# Analyses of alternative splicing landscapes in clear cell renal cell carcinomas reveal putative novel prognosis factors

**Pedro Nuno Brazão Faria**

Thesis to obtain the Master of Science Degree in

## Biomedical Engineering

Supervisor: Prof. Susana de Almeida Mendes Vinga Martins
Supervisor: Dr. Nuno Luís Barbosa Morais

### Examination Committee

Chairperson: Prof. João Pedro Estrela Rodrigues Conde
Supervisor: Dr. Nuno Luís Barbosa Morais
Member of the Committee: Prof. Sara Alexandra Cordeiro Madeira

**October 2014**

# Agradecimentos

Agradeço ao Dr. Nuno Morais, à Prof. Susana Vinga Martins e à Prof. Alexandra Carvalho pela imensa dedicação, disponibilidade e paciência demonstradas ao longo deste projeto.

À Dra. Ana Rita Grosso e ao Dr. Sérgio Almeida pela disponibilização dos dados de RNA-seq do TCGA.

À minha família que foi sempre um pilar durante todo o meu curso e vida. Agradeço especialmente aos meus pais e aos meus padrinhos pela generosidade e carinho que sempre demonstraram.

À Joana pelo seu inabalável otimismo e tremenda paciência.

A todos os colegas com quem partilhei as imensas horas passadas no Técnico.

Gostaria também de agradecer à Fundação para a Ciência e a Tecnologia (FCT) pelo apoio através do projeto CancerSys (EXPL/EMS-SIS/1954/2013).

# Abstract

The recent development of next-generation sequencing (NGS) largely improved our means to study transcriptomes. By RNA sequencing (RNA-seq, the use of NGS to sequence complementary DNA reversely transcribed from RNAs), one can not only quantify gene expression (GE) levels, with a higher resolution than microarrays, but also quantitatively reveal unknown transcripts and splicing isoforms. However, the use of RNA-Seq to find cancer transcriptomic signatures beyond GE has been very limited, partly due to a lack of accurate and efficient computational tools.

In this work, we have analysed GE, alternative splicing (AS) and associated patient survival using RNA-seq data from 138 clear cell renal cell carcinomas (ccRCCs) and 62 matched normal kidney samples from The Cancer Genome Atlas (TCGA) project, aiming to identify cancer-specific AS patterns as well as AS events that can potentially serve as prognostic factors. In addition, we have applied dimension reduction and regression methods in order to develop a cancer stage classifier based on AS patterns.

It was observed that, like GE, AS patterns primarily separate normal from tumour samples, with some exons exhibiting a normal/tumour *switch* pattern in their inclusion levels. This is the case, for example, for genes *CD44* and *FGFR2*, previously reported to undergo AS alterations in cancer. Interestingly, a considerable number of the identified cancer-specific AS patterns seem to facilitate an epithelial mesenchymal transition. Furthermore, several AS events appear to be associated with survival, being therefore identified as potential prognostic factors. Finally, the developed classifier revealed ineffective in the classification of the different cancer stages.

These results suggest a great potential of AS signatures derived from tumour transcriptomes in providing etiological leads for cancer progression and as a clinical tool. A deeper understanding of the contribution of splicing alterations to oncogenesis could lead to improved cancer prognosis and contribute to the development of RNA-based anticancer therapeutics, namely splicing-modulating small molecule compounds.



**Keywords:** RNA-seq; survival analysis; alternative splicing; cancer prognosis.

# Resumo

O desenvolvimento recente de ferramentas de sequenciação de nova geração (NGS) facilitou significativamente o estudo de transcriptomas. Com recurso à sequenciação de RNA (RNA-seq, o uso de NGS para sequenciar inversamente transcritos de DNA complementar a partir de RNAs) é possível não só quantificar os níveis de expressão génica com uma resolução superior à obtida através de microarrays mas também revelar e quantificar transcritos e isoformas previamente desconhecidos. Ainda assim, a utilização de RNA-seq na deteção de padrões transcriptómicos associados ao cancro tem sido muito limitada. Tal deve-se, em parte, à escassez de ferramentas precisas e computacionalmente eficientes.

Neste trabalho, analisou-se expressão génica, splicing alternativo e sobrevivência utilizando dados de RNA-seq (do projecto The Cancer Genome Atlas) de carcinomas renais de células claras pertencentes a 138 pacientes e de tecidos normais emparelhados pertencentes a 62 pacientes, com o intuito de identificar padrões de splicing alternativo específicos de tecido tumoral e eventos de splicing alternativo que sejam potenciais factores de prognóstico. Adicionalmente, utilizando métodos de redução dimensional e de regressão, tentou-se conceber um classificador que permitisse a classificar o estádio do cancro analisado tendo por base dados de splicing alternativo.

Observou-se que os padrões de splicing alternativo, tal como a expressão génica, estabelecem a distinção entre tecidos tumorais e normais, com certos exões evidenciando uma mudança drástica nos seus níveis de inclusão entre os dois grupos, sendo portanto potenciais biomarcadores da doença. É o caso, por exemplo, de genes como *CD44* e *FGFR2,* para os quais alterações de splicing alternativo já tinham sido anteriormente associadas a cancro. Curiosamente, um número considerável de padrões de splicing alternativo evidenciados em células tumorais parecem facilitar uma transição epitelial mesenquimal. Observou-se que vários eventos de splicing alternativo parecem estar associados a sobrevivência, constituindo potenciais novos fatores de prognóstico. Finalmente, o classificador concebido revelou-se ineficaz na classificação de estádios tumorais.

Estes resultados sugerem que existe um enorme potencial na utilização de padrões de splicing alternativo em cancro na compreensão da etiologia da progressão tumoral e como ferramenta clínica. Um melhor conhecimento do papel das alterações de splicing na oncogénese pode conduzir a melhorias no prognóstico em cancro e contribuir para o desenvolvimento de terapias anti-tumor baseadas em RNA, nomeadamente compostos moleculares moduladores de splicing.



