 206 specific to the Public catalog. Taken separately, the ENCODE peaks overlapped by 96% the Public regions, and 87% of the Public peaks overlapped ENCODE regions. Notably, 347 cell lines in 40 tissues have been taken under consideration. As not every ChIP-seq datasets are equal in terms of quality, they used four different metrics based on ENCODE ChIP-seq guidelines to retain high quality datasets for downstream analyses. First, they used the normalized strand cross-correlation coefficient (NSC) which is a normalized ratio between the fragment-length cross-correlation peak and the background cross-correlation, and the relative strand cross-correlation coefficient (RSC), a ratio between the fragment-length peak and the read-length peak to exclude low quality datasets. Moreover, they also used the FRiP, the fraction of reads in peaks, and the number of peaks identified in each dataset (min.100) to filter datasets. Every dataset that did not respect these parameters was excluded. After consistent peak calling, they identified 49 million peaks bound by transcription factors (80 million with ENCODE data included). These numbers include overlapping sites for identical TRs which were studied in various conditions.

ChIP-Atlas (Oki et al. 2018) is an online database containing ChIP-seq and DNase-seq datasets. Experiments targeting RNA polymerases, transcription factors, chromatin-remodeling factors and histone modification enzymes from the NCBI Sequence Read Archive (SRA) are included. The motivations under this work of uniformation are that although data on SRA are publicly available, only raw data are archived in most cases, metadata are often ambiguous and integrative analyses require skills for data mining and huge computational resources. ChIP-Atlas database contains 76,217 experiments consisting of 6 model organisms, human, mouse, rat, fruit fly, nematode and budding yeast. All experiments are analyzed using the same pipeline and no quality filters are applied, contrary to the ReMap database. Antigens and cell types are sorted into “antigen class” and “cell type classes”. Moreover, a brief description of each experiment is also furnished, including quality controls, allowing a punctual selection and extraction of data.

B. Recurrent accessible chromatin regions in distinct cancer types

Corges et al. (Corces et al. 2018) generated ATAC-seq data in 410 tumor samples from 404 donors of TCGA across 23 cancer types. Technical replicates were also done for 386 samples and only data that passed a minimum threshold of enrichment of signal over background were retained. Hg38 assembly was used for alignment and peak calling was performed with fixed-width option. Authors also defined “cancer type-specific peak sets”, combining all “sample peak sets” from a given cancer type into a cumulative peak set. Peaks from the merged peak set that were observed in at least two samples with a score per million ≥ 5 were considered as reproducible. Results of this analysis are reproducible, fixed-width peaks for a specific cancer type.

C. The Network of Cancer Genes: NCG

The Network of Cancer Genes (NCG) (Repana et al. 2019; Rambaldi et al. 2008; D’Antonio and Ciccarelli 2011; D’Antonio et al. 2012; An et al. 2014, 2016; Dressler et al. 2021) is a manually curated repository of genes whose somatic modification is known or predicted to have a cancer driver role. NCG6 (Repana et al. 2019), the version used in the present work, contains 2,372 genes that were collected from 275 publications, including two sources of known cancer genes and 273 cancer sequencing screens of 119 cancer types in 31 primary sites, for a total of 34,905 cancer donors. In addition to collecting cancer genes, NCG annotates their system-level properties, such as duplicability, evolutionary origin, RNA and protein expression, miRNA and protein interactions, protein function and essentiality.

References

Agirre, E., A. J. Oldfield, N. Bellora, A. Segelle, and R. F. Luco. 2021. “Splicing-Associated Chromatin Signatures: A Combinatorial and Position-Dependent Role for Histone Marks in Splicing Definition.” *Nature Communications* 12 (1). https://doi.org/10.1038/s41467-021-20979-x.

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2002. “Molecular Biology of the Cell.” https://www.ncbi.nlm.nih.gov/books/NBK21054/.

Anczuków, O., and A. R. Krainer. 2016. “Splicing-Factor Alterations in Cancers.” *RNA* 22 (9). https://doi.org/10.1261/rna.057919.116.

Anna, A., and G. Monika. 2018. “Splicing Mutations in Human Genetic Disorders: Examples, Detection, and Confirmation.” *Journal of Applied Genetics* 59 (3). https://doi.org/10.1007/s13353-018-0444-7.

An, O., G. M. Dall’Olio, T. P. Mourikis, and F. D. Ciccarelli. 2016. “NCG 5.0: Updates of a Manually Curated Repository of Cancer Genes and Associated Properties from Cancer Mutational Screenings.” *Nucleic Acids Research* 44 (D1). https://doi.org/10.1093/nar/gkv1123.

An, O., V. Pendino, M. D’Antonio, E. Ratti, M. Gentilini, and F. D. Ciccarelli. 2014. “NCG 4.0: The Network of Cancer Genes in the Era of Massive Mutational Screenings of Cancer Genomes.” *Database: The Journal of Biological Databases and Curation* 2014 (March). https://doi.org/10.1093/database/bau015.

Armstrong, C. W., P. J. Maxwell, C. W. Ong, K. M. Redmond, C. McCann, J. Neisen, G. A. Ward, et al. 2016. “PTEN Deficiency Promotes Macrophage Infiltration and Hypersensitivity of Prostate Cancer to IAP Antagonist/radiation Combination Therapy.” *Oncotarget* 7 (7). https://doi.org/10.18632/oncotarget.6955.

Arsura, M., A. Deshpande, S. R. Hann, and G. E. Sonenshein. 1995. “Variant Max Protein, Derived by Alternative Splicing, Associates with c-Myc in Vivo and Inhibits Transactivation.” *Molecular and Cellular Biology* 15 (12). https://doi.org/10.1128/MCB.15.12.6702.

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.

Baca, S. C., D. Y. Takeda, J. H. Seo, J. Hwang, S. Y. Ku, R. Arafeh, T. Arnoff, et al. 2021. “Reprogramming of the FOXA1 Cistrome in Treatment-Emergent Neuroendocrine Prostate Cancer.” *Nature Communications* 12 (1). https://doi.org/10.1038/s41467-021-22139-7.

Baek, D., and P. Green. 2005. “Sequence Conservation, Relative Isoform Frequencies, and Nonsense-Mediated Decay in Evolutionarily Conserved Alternative Splicing.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (36). https://doi.org/10.1073/pnas.0506139102.