**Palavras-chave:** RNA-seq; análise de sobrevivência; splicing alternativo; progressão tumoral

# Contents

# List of Figures

# List of Tables

# List of Acronyms and Symbols

$\Delta\widetilde{\Psi}$ **–** difference between two observations (typically in tumour and normal samples) of the median percentage spliced in in for a certain AS event

**Ψ –** percentage spliced in

**A3SS –** alternative 3' splice site

**A5SS –** alternative 5' splice site

**AFE –** alternative first exons

**ALE –** alternative last exons

**AUC –** area under the receiver operating characteristic curve

**AS –** alternative splicing

***BIRC5* –** Baculoviral IAP Repeat Containing 5 gene

**BPS –** branch point sequence

**ccRCC –** clear cell renal cell carcinoma

**CD44 –** CD44 Molecule (Indian Blood Group)

**cDNAs –** complementary DNA

$\boldsymbol{D}$ **–** Deviance

**DAVID –** Database for Annotation, Visualization and Integration Discovery

**DNASE1 –** Deoxyribonuclease I

**EGFR –** epidermal growth factor receptor

**EMT –** epithelial-mesenchymal transition

**ES –** enrichment score

**ESPR –** epithelial splicing regulatory protein

**FDR –** false discovery rate

**FGFR2 –** Fibroblast Growth Factor Receptor 2

***FOXM1* –** Forkhead Box M1 gene

**FPKM –** fragments per kilobase of transcript per million mapped reads

**GE –** Gene expression

**GPR132 –** G Protein-Coupled Receptor 132

**GSEA –** Gene Set Enrichment Analysis

**$H_0$ –** null hypothesis

**$H_1$ –** alternative hypothesis

**HLA –** human leukocyte antigen

**IUM –** initially unmapped reads

**KALRN –** Kalirin, RhoGEF Kinase

**K-S test –** Kolmogorov-Smirnov test

**MCF2L–** MCF.2 Cell Line Derived Transforming Sequence-Like

**MISO –** Mixture-of-isoforms

**MXE –** mutually exclusive exons

**NES –** normalized enrichment score

**NGS –**Next generation sequencing

**NMD –** nonsense-mediated mRNA decay

**PTPN6 –** Protein Tyrosine Phosphatase, Non-Receptor Type 6

**Poly(A)-tail –** poly-adenylated tail

**PPT –** polypyrimidine tract

**PXDN –** Peroxidasin Homolog (Drosophila)

**RAC1 –** Ras-Related C3 Botulinum Toxin Substrate 1

**RCC –**renal cell carcinoma

**RI –** retained introns

**RNA-seq –** RNA sequencing

**ROC –** Receiver operating characteristic

**S_TKc –** Serine/Threonine protein kinases catalytic domain

**SE –** skipped exons

**SH2 –** Src homology 2

**SYK –** Spleen Tyrosine Kinase

**TandemUTR –** tandem 3' UTRs

**TCGA –** The Cancer Genome Atlas

**TNFAI8 –** Tumour Necrosis Factor, Alpha-Induced Protein 8

**UTR –** untranslated region

# 1

# Introduction

## 1.1. Motivation

Cancer is a group of deadly diseases characterized by abnormal cell growth and the potential to invade or spread to other parts of the body. They can be assigned four general stages, according to the extent to which they have developed by spreading: I - localized cancer, usually curable; II - locally advanced, the cancer has spread or invaded beyond the boundaries of its original habitat; III- similar characteristics to stage II cancer, but more advanced; IV - the cancer has spread to other locations throughout the body (metastasis) [1]. Cancer prevalence is set to increase in years to come. According to GLOBOCAN estimates, 14.1 million new cancer cases and 8.2 million cancer-related deaths occurred in 2012. These numbers represent an increase of roughly 11% and 15% in the numbers of new cases and deaths respectively, registered in one year, when compared to 2008 records (12.7 million new cases and 7.1 cancer-related deaths) [2]. The development of new and more effective treatments, as well as more accurate diagnosis tools and methodologies, is thus a pressing matter.

A deep understanding of the triggers and mechanics involved in the oncogenic process is obviously crucial to achieve the aforementioned purposes. In recent years the extensive analysis conducted on a genetic level has made it clear that somatic mutations (mutations in DNA structure that are neither inherited nor passed to offspring), epigenetic changes (changes in the regulation of the expression of gene activity without alteration of genetic structure), and other genetic aberrations can drive human malignancies [3, 4, 5]. Specifically, gene expression (GE) alterations at a transcriptional level are being increasingly associated to oncogenesis and tumour progression. For instance, each of the hallmarks suggested by Hanahan and Weinberg to describe oncogenesis are associated with alterations of splicing patterns [6, 7]. Quantitative studies of transcriptomes are therefore deemed as one of the next major tools in the understanding of cancer biology [8].

The recent development of next-generation sequencing (NGS) technologies largely improved our means to study transcriptomes. By using RNA sequencing (RNA-seq, the use of NGS to sequence complementary DNA (cDNAs) reversely transcribed from RNAs) one can not only quantify GE levels, with a higher resolution than microarrays, but also identify new transcripts and provide quantitative measurements of alternatively spliced isoforms [8]. RNA-seq is therefore a potentially important tool in the establishment of the relation between splicing and cancer development. A deep understanding of the contribution of splicing to oncogenesis could lead to improved cancer prognosis and the development of a novel class of anticancer therapeutics: alternative-splicing inhibitors [7, 8].

## 1.2. Contributions

The main contributions of this work are:

- Identification of cancer-specific AS patterns that may serve as clinical diagnostic tools.
- Proposal of a methodology based on regression and dimension reduction methods to develop a dichotomous cancer stage classifier based on AS quantitation.

- Identification of survival-related AS events that may serve as clinical prognostic factors.

The above endings resulted in a poster presented in the European Conference on Computational Biology:

> Pedro Brazão-Faria, Alexandra M. Carvalho, Susana Vinga and Nuno L. Barbosa-Morais (2014) Analyses of alternative splicing landscapes in clear cell renal cell carcinomas reveal putative novel prognosis factors. ECCB'14 - 13th European Conference on Computational Biology. 7-10 Sep. Strasbourg, France.

## 1.3. Outline

In Chapter 2, we do a literary review on the concepts and methods necessary to both understand and develop this project. In Chapter 3, a methodology to identify cancer- and stage-specific AS patterns, as well as independent AS prognostic factors, is proposed. To that end we have analysed GE, alternative splicing (AS) and associated patient survival using RNA-seq data from 138 clear cell renal cell carcinomas (ccRCCs) and 62 matched normal kidney samples from The Cancer Genome Atlas (TCGA) project. In Chapter 4, we present our analyses' results from biological and technical standpoints. Finally, in Chapter 5, some conclusions on the results are gathered, allowing a perspective on future work possibilities.