Barbie, D. A., P. Tamayo, J. S. Boehm, S. Y. Kim, S. E. Moody, I. F. Dunn, A. C. Schinzel, et al. 2009. “Systematic RNA Interference Reveals That Oncogenic KRAS-Driven Cancers Require TBK1.” *Nature* 462 (7269). https://doi.org/10.1038/nature08460.

Barfeld, S. J., A. Urbanucci, H. M. Itkonen, L. Fazli, J. L. Hicks, B. Thiede, P. S. Rennie, S. Yegnasubramanian, A. M. DeMarzo, and I. G. Mills. 2017. “C-Myc Antagonises the Transcriptional Activity of the Androgen Receptor in Prostate Cancer Affecting Key Gene Networks.” *EBioMedicine* 18 (April). https://doi.org/10.1016/j.ebiom.2017.04.006.

Bernardo, G. M., and R. A. Keri. 2012. “FOXA1: A Transcription Factor with Parallel Functions in Development and Cancer.” *Bioscience Reports* 32 (2). https://doi.org/10.1042/BSR20110046.

Bissonnette, R. P., A. McGahon, A. Mahboubi, and D. R. Green. 1994. “Functional Myc-Max Heterodimer Is Required for Activation-Induced Apoptosis in T Cell Hybridomas.” *The Journal of Experimental Medicine* 180 (6). https://doi.org/10.1084/jem.180.6.2413.

Boija, A., I. A. Klein, B. R. Sabari, A. Dall’Agnese, E. L. Coffey, A. V. Zamudio, C. H. Li, et al. 2018. “Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains.” *Cell* 175 (7). https://doi.org/10.1016/j.cell.2018.10.042.

Briese, M., N. Haberman, C. R. Sibley, R. Faraway, A. S. Elser, A. M. Chakrabarti, Z. Wang, et al.

2019. “A Systems View of Spliceosomal Assembly and Branchpoints with iCLIP.” *Nature Structural & Molecular Biology* 26 (10). https://doi.org/10.1038/s41594-019-0300-4.

Cerami, E., J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, et al. 2012. “The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data.” *Cancer Discovery* 2 (5). https://doi.org/10.1158/2159-8290.CD-12-0095.

Cereda, M., G. Gambardella, L. Benedetti, F. Iannelli, D. Patel, G. Basso, R. F. Guerra, et al. 2016. “Patients with Genetically Heterogeneous Synchronous Colorectal Cancer Carry Rare Damaging Germline Mutations in Immune-Related Genes.” *Nature Communications* 7 (July). https://doi.org/10.1038/ncomms12072.

Cereda, M., U. Pozzoli, G. Rot, P. Juvan, A. Schweitzer, T. Clark, and J. Ule. 2014. “RNAmotifs: Prediction of Multivalent RNA Motifs That Control Alternative Splicing.” *Genome Biology* 15 (1). https://doi.org/10.1186/gb-2014-15-1-r20.

Chakravarthy, A., A. Furness, K. Joshi, E. Ghorani, K. Ford, M. J. Ward, E. V. King, et al. 2018. “Pan-Cancer Deconvolution of Tumour Composition Using DNA Methylation.” *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-05570-1.

Chèneby, J., M. Gheorghe, M. Artufel, A. Mathelier, and B. Ballester. 2018. “ReMap 2018: An Updated Atlas of Regulatory Regions from an Integrative Analysis of DNA-Binding ChIP-Seq Experiments.” *Nucleic Acids Research* 46 (D1). https://doi.org/10.1093/nar/gkx1092.

Chèneby, J., Z. Ménétrier, M. Mestdagh, T. Rosnet, A. Douida, W. Rhalloussi, A. Bergon, F. Lopez, and B. Ballester. 2020. “ReMap 2020: A Database of Regulatory Regions from an Integrative Analysis of Human and Arabidopsis DNA-Binding Sequencing Experiments.” *Nucleic Acids Research* 48 (D1). https://doi.org/10.1093/nar/gkz945.

Chen, L., and D. Guo. 2017. “The Functions of Tumor Suppressor PTEN in Innate and Adaptive Immunity.” *Cellular & Molecular Immunology* 14 (7). https://doi.org/10.1038/cmi.2017.30.

Chen, Y., P. Chi, S. Rockowitz, P. J. Iaquinta, T. Shamu, S. Shukla, D. Gao, et al. 2013. “ETS Factors Reprogram the Androgen Receptor Cistrome and Prime Prostate Tumorigenesis in Response to PTEN Loss.” *Nature Medicine* 19 (8). https://doi.org/10.1038/nm.3216.

Chen, Y., M. Huang, X. Liu, Y. Huang, C. Liu, J. Zhu, G. Fu, Z. Lei, and X. Chu. 2021. “Alternative Splicing of mRNA in Colorectal Cancer: New Strategies for Tumor Diagnosis and Treatment.” *Cell Death & Disease* 12 (8). https://doi.org/10.1038/s41419-021-04031-w.

Chen, Z., X. Lan, J. M. Thomas-Ahner, D. Wu, X. Liu, Z. Ye, L. Wang, et al. 2015. “Agonist and Antagonist Switch DNA Motifs Recognized by Human Androgen Receptor in Prostate Cancer.” *The EMBO Journal* 34 (4). https://doi.org/10.15252/embj.201490306.

Cieśla, M., P. C. T. Ngoc, E. Cordero, Á. S. Martinez, M. Morsing, S. Muthukumar, G. Beneventi, et al. 2021. “Oncogenic Translation Directs Spliceosome Dynamics Revealing an Integral Role for SF3A3 in Breast Cancer.” *Molecular Cell* 81 (7). https://doi.org/10.1016/j.molcel.2021.01.034.

Coomer, A. O., F. Black, A. Greystoke, J. Munkley, and D. J. Elliott. 2019. “Alternative Splicing in Lung Cancer.” *Biochimica et Biophysica Acta, Gene Regulatory Mechanisms* 1862 (11-12). https://doi.org/10.1016/j.bbagrm.2019.05.006.

Corces, M. R., J. M. Granja, S. Shams, B. H. Louie, J. A. Seoane, W. Zhou, T. C. Silva, et al. 2018. “The Chromatin Accessibility Landscape of Primary Human Cancers.” *Science* 362 (6413). https://doi.org/10.1126/science.aav1898.