# 2

# Background

## 2.1. Alternative Splicing

GE is the process by which information from a gene is used in the synthesis of a functional gene product. The first of its essential steps is transcription, by which a RNA molecule is synthetized from a DNA template. Upstream to the DNA transcription unit is the TATA-box, a smaller region (25-30 bps) that helps to position the complexes involved in transcription. Several transcription factor proteins (TFIID, TFIIA, and TFIIB) bind to specific DNA sequences in this region, preparing DNA to the binding of a suitable RNA polymerase. In eukaryotic cells there are three types of RNA polymerase (I, II and III) that are responsible for transcribing different types of genes. RNA polymerase I and III are responsible for the transcription of RNAs that play structural and catalytic roles (transfer RNAs, ribosomal RNAs and others). RNA polymerase II transcribes the majority of eukaryotic genes, including all those that encode proteins [9]. Once bound to the DNA, RNA polymerase starts to synthetize the premature messenger RNA (pre-mRNA) strand, releasing it when the end of the transcription unit is reached [10]. Most pre-mRNAs are composed by introns and exons. Introns are non-coding regions that are generally removed during the splicing process (see below). Exons are the coding regions. These can be alternative exons, that may or may not be included in an isoform (i.e. any of several different mRNA forms from the same gene), or constitutive exons, that are included in all isoforms [11].

Splicing is part of pre-mRNA processing, which also comprises the adding of a methylated cap, soon after RNA polymerase begins transcription, and a poly-adenylated tail (poly(A)-tail) [12] at the end. Capping and polyadenylation increase the stability of the mRNA and ensure that both ends of the mRNA are present and that the message is therefore complete before protein synthesis begins [15]. Splicing was first observed by Richard J. Roberts and Phillip A. Sharp in 1977. This remarkable discovery led to the awarding of the 1993 Nobel Prize in Physiology or Medicine [13]. In eukaryotes, splicing is performed by the spliceosome, a large complex molecular machine built in several steps. The spliceosome excises the intron, which is well defined by nucleotide sequences. The beginning of the intron (also known as the 5’ splice site) is defined by the GU nucleotide sequence. The end of the intron (also known as the 3’ splice site) is defined by the AG nucleotide sequence and a variable length of upstream polypyrimidines, called the polypyrimidine tract (PPT), which serves the dual function of recruiting factors to the 3' splice site and to the branch point sequence (BPS). The BPS contains the conserved Adenosine required for the first step of splicing (Figure 1). The formation of the spliceosome requires the activity of at least 170 distinct proteins and 5 U-rich small nuclear RNAs (snRNAs) (U1, U2, U4, U5 and U6) that are the core of the major spliceosome. There is also a minor spliceosome that, by using a similar mechanism but some different snRNAs (for example, U11, U12, U4atac, and U6atac in place of U1, U2, U4, and U6), excises less than 1% of introns. The assembly of the spliceosome starts with two small nuclear ribonucleoproteins (snRNPs) binding to the intron through their snRNA components: U1 binds to the GU site and U2 binds to a location nearby, interacting with the BPS. The PPT functions as a binding platform for the U2 snRNP auxiliary factor (U2AF). Then three other snRNPs (U4, U5 and U6) and interactions between their protein components drive the assembly of the complete spliceosome. The complex undergoes a conformational change and splicing may begin. The GU site is cleaved and forms a lariat

with the A nucleotide at the branch point. Finally, the intron is cleaved at the AG sequence and the exons are ligated together [12, 14].

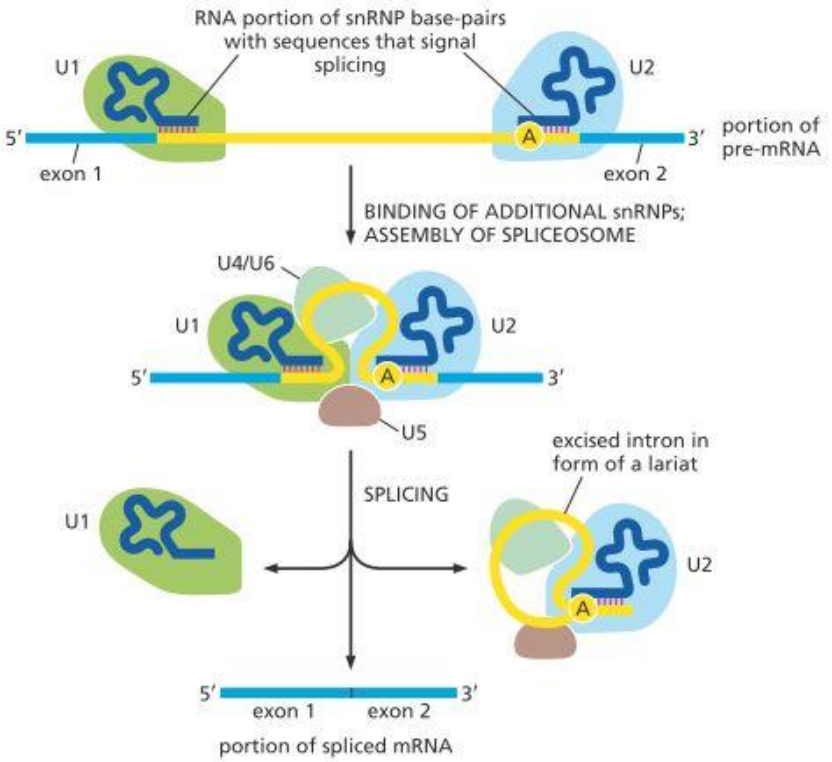



**Figure 1 –** Splicing: U1 binds to the 5' splice site (GU nucleotide sequence) and U2 binds to BPS. Three other small nuclear ribonucleoproteins (U4, U5 and U6) and interactions between their protein components drive the assembly of the complete spliceosome. The product of the splicing process is a functional mRNA [9].

It is important to stress that the splicing of a pre-mRNA does not always give origin to the same isoform, due to AS. AS was first reported by P. Early and colleagues in 1980, when they observed that two different mRNAs could be produced from a single immunoglobulin μ gene [16]. Today, more than 30 years later, the scientific community has a deeper but not yet complete understanding of how AS is regulated. Based on high-throughput sequencing assays, it is estimated that at least 95% of human pre-mRNA undergo AS [17].

There are several AS basic mechanisms, described in Table 1 [18]. The first two mechanisms are characterized by the use of alternative splice junctions, thus changing the boundary of the upstream (5' splice-site choice) or downstream (3' splice-site choice) exon. The other mechanisms have self-explanatory names. The mature mRNA is originated through the combination of one or more of these mechanisms, being exon skipping (or cassette exon) the most common of these in mammals [13]. Naturally, a gene containing a larger number of alternative exons can give origin to a larger set of different mRNAs since the number of possible combinations is greater [20].