D’Antonio, M., and F. D. Ciccarelli. 2011. “Modification of Gene Duplicability during the Evolution of Protein Interaction Network.” *PLoS Computational Biology* 7 (4). https://doi.org/10.1371/journal.pcbi.1002029.

D’Antonio, M., V. Pendino, S. Sinha, and F. D. Ciccarelli. 2012. “Network of Cancer Genes (NCG 3.0): Integration and Analysis of Genetic and Network Properties of Cancer Genes.” *Nucleic Acids Research* 40 (Database issue). https://doi.org/10.1093/nar/gkr952.

David, C. J., M. Chen, M. Assanah, P. Canoll, and J. L. Manley. 2010. “HnRNP Proteins Controlled by c-Myc Deregulate Pyruvate Kinase mRNA Splicing in Cancer.” *Nature* 463 (7279). https://doi.org/10.1038/nature08697.

Del Giudice, M., S. Peirone, S. Perrone, F. Priante, F. Varese, E. Tirtei, F. Fagioli, and M. 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.

Demichelis, F., P. Magni, P. Piergiorgi, M. A. Rubin, and R. Bellazzi. 2006. "A Hierarchical Naïve Bayes Model for Handling Sample Heterogeneity in Classification Problems: An Application to Tissue Microarrays." *BMC Bioinformatics* 7 (November). https://doi.org/10.1186/1471-2105-7-514.

Ding, Y., G. Feng, and M. Yang. 2020. "Prognostic Role of Alternative Splicing Events in Head and Neck Squamous Cell Carcinoma." *Cancer Cell International* 20 (May). https://doi.org/10.1186/s12935-020-01249-0.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1). https://doi.org/10.1093/bioinformatics/bts635.

Donaghey, J., S. Thakurela, J. Charlton, J. S. Chen, Z. D. Smith, H. Gu, R. Pop, et al. 2018. "Genetic Determinants and Epigenetic Effects of Pioneer-Factor Occupancy." *Nature Genetics* 50 (2). https://doi.org/10.1038/s41588-017-0034-3.

Dressler, Lisa, Michele Bortolomeazzi, Mohamed Reda Keddar, Hrvoje Misetic, Giulia Sartini, Amelia Acha-Sagredo, Lucia Montorsi, et al. 2021. "Comparative Assessment of Genes Driving Cancer and Somatic Evolution in Noncancer Tissues: An Update of the NCG Resource." *bioRxiv*. https://doi.org/10.1101/2021.08.31.458177.

El Marabti, E., and I. Younis. 2018. "The Cancer Spliceome: Reprograming of Alternative Splicing in Cancer." *Frontiers in Molecular Biosciences* 5 (September). https://doi.org/10.3389/fmolb.2018.00080.

Expert-Bezançon, A., A. Sureau, P. Durosay, R. Salesse, H. Groeneveld, J. P. Lecaer, and J. Marie. 2004. "hnRNP A1 and the SR Proteins ASF/SF2 and SC35 Have Antagonistic Functions in Splicing of Beta-Tropomyosin Exon 6B." *The Journal of Biological Chemistry* 279 (37). https://doi.org/10.1074/jbc.M405377200.

Fei, T., W. Li, J. Peng, T. Xiao, C. H. Chen, A. Wu, J. Huang, C. Zang, X. S. Liu, and M. Brown. 2019. "Deciphering Essential Cistromes Using Genome-Wide CRISPR Screens." *Proceedings of the National Academy of Sciences of the United States of America* 116 (50). https://doi.org/10.1073/pnas.1908155116.

Feng, H., S. Bao, M. A. Rahman, S. M. Weyn-Vanhentenryck, A. Khan, J. Wong, A. Shah, E. D. Flynn, A. R. Krainer, and C. Zhang. 2019. "Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites." *Molecular Cell* 74 (6). https://doi.org/10.1016/j.molcel.2019.02.002.

Fisher, R. A. 1992. "Statistical Methods for Research Workers." In *Breakthroughs in Statistics*, 66–70. Springer, New York, NY.

Foster, John G., Rebecca Arkell, Marco Del Giudice, Chinedu Anene, Andrea Lauria, John D. Kelly, Nicholas R. Lemoine, Salvatore Oliviero, Matteo Cereda, and Prabhakar Rajan. 2018. "Dysregulation of Splicing-Related Proteins in Prostate Cancer Is Controlled by FOXA1." *bioRxiv*. https://doi.org/10.1101/509034.

Frankish, A., M. Diekhans, A. M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, et al. 2019. "GENCODE Reference Annotation for the Human and Mouse Genomes." *Nucleic Acids Research* 47 (D1). https://doi.org/10.1093/nar/gky955.

Fu, X., R. Pereira, C. De Angelis, J. Veeraraghavan, S. Nanda, L. Qin, M. L. Cataldo, et al. 2019. "FOXA1 Upregulation Promotes Enhancer and Transcriptional Reprogramming in Endocrine-Resistant Breast Cancer." *Proceedings of the National Academy of Sciences of the United States of America* 116 (52). https://doi.org/10.1073/pnas.1911584116.

Gambardella, G., M. Cereda, L. Benedetti, and F. D. Ciccarelli. 2017. "MEGA-V: Detection of Variant Gene Sets in Patient Cohorts." *Bioinformatics* 33 (8). https://doi.org/10.1093/bioinformatics/btw809.

Gasser, E., H. C. Johannssen, T. Rülicke, H. U. Zeilhofer, and M. Stoffel. 2016. "Foxa1 Is Essential for Development and Functional Integrity of the Subthalamic Nucleus." *Scientific Reports* 6 (December). https://doi.org/10.1038/srep38611.

Gehring, N. H., and J. Y. Roignant. 2021. "Anything but Ordinary - Emerging Splicing Mechanisms in Eukaryotic Gene Regulation." *Trends in Genetics: TIG* 37 (4). https://doi.org/10.1016/j.tig.2020.10.008.

George, S., D. Miao, G. D. Demetri, D. Adeegbe, S. J. Rodig, S. Shukla, M. Lipschitz, et al. 2017.

"Loss of PTEN Is Associated with Resistance to Anti-PD-1 Checkpoint Blockade Therapy in Metastatic Uterine Leiomyosarcoma." *Immunity* 46 (2). https://doi.org/10.1016/j.immuni.2017.02.001.