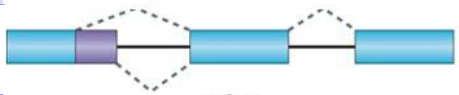
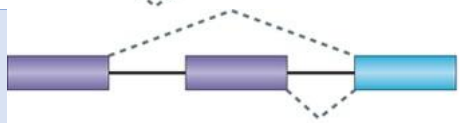
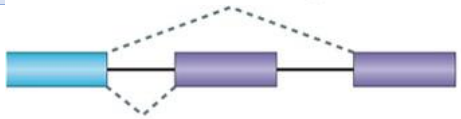
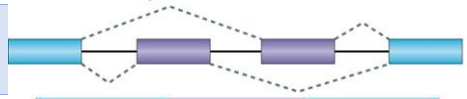
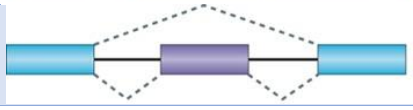
| AS mechanism | Acronym | Schematic representation |
|---|---|---|
| Alternative 3´ splice-site selection | A3SS |  |
| Alternative 5´ splice-site selection | A5SS |  |
| Alternative first exon | AFE |  |
| Alternative last exon | ALE |  |
| Mutually exclusive exons | MXE |  |
| Intron Retention | RI |  |
| Skipped exon | SE |  |

**Table 1 -** Schematics and acronyms of the different types of AS [21].

The biomechanics involved in splice-site selection, and therefore AS, remains incompletely understood. However, it is known that this selection is done by an interplay between RNA-binding proteins and RNA regulatory sequences. Most of the identified RNA-binding proteins are ubiquitously expressed, even though their abundances can differ between tissues. The most studied splicing regulators belong to the SR-protein and hnRNP families. The SR-protein family members contain domains composed of extensive repeats of serine (S) and arginine (R), hence the name, and are thought to activate splicing by binding mostly to exonic sequences and recruiting spliceosomal components. The hnRNP (heterogeneous nuclear ribonucleoprotein) family members, on the other hand, are thought to repress splicing by binding to both exonic and intronic sequences and interfering with the ability of the spliceosome to act on those sites [20].

The splicing patterns, i.e. the relative proportion of different mRNA isoforms originated from the same gene, evidenced by a cell vary in accordance to a large number of factors, such as the type of tissue in which the cell is inserted, the developmental stage of the cell or its differentiation level [20]. Given this tissue-specificity of AS, cellular programs altered in oncogenesis may also be associated with deregulated splicing patterns [7].

## 2.2. Cancer hallmarks and splicing

Hanahan and Weingberg proposed 8 cancer hallmarks: limitless replicative potential, survival (evading apoptosis), invasion and metastasis, immune escape, insensitivity to growth inhibitors, growth factor self-sufficiency, cellular hyperenergetics, and angiogenesis (Figure 2) [6]. As the cell moves through the oncogenic process, its splicing patterns are altered. In fact, each of the 8 suggested hallmarks of

cancer is associated with switches in AS [6]. Therefore, it is reasonable to assume that AS can play a crucial part in tumour evolution.

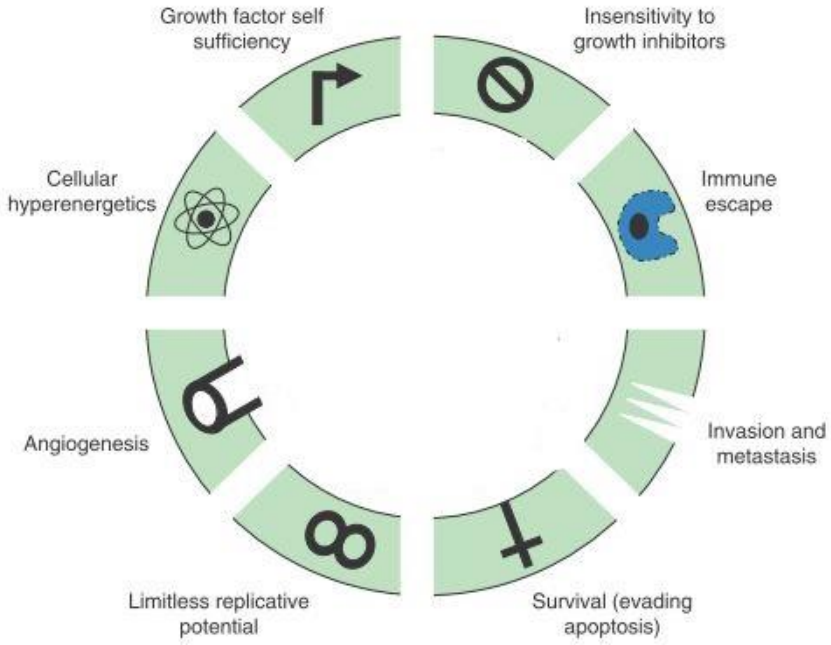



**Figure 2-** The Hallmarks of cancer [7].

## 2.2.1. Limitless replicative potential

The ability to avoid the cellular senescence process resulting from shortening of telomeres (region of repetitive nucleotide sequences at each end of a chromatid, which protects the end of the chromosome from deterioration or from fusion with neighbouring chromosomes) is a key property of cancer cells. Activation of telomerase and high activity is thought to occur in over 90% of cancers [22]. It is not clear how AS regulates this process but several splicing isoforms of hTERT, an important component of telomerase activity, have been reported. These result in shorter proteins that are thought to affect the expression and activity of telomerase through dominant-negative properties (by acting antagonistically to its wild-type). These abnormal splicing isoforms have been detected in various cancers, such as breast and gastric cancer and during lymphoma development [7].

## 2.2.2. Survival by apoptosis evasion

Apoptosis is the process of programmed cell death. This process is triggered by a number of stimuli, internal or external, that may lead to the malfunctioning of the cell. Cancer cells, however, are able to avoid apoptosis.

This is the most studied cancer hallmark in relation to AS [7]. Research has focused on two splicing variants of Bcl-x which result from an alternative 5' splice-site choice producing a pro-apoptotic short isoform (Bcl-xs) and an antiapoptotic long one (Bcl-xl). The production of these two isoforms is unbalanced in a large number of cancer cell lines and human cancer samples [23, 24, 25]. The regulation of the production of these two isoforms is done by several splicing factors (for example, hnRNPs A1 and H/F, Sam 68), conditions and signalling pathways (for example, ceramide, which upregulates Bcl-xs, and protein kinase C, which downregulates the same isoform) [7].