Gerhardt, J., M. Montani, P. Wild, M. Beer, F. Huber, T. Hermanns, M. Müntener, and G. Kristiansen. 2012. "FOXA1 Promotes Tumor Progression in Prostate Cancer and Represents a Novel Hallmark of Castration-Resistant Prostate Cancer." *The American Journal of Pathology* 180 (2). https://doi.org/10.1016/j.ajpath.2011.10.021.

Ghigna, C., C. Valacca, and G. Biamonti. 2008. "Alternative Splicing and Tumor Progression." *Current Genomics* 9 (8). https://doi.org/10.2174/138920208786847971.

Goeman, J. J., S. A. van de Geer, F. de Kort, and H. C. van Houwelingen. 2004. "A Global Test for Groups of Genes: Testing Association with a Clinical Outcome." *Bioinformatics* 20 (1). https://doi.org/10.1093/bioinformatics/btg382.

Grandori, C., J. Mac, F. Siëbelt, D. E. Ayer, and R. N. Eisenman. 1996. "Myc-Max Heterodimers Activate a DEAD Box Gene and Interact with Multiple E Box-Related Sites in Vivo." *The EMBO Journal* 15 (16). https://pubmed.ncbi.nlm.nih.gov/8861962/.

Griffon, A., Q. Barbier, J. Dalino, J. van Helden, S. Spicuglia, and B. Ballester. 2015. "Integrative Analysis of Public ChIP-Seq Experiments Reveals a Complex Multi-Cell Regulatory Landscape." *Nucleic Acids Research* 43 (4). https://doi.org/10.1093/nar/gku1280.

Hammal, F., P. de Langen, A. Bergon, F. Lopez, and B. Ballester. 2021. "ReMap 2022: A Database of Human, Mouse, Drosophila and Arabidopsis Regulatory Regions from an Integrative Analysis of DNA-Binding Sequencing Experiments." *Nucleic Acids Research*, November. https://doi.org/10.1093/nar/gkab996.

Hänzelmann, S., R. Castelo, and J. Guinney. 2013. "GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data." *BMC Bioinformatics* 14 (January). https://doi.org/10.1186/1471-2105-14-7.

Hawe, J. S., F. J. Theis, and M. Heinig. 2019. "Inferring Interaction Networks From Multi-Omics Data." *Frontiers in Genetics* 10 (June). https://doi.org/10.3389/fgene.2019.00535.

Hebenstreit, D., M. Fang, M. Gu, V. Charoensawan, A. van Oudenaarden, and S. A. Teichmann. 2011. "RNA Sequencing Reveals Two Major Classes of Gene Expression Levels in Metazoan Cells." *Molecular Systems Biology* 7 (June). https://doi.org/10.1038/msb.2011.28.

Hegele, A., A. Kamburov, A. Grossmann, C. Sourlis, S. Wowro, M. Weimann, C. L. Will, V. Pena, R. Lührmann, and U. Stelzl. 2012. "Dynamic Protein-Protein Interaction Wiring of the Human Spliceosome." *Molecular Cell* 45 (4). https://doi.org/10.1016/j.molcel.2011.12.034.

Helman, P., R. Veroff, Atlas SR, and C. Willman. 2004. "A Bayesian Network Classification Methodology for Gene Expression Data." *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 11 (4). https://doi.org/10.1089/cmb.2004.11.581.

Hug, N., D. Longman, and J. F. Cáceres. 2016. "Mechanism and Regulation of the Nonsense-Mediated Decay Pathway." *Nucleic Acids Research* 44 (4). https://doi.org/10.1093/nar/gkw010.

Imamura, Y., S. Sakamoto, T. Endo, T. Utsumi, M. Fuse, T. Suyama, K. Kawamura, et al. 2012. "FOXA1 Promotes Tumor Progression in Prostate Cancer via the Insulin-like Growth Factor Binding Protein 3 Pathway." *PloS One* 7 (8). https://doi.org/10.1371/journal.pone.0042456.

"Introduction to Bivariate and Multivariate Analysis." n.d. Accessed December 28, 2021a. https://books.google.com/books/about/Introduction_to_Bivariate_and_Multivaria.html?hl=it&id=-hfvAAAAMAAJ.

———. n.d. Accessed December 28, 2021b. https://books.google.com/books/about/Introduction_to_Bivariate_and_Multivaria.html?hl=it&id=-hfvAAAAMAAJ.

Jiménez-Vacas, J. M., V. Herrero-Aguayo, A. J. Montero-Hidalgo, E. Gómez-Gómez, A. C. Fuentes-Fayos, A. J. León-González, P. Sáez-Martínez, et al. 2020. "Dysregulation of the Splicing Machinery Is Directly Associated to Aggressiveness of Prostate Cancer." *EBioMedicine* 51 (January). https://doi.org/10.1016/j.ebiom.2019.11.008.

Jin, H. J., J. C. Zhao, I. Ogden, R. C. Bergan, and J. Yu. 2013. "Androgen Receptor-Independent Function of FoxA1 in Prostate Cancer Metastasis." *Cancer Research* 73 (12). https://doi.org/10.1158/0008-5472.CAN-12-3468.

Jin, H. J., J. C. Zhao, L. Wu, J. Kim, and J. Yu. 2014. “Cooperativity and Equilibrium with FOXA1 Define the Androgen Receptor Transcriptional Program.” *Nature Communications* 5 (May). https://doi.org/10.1038/ncomms4972.

Jones, D., M. Wade, S. Nakjang, L. Chaytor, J. Grey, C. N. Robson, and L. Gaughan. 2015. “FOXA1 Regulates Androgen Receptor Variant Activity in Models of Castrate-Resistant Prostate Cancer.” *Oncotarget* 6 (30). https://doi.org/10.18632/oncotarget.4927.

Kahles, A., K. V. Lehmann, N. C. Toussaint, M. Hüser, S. G. Stark, T. Sachsenberg, O. Stegle, O. Kohlbacher, C. Sander, and G. Rätsch. 2018. “Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients.” *Cancer Cell* 34 (2). https://doi.org/10.1016/j.ccell.2018.07.001.

Kajita, M., K. N. McClinic, and P. A. Wade. 2004. “Aberrant Expression of the Transcription Factors Snail and Slug Alters the Response to Genotoxic Stress.” *Molecular and Cellular Biology* 24 (17). https://doi.org/10.1128/MCB.24.17.7559-7566.2004.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. 2002. “The Human Genome Browser at UCSC.” *Genome Research* 12 (6). https://doi.org/10.1101/gr.229102.

Keren, Hadas, Galit Lev-Maor, and Gil Ast. 2010. “Alternative Splicing and Evolution: Diversification, Exon Definition and Function.” *Nature Reviews. Genetics* 11 (5): 345–55.

Koedoot, E., M. Smid, J. A. Foekens, J. W. M. Martens, S. E. Le Dévédec, and B. van de Water. 2019. “Co-Regulated Gene Expression of Splicing Factors as Drivers of Cancer Progression.” *Scientific Reports* 9 (1). https://doi.org/10.1038/s41598-019-40759-4.

Koh, C. M., M. Bezzi, D. H. Low, W. X. Ang, S. X. Teo, F. P. Gay, M. Al-Haddawi, et al. 2015. “MYC Regulates the Core Pre-mRNA Splicing Machinery as an Essential Step in Lymphomagenesis.” *Nature* 523 (7558). https://doi.org/10.1038/nature14351.

Kurosaki, T., M. W. Popp, and L. E. Maquat. 2019. “Quality and Quantity Control of Gene Expression by Nonsense-Mediated mRNA Decay.” *Nature Reviews. Molecular Cell Biology* 20 (7). https://doi.org/10.1038/s41580-019-0126-2.

Lambert, S. A., A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch. 2018. “The Human Transcription Factors.” *Cell* 172 (4). https://doi.org/10.1016/j.cell.2018.01.029.

Lardelli, R. M., J. X. Thompson, J. R. Yates, and S. W. Stevens. 2010. “Release of SF3 from the Intron Branchpoint Activates the First Step of Pre-mRNA Splicing.” *RNA* 16 (3). https://doi.org/10.1261/rna.2030510.

Lauria, A., S. Peirone, Giudice, F. Priante, P. Rajan, M. Caselle, S. Oliviero, and M. Cereda. 2020. “Identification of Altered Biological Processes in Heterogeneous RNA-Sequencing Data by Discretization of Expression Profiles.” *Nucleic Acids Research* 48 (4). https://doi.org/10.1093/nar/gkz1208.

Leavy, O. 2015. “Regulatory T Cells. The PTEN Stabilizer.” *Nature Reviews. Immunology* 15 (2). https://doi.org/10.1038/nri3809.

Lee, E., H. Y. Chuang, J. W. Kim, T. Ideker, and D. Lee. 2008. “Inferring Pathway Activity toward Precise Disease Classification.” *PLoS Computational Biology* 4 (11). https://doi.org/10.1371/journal.pcbi.1000217.

Lee, S. C., and O. Abdel-Wahab. 2016. “Therapeutic Targeting of Splicing in Cancer.” *Nature Medicine* 22 (9). https://doi.org/10.1038/nm.4165.

Liao, Y., G. K. Smyth, and W. Shi. 2014. “featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7). https://doi.org/10.1093/bioinformatics/btt656.

Littell, R., and J. L. Folks. 1971. “Asymptotic Optimality of Fisher’s Method of Combining Independent Tests.” https://doi.org/10.1080/01621459.1971.10482347.

Liu, Huan, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash. 2002. “Discretization: An Enabling Technique.” *Data Mining and Knowledge Discovery* 6 (4): 393–423.

Li, Y., N. Sahni, R. Pancsa, D. J. McGrail, J. Xu, X. Hua, J. Coulombe-Huntington, et al. 2017. “Revealing the Determinants of Widespread Alternative Splicing Perturbation in Cancer.” *Cell Reports* 21 (3). https://doi.org/10.1016/j.celrep.2017.09.071.

Love, M. I., W. Huber, and S. Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12). https://doi.org/10.1186/s13059-014-0550-8.

Lupien, M., J. Eeckhoute, C. A. Meyer, Q. Wang, Y. Zhang, W. Li, J. S. Carroll, X. S. Liu, and M. Brown. 2008. "FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription." *Cell* 132 (6). https://doi.org/10.1016/j.cell.2008.01.018.

"Machine Learning on Big Data: Opportunities and Challenges." 2017. *Neurocomputing* 237 (May): 350–61.

Marchand, V., M. Santerre, C. Aigueperse, L. Fouillen, J. M. Saliou, A. Van Dorsselaer, S. Sanglier-Cianférani, C. Branlant, and Y. Motorin. 2011. "Identification of Protein Partners of the Human Immunodeficiency Virus 1 Tat/rev Exon 3 Leads to the Discovery of a New HIV-1 Splicing Regulator, Protein hnRNP K." *RNA Biology* 8 (2). https://doi.org/10.4161/rna.8.2.13984.

Mazin, Pavel V., Philipp Khaitovich, Margarida Cardoso-Moreira, and Henrik Kaessmann. 2021. "Alternative Splicing during Mammalian Organ Development." *Nature Genetics* 53 (6): 925–34.

McLachlan, Geoffrey J., and David Peel. 2000. *Finite Mixture Models*.

Mobley, Arie. 2019. *Neural Stem Cells and Adult Neurogenesis*. 1st ed. Academic Press.

"Modeling Survival Data: Extending the Cox Model." n.d. Accessed December 28, 2021. https://link.springer.com/book/10.1007%2F978-1-4757-3294-8.

Müller-McNicoll, M., O. Rossbach, J. Hui, and J. Medenbach. 2019. "Auto-Regulatory Feedback by RNA-Binding Proteins." *Journal of Molecular Cell Biology* 11 (10). https://doi.org/10.1093/jmcb/mjz043.

Munkley, J., L. Li, S. R. G. Krishnan, G. Hysenaj, E. Scott, C. Dalgliesh, H. Z. Oo, et al. 2019. "Androgen-Regulated Transcription of ESRP2 Drives Alternative Splicing Patterns in Prostate Cancer." *eLife* 8 (September). https://doi.org/10.7554/eLife.47678.

Murray, J. I., R. B. Voelker, K. L. Henscheid, M. B. Warf, and J. A. Berglund. 2008. "Identification of Motifs That Function in the Splicing of Non-Canonical Introns." *Genome Biology* 9 (6). https://doi.org/10.1186/gb-2008-9-6-r97.