Cancer cells manipulate AS in order to enhance the expression of pro-apoptotic isoforms and diminish antiapoptotic isoform expression. RBM5, a RNA-binding protein with properties of a splicing factor, is abnormally expressed in lung and breast tumours. This protein regulates Fas receptor exon 6 splicing, giving origin to either the membrane-bound Fas receptor with pro-apoptotic function or the soluble form of the receptor, which is antiapoptotic. Part of intron 8 retention results in a splice isoform of caspase 8 (caspase 8L), which has antiapoptotic properties. Exclusion of exons 3, 4, 5 and 6 in caspase 9 results in a smaller protein with antiapoptotic functions, reported to be expressed in several cancer cell lines [26, 27]. The exclusion of caspase 9 exon is regulated by SRSF1 and SRSF2 (SC35), which also regulates Bcl-x splicing-ceramide. Also, the transcription factor E2F1 and splicing factor SRSF2 coordinated action regulates splicing switches between pro and antiapoptotic isoforms of four genes: c-flip, caspase 8, caspase 9 and Bcl-x [7].

A study by the Chabot Laboratory showed that 20 anticancer drugs are able to shift splicing patterns of several apoptotic genes towards promoting apoptosis in several cancer cell lines [28]. This underscores that the understanding of the relationship between AS and cancer is important not only for understanding cancer biology, but also for developing new effective therapeutics.

## 2.2.3. Invasion and metastasis

Over 90% of cancer-related deaths are due to metastasis [29]. The metastatic process is very complex: cells need to be able to leave the primary tumour, intravasate, survive in blood, extravasate and colonise the target tissue. To do so, an incredible phenotypic plasticity is needed. This plasticity can be achieved through epithelial-mesenchymal transition (EMT) and the reverse transitions. EMT is a process by which epithelial cells lose their cell polarity and cell-cell adhesion, and gain migratory and invasive properties to become mesenchymal stem cells. As one would expect, such a complex process cannot happen without highly complex changes in GE. EMT and its reverse process occur normally during embryogenesis and wound healing. However, these processes are hijacked in cancer progression steps, including invasion and angiogenesis [7]. The numerous associated GE changes are controlled by certain transcription factors (for example, twist, snail, zeb 1 and 2). Recently a novel epithelial-specific splice factor, epithelial splicing regulatory protein (ESPR) (isoforms 1 and 2), has been shown to be a master regulator of AS events induced during EMT [18]. ESRP expression is controlled by several transcription factors, including the aforementioned snail and TGF-β, a major regulator of EMT and EMT-

associated transcription factors [7]. TGF-β AS variants have been reported to be heterogeneously expressed in prostate cancer cells [31].

The list of EMT-related genes that have AS variants associated with cancer progression is a long one. For instance, skipping of exon 11 of E-cadherin, a cell-to-cell adhesion molecule that is downregulated in EMT, results in a splice variant that is upregulated in several cancers. Another interesting example is Rac1b, a splice isoform used as a signalling mediator instead of the final effector, whose expression is stimulated by metalloproteinase-3 and in turn upregulates Snail and induces EMT. It is important to refer that many AS events in EMT-related genes have been shown to be under the control of SRSF1 [7].

### 2.2.4. Immune escape

Tumour cells are identified by the organism as abnormal phenotypes, which activate immune responses. However, tumour cells are able to avoid recognition and destruction by immune cells. One of the mechanisms used by these cells to evade immune response is the use of unusual human leukocyte antigen (HLA) molecules such as HLA-G, which inhibits immunocompetent cells. This molecule is not expressed under normal conditions but it is highly expressed in various tumour types [7].

### 2.2.5. Insensitivity to growth inhibitors

The oncogenic process is also responsible for disrupting the normal cellular growth regulation. The most studied molecules in this class are tumour suppressors p53 and retinoblastoma protein [7].

The tumour suppressor p53 is a transcription factor that coordinates cell-cycle arrest in responses to many cellular stresses and injuries. A dominant-negative splice variant of p53 is DNp53. This variant lacks the first 40 amino acids of the wild type but it is still able to bind DNA, thus competing against wild-type p53 and affecting its normal function. Additionally, p53 is stabilised by SRSF1 binding to RPL5/MDM2 ribosomal-protein complex in the cytoplasm, so phosphorylation and nuclear localisation of SRSF1 are likely to lead to p53 degradation and proliferation [7]. This alteration has been detected in retinoblastoma [32, 33].

### 2.2.6. Growth factor self-sufficiency

In a normal state, the proliferation of cells is limited, being controlled by a complex network of signalling pathways responding to growth factors and their receptors. Through the abnormal modification of these pathways and expression of their messengers and effectors, tumours cells are able to limitlessly

replicate themselves. To achieve this limitless replicative potential, splicing patterns are often modified in order to preferentially express isoforms that promote and maintain the cell proliferation [7].

One of the growth factors playing a central role in the control of cell proliferation is the epidermal growth factor receptor (EGFR), which is a member of the receptor tyrosine kinases family. EGFR achieves this control through activation by EGF-ligand binding and downstream signalling mediators such as Akt, JAK/STAT or ERK. In several cancers, such as gliomas, prostate and ovarian cancers, a splice variant lacking exon 4 (de4 EGFR) is highly expressed. The missing exon translates functionally into a receptor that is constitutively active and promotes proliferation [7].

Another example is the BRaf mutation in over 50% of melanomas. BRaf is a member of the Raf kinase family of growth signal transduction protein kinases. Inhibitors against this mutant BRaf have been developed and are in clinical use. However, these isoforms develop resistance against these inhibitors [34].

KRas mutation is also associated with many cancers (most prominently colon cancer). Kras is a member of the Ras family of GTPases, which is an important element in cell proliferation control, differentiation and migration. Two splice isoforms that include alternate cassette exon 4 (KRas 4A and 4B) are strongly correlated with several colon cancer properties, such as left colon location, size of the tumour and histological subtype [7].

Similarly, mutations and deregulated activity of the PTEN tumour suppressor, a phosphatase essential for regulating the cell cycle, are described in many cancers. Two splice variants of PTEN, characterised by intron 3 and intron 5 retention, are strongly associated with breast cancer [7.

## 2.2.7. Cellular hyperenergetics

Under normal conditions the primary process by which cells produce energy is oxidative phosphorylation. When lacking proper oxygen supply, the glycolytic pathway is switched on. Cancer cells, independently of the amount of oxygen available, use glucose as their primary energetic source, via a process termed as aerobic glycolysis. This process, although much less efficient than oxidative phosphorylation, is used by tumour cells to produce needed intermediates to supply the high demands of biosynthesis [7].