Nair, S. K., and S. K. Burley. 2003. "X-Ray Structures of Myc-Max and Mad-Max Recognizing DNA. Molecular Bases of Regulation by Proto-Oncogenic Transcription Factors." *Cell* 112 (2). https://doi.org/10.1016/s0092-8674(02)01284-9.

Oki, S., T. Ohta, G. Shioi, H. Hatanaka, O. Ogasawara, Y. Okuda, H. Kawaji, R. Nakaki, J. Sese, and C. Meno. 2018. "ChIP-Atlas: A Data-Mining Suite Powered by Full Integration of Public ChIP-Seq Data." *EMBO Reports* 19 (12). https://doi.org/10.15252/embr.201846255.

Papasaikas, P., J. R. Tejedor, L. Vigevani, and J. Valcárcel. 2015. "Functional Splicing Network Reveals Extensive Regulatory Potential of the Core Spliceosomal Machinery." *Molecular Cell* 57 (1). https://doi.org/10.1016/j.molcel.2014.10.030.

Parolia, A., M. Cieslik, S. C. Chu, L. Xiao, T. Ouchi, Y. Zhang, X. Wang, et al. 2019. "Distinct Structural Classes of Activating FOXA1 Alterations in Advanced Prostate Cancer." *Nature* 571 (7765). https://doi.org/10.1038/s41586-019-1347-4.

Paschalis, A., A. Sharp, J. C. Welti, A. Neeb, G. V. Raj, J. Luo, S. R. Plymate, and J. S. de Bono. 2018. "Alternative Splicing in Prostate Cancer." *Nature Reviews. Clinical Oncology* 15 (11). https://doi.org/10.1038/s41571-018-0085-0.

Pelengaris, S., M. Khan, and G. Evan. 2002. "C-MYC: More than Just a Matter of Life and Death." *Nature Reviews. Cancer* 2 (10). https://doi.org/10.1038/nrc904.

Peng, W., J. Q. Chen, C. Liu, S. Malu, C. Creasy, M. T. Tetzlaff, C. Xu, et al. 2016. "Loss of PTEN Promotes Resistance to T Cell-Mediated Immunotherapy." *Cancer Discovery* 6 (2). https://doi.org/10.1158/2159-8290.CD-15-0283.

Pervouchine, D., Y. Popov, A. Berry, B. Borsari, A. Frankish, and R. Guigó. 2019. "Integrative Transcriptomic Analysis Suggests New Autoregulatory Splicing Events Coupled with Nonsense-Mediated mRNA Decay." *Nucleic Acids Research* 47 (10). https://doi.org/10.1093/nar/gkz193.

Phillips, J. W., Y. Pan, B. L. Tsai, Z. Xie, L. Demirdjian, W. Xiao, H. 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). https://doi.org/10.1073/pnas.1915975117.

Planell, N., V. Lagani, P. Sebastian-Leon, F. van der Kloet, E. Ewing, N. Karathanasis, A. Urdangarin, et al. 2021. “STATegra: Multi-Omics Data Integration - A Conceptual Scheme With a Bioinformatics Pipeline.” *Frontiers in Genetics* 12 (March). https://doi.org/10.3389/fgene.2021.620453.

Pozzoli, U., and M. Sironi. 2005. “Silencers Regulate Both Constitutive and Alternative Splicing Events in Mammals.” *Cellular and Molecular Life Sciences: CMLS* 62 (14). https://doi.org/10.1007/s00018-005-5030-6.

Qiu, M., W. Bao, J. Wang, T. Yang, X. He, Y. Liao, and X. Wan. 2014. “FOXA1 Promotes Tumor Cell Proliferation through AR Involving the Notch Pathway in Endometrial Cancer.” *BMC Cancer* 14 (February). https://doi.org/10.1186/1471-2407-14-78.

Qiu, Xintao, Nadia Boufaied, Tarek Hallal, Avery Feit, Anna de Polo, Adrienne M. Luoma, Janie Larocque, et al. 2021. “MYC Drives Aggressive Prostate Cancer by Disrupting Transcriptional Pause Release at Androgen Receptor Targets.” *bioRxiv*. https://doi.org/10.1101/2021.04.23.441016.

Quinlan, A. R., and I. M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6). https://doi.org/10.1093/bioinformatics/btq033.

Ramanand, S. G., Y. Chen, J. Yuan, K. Daescu, M. B. Lambros, K. E. Houlahan, S. Carreira, et al. 2020. “The Landscape of RNA Polymerase II-Associated Chromatin Interactions in Prostate Cancer.” *The Journal of Clinical Investigation* 130 (8). https://doi.org/10.1172/JCI134260.

Rambaldi, D., F. M. Giorgi, F. Capuani, A. Ciliberto, and F. D. Ciccarelli. 2008. “Low Duplicability and Network Fragility of Cancer Genes.” *Trends in Genetics: TIG* 24 (9). https://doi.org/10.1016/j.tig.2008.06.003.

Ray, D., H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, et al. 2013. “A Compendium of RNA-Binding Motifs for Decoding Gene Regulation.” *Nature* 499 (7457). https://doi.org/10.1038/nature12311.

Repana, D., J. Nulsen, L. Dressler, M. Bortolomeazzi, S. K. Venkata, A. Tourna, A. Yakovleva, T. Palmieri, and F. D. Ciccarelli. 2019. “The Network of Cancer Genes (NCG): A Comprehensive Catalogue of Known and Candidate Cancer Genes from Cancer Sequencing Screens.” *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-018-1612-0.

Rhie, S. K., A. A. Perez, F. D. Lay, S. Schreiner, J. Shi, J. Polin, and P. J. Farnham. 2019. “A High-Resolution 3D Epigenomic Map Reveals Insights into the Creation of the Prostate Cancer Transcriptome.” *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-12079-8.

Rizzo, F., M. Nizzardo, S. Vashisht, E. Molteni, V. Melzi, M. Taiana, S. Salani, et al. 2019. “Key Role of SMN/SYNCRIP and RNA-Motif 7 in Spinal Muscular Atrophy: RNA-Seq and Motif Analysis of Human Motor Neurons.” *Brain: A Journal of Neurology* 142 (2). https://doi.org/10.1093/brain/awy330.

Robinson, J. L., and J. S. Carroll. 2012. “FoxA1 Is a Key Mediator of Hormonal Response in Breast and Prostate Cancer.” *Frontiers in Endocrinology* 3 (May). https://doi.org/10.3389/fendo.2012.00068.

Robinson, J. L., T. E. Hickey, A. Y. Warren, S. L. Vowler, T. Carroll, A. D. Lamb, N. Papoutsoglou, D. E. Neal, W. D. Tilley, and J. S. Carroll. 2014. “Elevated Levels of FOXA1 Facilitate Androgen Receptor Chromatin Binding Resulting in a CRPC-like Phenotype.” *Oncogene* 33 (50). https://doi.org/10.1038/onc.2013.508.

Robinson, D. J. McCarthy, and G. K. Smyth. 2010. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1). https://doi.org/10.1093/bioinformatics/btp616.

Rot, G., Z. Wang, I. Huppertz, M. Modic, T. Lenče, M. Hallegger, N. Haberman, T. Curk, C. von Mering, and J. Ule. 2017. “High-Resolution RNA Maps Suggest Common Principles of Splicing and Polyadenylation Regulation by TDP-43.” *Cell Reports* 19 (5). https://doi.org/10.1016/j.celrep.2017.04.028.

“Royal Statistical Society Publications.” n.d. Accessed December 29, 2021. https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1977.tb01600.x.

Sahu, B., M. Laakso, K. Ovaska, T. Mirtti, J. Lundin, A. Rannikko, A. Sankila, et al. 2011. “Dual Role of FoxA1 in Androgen Receptor Binding to Chromatin, Androgen Signalling and Prostate Cancer.” *The EMBO Journal* 30 (19). https://doi.org/10.1038/emboj.2011.328.

Saraiva-Agostinho, N., and N. L. Barbosa-Morais. 2019. "Psichomics: Graphical Application for Alternative Splicing Quantification and Analysis." *Nucleic Acids Research* 47 (2). https://doi.org/10.1093/nar/gky888.

Schafer, S., K. Miao, C. C. Benson, M. Heinig, S. A. Cook, and N. Hubner. 2015. "Alternative Splicing Signatures in RNA-Seq Data: Percent Spliced in (PSI)." *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]* 87 (October). https://doi.org/10.1002/0471142905.hg1116s87.

Seachrist, D. D., L. J. Anstine, and R. A. Keri. 2021. "FOXA1: A Pioneer of Nuclear Receptor Action in Breast Cancer." *Cancers* 13 (20). https://doi.org/10.3390/cancers13205205.

Shah, K., T. Gagliano, L. Garland, T. O'Hanlon, D. Bortolotti, V. Gentili, R. Rizzo, G. Giamas, and M. Dean. 2020. "Androgen Receptor Signaling Regulates the Transcriptome of Prostate Cancer Cells by Modulating Global Alternative Splicing." *Oncogene* 39 (39). https://doi.org/10.1038/s41388-020-01429-2.

Shah, N., and M. Brown. 2019. "The Sly Oncogene: FOXA1 Mutations in Prostate Cancer." *Cancer Cell* 36 (2). https://doi.org/10.1016/j.ccell.2019.07.005.

Shiraishi, Y., K. Kataoka, K. Chiba, A. Okada, Y. Kogure, H. Tanaka, S. Ogawa, and S. Miyano. 2018. "A Comprehensive Characterization of Cis-Acting Splicing-Associated Variants in Human Cancer." *Genome Research* 28 (8). https://doi.org/10.1101/gr.231951.117.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43). https://doi.org/10.1073/pnas.0506580102.

Subramanian, I., S. Verma, S. Kumar, A. Jere, and K. Anamika. 2020. "Multi-Omics Data Integration, Interpretation, and Its Application." *Bioinformatics and Biology Insights* 14 (January). https://doi.org/10.1177/1177932219899051.

Sutandy, F. X. R., S. Ebersberger, L. Huang, A. Busch, M. Bach, H. S. Kang, J. Fallmann, et al. 2018. "In Vitro iCLIP-Based Modeling Uncovers How the Splicing Factor U2AF2 Relies on Regulation by Cofactors." *Genome Research* 28 (5). https://doi.org/10.1101/gr.229757.117.

Su, Z., J. Wang, J. Yu, X. Huang, and X. Gu. 2006. "Evolution of Alternative Splicing after Gene Duplication." *Genome Research* 16 (2). https://doi.org/10.1101/gr.4197006.

Teng, M., S. Zhou, C. Cai, M. Lupien, and H. H. He. 2021. "Pioneer of Prostate Cancer: Past, Present and the Future of FOXA1." *Protein & Cell* 12 (1). https://doi.org/10.1007/s13238-020-00786-8.

"The Genotype-Tissue Expression (GTEx) Project." 2013. *Nature Genetics* 45 (6). https://doi.org/10.1038/ng.2653.

"The Molecular Taxonomy of Primary Prostate Cancer." 2015. *Cell* 163 (4). https://doi.org/10.1016/j.cell.2015.10.025.

Thomas, J. D., J. T. Polaski, Q. Feng, E. J. De Neef, E. R. Hoppe, M. V. McSharry, J. Pangallo, et al. 2020. "RNA Isoform Screens Uncover the Essentiality and Tumor-Suppressor Activity of Ultraconserved Poison Exons." *Nature Genetics* 52 (1). https://doi.org/10.1038/s41588-019-0555-z.

Thompson, M. G., R. Muñoz-Moreno, P. Bhat, R. Roytenberg, J. Lindberg, M. R. Gazzara, M. J. Mallory, et al. 2018. "Co-Regulatory Activity of hnRNP K and NS1-BP in Influenza and Human mRNA Splicing." *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-04779-4.

Thorsson, V., D. L. Gibbs, S. D. Brown, D. Wolf, D. S. Bortone, Ou Yang Th, E. Porta-Pardo, et al. 2018. "The Immune Landscape of Cancer." *Immunity* 48 (4). https://doi.org/10.1016/j.immuni.2018.03.023.

Tilot, A. K., G. Bebek, F. Niazi, J. B. Altemus, T. Romigh, T. W. Frazier, and C. Eng. 2016. "Neural Transcriptome of Constitutional Pten Dysfunction in Mice and Its Relevance to Human Idiopathic Autism Spectrum Disorder." *Molecular Psychiatry* 21 (1). https://doi.org/10.1038/mp.2015.17.