PKM (pyruvate kinase), which has two splice isoforms, PKM1 and PKM2, is abnormally expressed in tumour cells. PKM1, which stimulates oxidative phosphorylation, is normally expressed in adult life, whereas PKM2 is a promoter of aerobic glycolysis that is normally only expressed during embryonic development. However, PKM2 is reported to be re-expressed in numerous cancers. The regulation of PKM isoforms ratio involves c-Myc pathway and ribonucleoproteins hnRNP A1, A2 and PTB [7].

## 2.2.8. Angiogenesis

Angiogenesis is the process by which new blood vessels are created. This is an import hallmark of cancer, since the creation of new blood vessels allows for a more efficient vascularization of the tumour.

In almost every form of cancer the main angiogenic molecules are VEGFs. These molecules act as principal mediators of metastasis through the lymphatic system. The VEGF family of ligands and receptors are regulated by AS. The most studied members of this family are VEGF-A isoforms (angiogenic $VEGF_{xxx}$ and $VEGF_{xxx}b$), VEGF-C and placental growth factor. The most well-characterised splice variants contributing to angiogenesis result from alternative splice sites found in the terminal exon 8 of the VEGF-A. One splice variant ($VEGF\text{-}A_{165}$) is highly angiogenic and upregulated in tumours, whereas the other splice variant ($VEGF\text{-}A_{165}b$) is expressed in normal tissues and downregulated in colon, renal and prostate cancer and metastatic melanoma [7].

The splicing pattern of VEGF is altered in various types of cancers. For instance, $VEGF_{165}b$ was first detected in renal cortex, yet it was not present in renal carcinoma [7]. In non-VHL renal cell carcinoma, VEGF is spliced exclusively as the pro-angiogenic forms, thus disrupting the 50-50 expression proportion of $VEGF_{165}b$ and $VEGF_{165}$ found in normal renal glomerular epithelial cells. There is downregulation of $VEGF_{165}b$ and/or upregulation of $VEGF_{165}$ in renal, prostate, melanoma, neuroblastoma, colorectal and bladder cancers [35].

## 2.3. Renal clear cell carcinoma

Kidney cancer, or renal cell carcinoma (RCC) are a common group of chemotherapy resistant diseases [36]. RCC is the twelfth most common cancer in the world and in Europe alone it is estimated that each year there are approximately 102 000 new cases and 45 000 deaths [37, 38]. The most common type of RCC is ccRCC, which underlies alterations in genes controlling cellular oxygen sensing (for example, VHL) and the maintenance of chromatin states (for example, PBRM1) [36].

A recent study revealed that, although the analysed ccRCC samples had fewer somatic copy number alterations (i.e. variations in the number of copies of sections of genomic DNA) than most cancers, when those were observed they more commonly involved the entire chromosome or chromosomal arms rather than focal events. The most frequent event involved the loss of chromosome 3p that encompassed all of the four most commonly mutated genes (*VHL*, *PBRM1*, *BAP1* and *SETD2*) [36].

The same study also observed arm level losses on chromosome 14q, associated with the loss of *HIF1A* (that plays an essential role in cellular and systemic responses to hypoxia) in 45% of the studied samples. Gains of 5q were observed in 67% of samples and additional focal amplifications refined the region of interest to 60 genes in 5q35. Focal amplification also implicated *PRKCI* (a protein kinase C member), and the *MDS1* and *EVI1* complex locus MECOM at 3p26, the p53 regulator *MDM4* at 1q32, *MYC* at 8q24 and *JAK2* on 9p24. Focally deleted regions included the tumour suppressor genes

*CDKN2A* at 9p21 and *PTEN* at 10q23, putative tumour suppressor genes *NEGR1* at 1p31, *QKI* at 6q26, and *CADM2* at 3p12 and the genes that are frequently deleted in cancer, *PTPRD* at 9p23 and *NRXN3* at 14q24 [36].

The same study also identified nineteen significantly mutated genes, with *VHL*, *PBRM1*, *SETD2*, *KDM5C*, *PTEN*, *BAP1*, *MTOR* and *TP53* representing the eight most extreme members [36].

## 2.4. Transcriptome studies

Each hallmark of cancer is associated with alterations in splicing patterns, suggesting that splicing regulation may play a crucial part in tumour evolution [7]. Transcriptome studies may therefore reveal keys to the understanding of cancer biology [8].

Recent advancements in NGS technologies are revolutionizing cancer genomic studies. Such methodologies can also be used to profile cancer transcriptomes and most molecular oncogenic mechanisms ultimately involve transcriptomic variation.

The analyses of GE and AS from RNA-seq datasets (comprising millions of transcriptomic sequence reads) can be performed using established bioinformatics tools, such as TopHat, Cufflinks and MISO.

The RNA-seq technology, as well as TopHat, Cufflinks and MISO, are described in the next sections. In addition, a summarized description of these tools is available on Appendix A.

### 2.4.1. RNA-seq

RNA-seq is a recently developed technology for transcriptome profiling that uses NGS to sequence DNA molecules reversely transcribed from RNAs. By using RNA-seq one can not only measure GE levels with an unmatched precision but also discover and quantify previously unknown transcripts and splicing isoforms [8, 39].

The RNA-seq protocol can be divided into 2 major steps:  cDNA library construction and sequencing.

#### 2.4.1.1. Library construction

The first step is to extract RNA from the sample to be studied. Generally the RNA molecules go through a selection process in order to guarantee that the sequencing capacity is mostly used on RNAs of interest. That selection process thus varies according to the molecules one intends to study. For instance, when the goal is to study mRNAs, oligo-dT primers are used to select the RNAs with poly-A tails. When the study target are micro RNAs (miRNAs), a size selection is the adopted selection method [8].

The subset is then fragmented into short pieces usually by RNA hydrolysis or nebulization [39]. Thereafter the fragments are transcribed into double-stranded cDNAs, which go through end-repair, 3' adenylation and adaptor ligation. Next the cDNAs go through PCR amplification and size selection, since the sequencing read length is limited. Finally, the cDNAs are denatured and ready for sequencing. The process is schematized in Figure 3.