Tomfohr, J., J. Lu, and T. B. Kepler. 2005. "Pathway Level Analysis of Gene Expression Using Singular Value Decomposition." *BMC Bioinformatics* 6 (September). https://doi.org/10.1186/1471-2105-6-225.

Ule, J., and B. J. Blencowe. 2019. "Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution." *Molecular Cell* 76 (2).

https://doi.org/10.1016/j.molcel.2019.09.017.

Ule, J., and R. B. Darnell. 2006. “RNA Binding Proteins and the Regulation of Neuronal Synaptic Plasticity.” *Current Opinion in Neurobiology* 16 (1). https://doi.org/10.1016/j.conb.2006.01.003.

Ullrich, S., and R. Guigó. 2020. “Dynamic Changes in Intron Retention Are Tightly Associated with Regulation of Splicing Factors and Proliferative Activity during B-Cell Development.” *Nucleic Acids Research* 48 (3). https://doi.org/10.1093/nar/gkz1180.

Urbanski, Laura, Mattia Brugiolo, Sunghee Park, Brittany Angarola, Nathan K. Leclair, Phil Palmer, Sangram Keshari Sahu, and Olga Anczuków. 2021. “MYC Regulates a Pan-Cancer Network of Co-Expressed Oncogenic Splicing Factors.” *bioRxiv*. https://doi.org/10.1101/2021.11.24.469558.

Van Nostrand, E. L., P. Freese, G. A. Pratt, X. Wang, X. Wei, R. Xiao, S. M. Blue, et al. 2020. “A Large-Scale Binding and Functional Map of Human RNA-Binding Proteins.” *Nature* 583 (7818). https://doi.org/10.1038/s41586-020-2077-3.

Van Nostrand, E. L., G. A. Pratt, B. A. Yee, E. C. Wheeler, S. M. Blue, J. Mueller, S. S. Park, et al. 2020. “Principles of RNA Processing from Analysis of Enhanced CLIP Maps for 150 RNA Binding Proteins.” *Genome Biology* 21 (1). https://doi.org/10.1186/s13059-020-01982-9.

Vaquerizas, J. M., S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. 2009. “A Census of Human Transcription Factors: Function, Expression and Evolution.” *Nature Reviews. Genetics* 10 (4). https://doi.org/10.1038/nrg2538.

Vasaikar, S. V., P. Straub, J. Wang, and B. Zhang. 2018. “LinkedOmics: Analyzing Multi-Omics Data within and across 32 Cancer Types.” *Nucleic Acids Research* 46 (D1). https://doi.org/10.1093/nar/gkx1090.

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.

Vorontsov, I. E., I. V. Kulakovskiy, and V. J. Makeev. 2013. “Jaccard Index Based Similarity Measure to Compare Transcription Factor Binding Site Models.” *Algorithms for Molecular Biology: AMB* 8 (1). https://doi.org/10.1186/1748-7188-8-23.

Walhout, A. J., J. M. Gubbels, R. Bernards, P. C. van der Vliet, and H. T. Timmers. 1997. “C-Myc/Max Heterodimers Bind Cooperatively to the E-Box Sequences Located in the First Intron of the Rat Ornithine Decarboxylase (ODC) Gene.” *Nucleic Acids Research* 25 (8). https://doi.org/10.1093/nar/25.8.1493.

Wang, Q., J. Armenia, C. Zhang, A. V. Penson, E. Reznik, L. Zhang, T. Minet, et al. 2018. “Unifying Cancer and Normal RNA Sequencing Data from Different Sources.” *Scientific Data* 5 (April). https://doi.org/10.1038/sdata.2018.61.

Wang, Z., and C. B. Burge. 2008. “Splicing Regulation: From a Parts List of Regulatory Elements to an Integrated Splicing Code.” *RNA* 14 (5). https://doi.org/10.1261/rna.876308.

Wickramasinghe, V. O., M. Gonzàlez-Porta, D. Perera, A. R. Bartolozzi, C. R. Sibley, M. Hallegger, J. Ule, J. C. Marioni, and A. R. Venkitaraman. 2015. “Regulation of Constitutive and Alternative mRNA Splicing across the Human Transcriptome by PRPF8 Is Determined by 5’ Splice Site Strength.” *Genome Biology* 16 (1). https://doi.org/10.1186/s13059-015-0749-3.

Wolf, D. M., M. E. Lenburg, C. Yau, A. Boudreau, and L. J. van ’t Veer. 2014. “Gene Co-Expression Modules as Clinically Relevant Hallmarks of Breast Cancer Diversity.” *PloS One* 9 (2). https://doi.org/10.1371/journal.pone.0088309.

Wu, D., E. Lim, F. Vaillant, M. L. Asselin-Labat, J. E. Visvader, and G. K. Smyth. 2010. “ROAST: Rotation Gene Set Tests for Complex Microarray Experiments.” *Bioinformatics* 26 (17). https://doi.org/10.1093/bioinformatics/btq401.

Xu, T. P., Y. F. Wang, W. L. Xiong, P. Ma, W. Y. Wang, W. M. Chen, Huang, et al. 2017. “E2F1 Induces TINCR Transcriptional Activity and Accelerates Gastric Cancer Progression via Activation of TINCR/STAU1/CDKN2B Signaling Axis.” *Cell Death & Disease* 8 (6). https://doi.org/10.1038/cddis.2017.205.

Zarnack, K., J. König, M. Tajnik, I. Martincorena, S. Eustermann, I. Stévant, A. Reyes, S. Anders, N. M. Luscombe, and J. Ule. 2013. “Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements.” *Cell* 152 (3). https://doi.org/10.1016/j.cell.2012.12.023.

Zhang, D., Q. Hu, X. Liu, Y. Ji, H. P. Chao, Y. Liu, A. Tracz, et al. 2020. “Intron Retention Is a

Hallmark and Spliceosome Represents a Therapeutic Vulnerability in Aggressive Prostate Cancer." *Nature Communications* 11 (1). https://doi.org/10.1038/s41467-020-15815-7.

Zhang, Y., J. Qian, C. Gu, and Y. Yang. 2021. "Alternative Splicing and Cancer: A Systematic Review." *Signal Transduction and Targeted Therapy* 6 (1). https://doi.org/10.1038/s41392-021-00486-7.