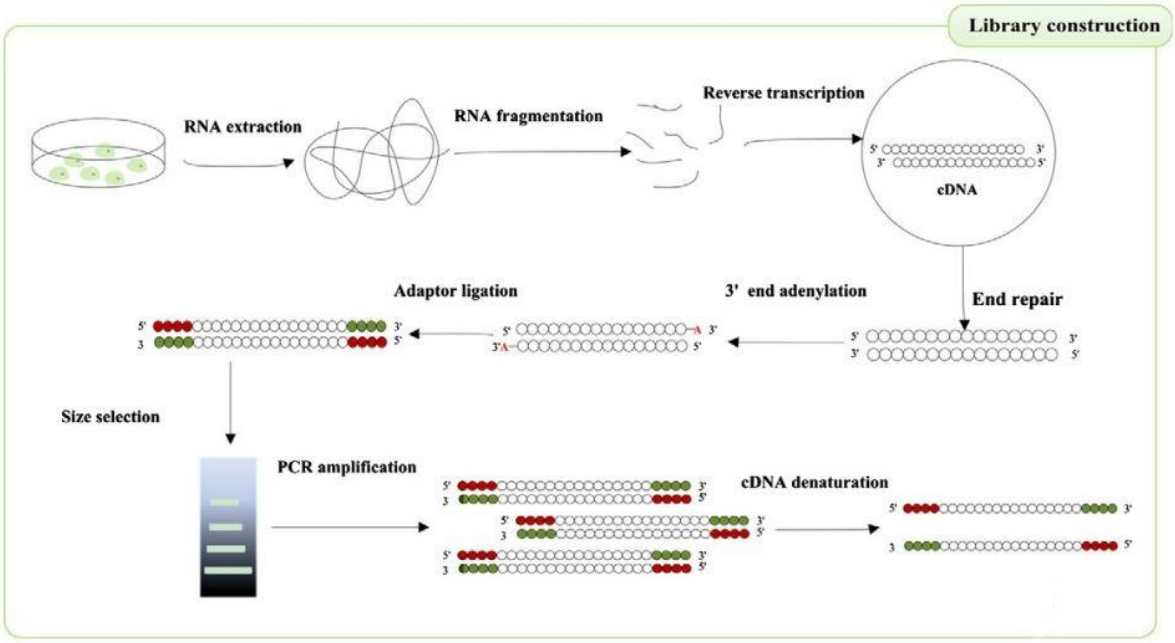



**Figure 3-** cDNA library construction phases. The extracted RNA is fragmented and reversely transcripted. The originating cDNA suffers end-repair, adaptor ligation, followed by PCR amplification, and ultimately it is denatured [8].

## 2.4.1.2. Sequencing

There are several NGS platforms available, having different sequencing lengths, costs per nucleotide and throughput. The most used platform is Illumina/Solexa, which has a very high throughput, a reasonable cost per nucleotide and sequencing lengths from 30 to 120 bases [8].

Illumina/Solexa reads transpire in the following manner, schematized in Figure 4: the single-strand cDNAs adhere to a flowcell by flexible linkers. After a number of bridge PCR cycles, clusters are formed and these are sequenced by synthesis at each cluster in parallel. Fluorescent complementing dNTPs are incorporated with the single-strand DNAs and a series of high-resolution digital images is captured. Using base calling software and image processing the sequences are read from the captured images. The obtained reads are commonly saved in a FASTQ format file (.fq extension), with nucleotide letters and quality scores accompanying each letter. Resorting to bioinformatics tools, such as TopHat [40], FASTQ files can be used to calculate the composition and abundance of RNAs by aligning the reads to a reference genome and then counting the number of reads mapping to each gene or transcript [8].

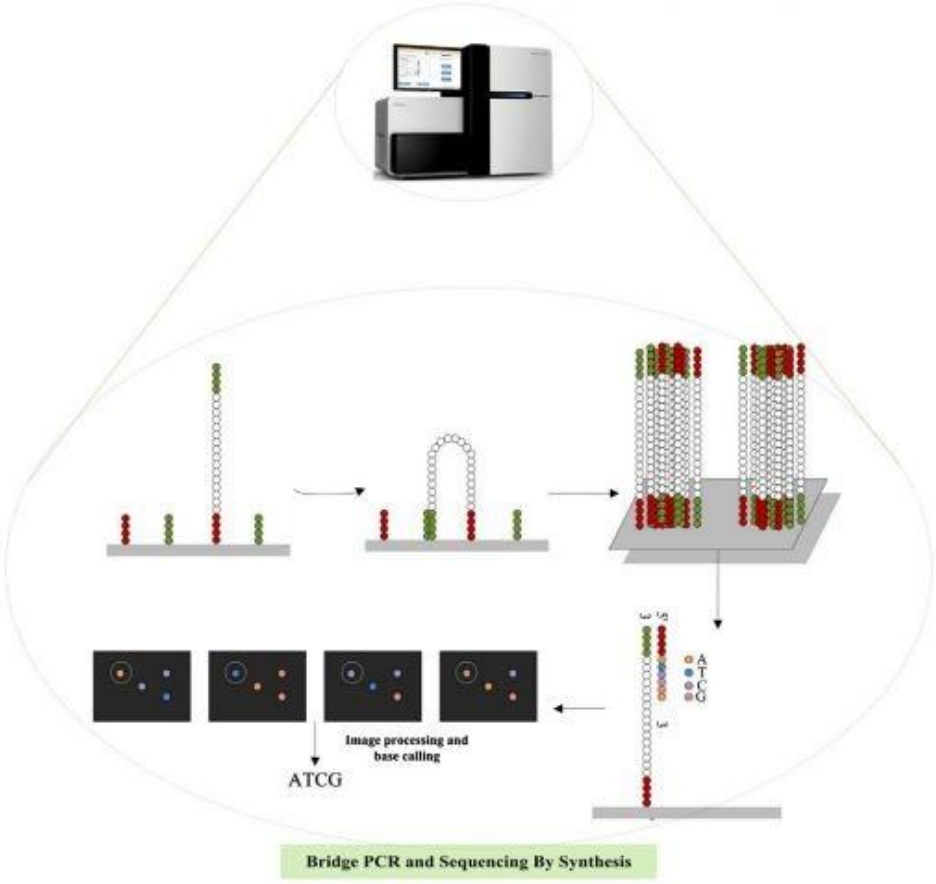



**Figure 4-** cDNA sequencing phases using Illumina/Solexa. Single-strand DNAs adhere to the flowcell by flexible linkers. They grow into clusters and, after a number of bridge PCR cycles, fluorescenced dNTPs are incorporated with the single-strand DNAs in the clusters according to nucleotide complementation. The sequences are read out from these images by image-processing and base-calling software [8].

## 2.4.2. TopHat

TopHat is a fast splice junction mapper for RNA-Seq reads. Contrary to most of the other currently available software for aligning RNA-seq data to a genome, TopHat does not rely on known splice junctions [40]. This particularity makes TopHat able to identify previously unknown splice variants of genes. TopHat is an efficient alignment software, mapping nearly 2.2 million reads per CPU hour (corresponding to TopHat using 100% of CPU during one hour) [40]. TopHat is implemented in C++ and Python, and runs on Linux and Mac OS X operating systems. It makes substantial use of Bowtie (an ultrafast memory-efficient short read aligner [41]), Maq (that builds mapping assemblies by aligning short reads to reference sequences [42]) and the SeqAn library (an open source C++ library of efficient algorithms and data structures for the analysis of sequences with the focus on biological data [43]).

TopHat finds junctions by mapping reads from FASTQ files to a reference genome. This is an efficient process that divides itself into 2 stages.

Firstly, all reads are mapped against a reference genome using Bowtie. Those that do not map to the

genome are set aside as *initially unmapped reads* (IUM reads). For each read, TopHat allows Bowtie to report one or more alignments (so called multireads). To avoid alignments to low complexity sequences, a maximum number of reported alignments is set (by default 10). When Bowtie reports more alignments than the maximum allowed, the read is simply discarded [40]. Note that the reported alignments are not mismatch-free. By default, 2 mismatches are allowed for the 5' end of each read. For the 3' end of each read additional mismatches are allowed. The accuracy of each base calling is classified using Phred quality scores, defined as a property which is logarithmically related to the base-calling error probabilities. These quality scores range from 4 to about 60, with higher values corresponding to higher quality. The maximum permitted total of quality values at mismatched read positions is the *quality-weighted hamming distance* or *Q-distance* which cannot be higher than 70 by default for reads in the 3' end (this threshold can be changed by the user) [40, 44, 45].

The second stage consists of assembling the mapped reads, resorting to the assembly module in Maq. First off, TopHat determines the island sequences (close sequences from the sparse consensus that are assumed to be putative exons). To accomplish this task, TopHat invokes Maq's *assemble* command, which generates a compact consensus file containing the called bases and the corresponding bases in the reference genome. These islands may have incorrect base calls, making them *pseudoconsensus*. Incorrect base calling might be due to sequencing errors in low-coverage regions [40]. TopHat includes a small amount of flanking sequence from the reference genome on both sides of each island (45 bp, by default), to guarantee that donor and acceptor sites from flanking introns will be captured. These sequences are added because most reads covering the ends of exons will also span splice junctions. Thus the ends of exons in the pseudoconsensus will initially be covered by few reads and, as a result, an exon's pseudoconsensus will likely be missing a small amount of sequence on each end [40].

Another important parameter of TopHat is the longest coverage gap allowed in a single island. Exons adjacent to a gap with a length shorter than the predefine value are merged into a single exon. By default, this value is 6 bp but, since introns shorter than 70 bps are rare in mammalian genomes, any smaller value is an acceptable threshold [40].

To map reads to splice junctions, TopHat first enumerates all canonical donor and acceptor sites within the island sequences. Then it considers all pairings of these sites that could form canonical GT-AG introns between close islands. These possible introns are checked against the IUM reads looking for reads that span the splice junction. By default, TopHat only examines introns longer than 70 bp and shorter than 20 000 bp, since this is the length of most known eukaryotic introns. If the user opts to, these values can be changed [40]. For each splice junction, TopHat searches the IUM reads to find reads that span junctions using a seed-and-extend approach. The seed is formed by combining a small amount of sequence upstream of the donor and downstream of the acceptor, as seen in figure 5 (seed - dark grey sequences) [40].

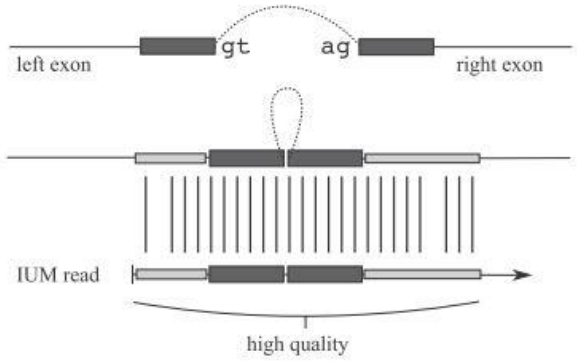



**Figure 5-** Seed-and-extend strategy adopted by TopHat. The seed, in dark grey, is originated from the combination of a small amount of sequence from both the acceptor and donor of the junction [40].

To improve running time and avoid reporting false positives, TopHat rejects donor-acceptor pairs from the same island, unless this island is highly covered. Figure 6 shows two alternative transcripts from one gene, one transcript having an intron that overlaps an untranslated region (UTR) of the other transcript. This region is highly covered, making it clear that both transcripts are present in the RNA-seq and so TopHat reports this whole region as a single island [40].

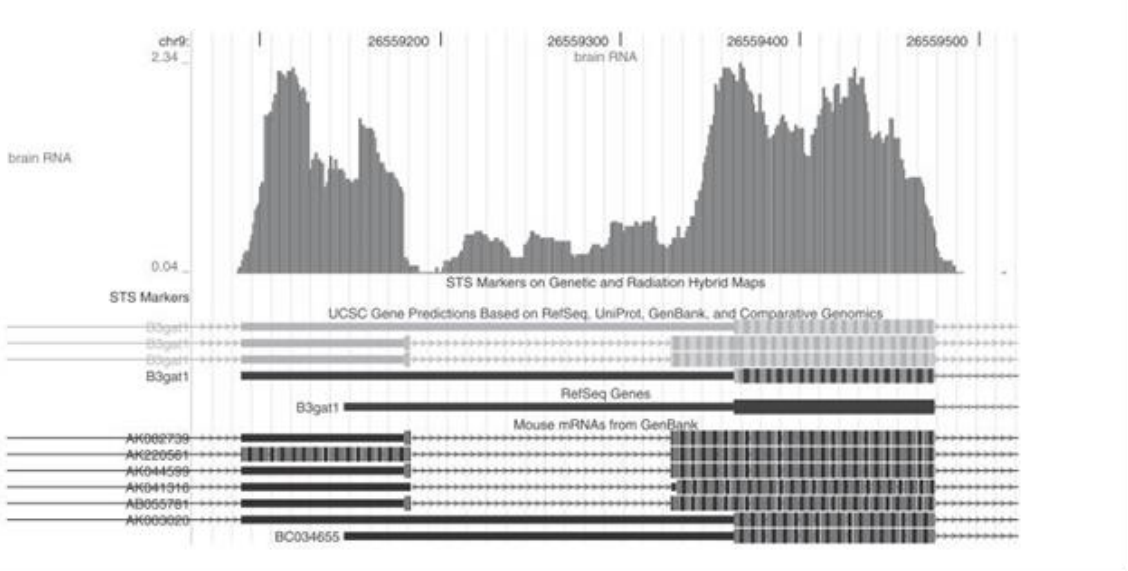



**Figure 6-** Example where donor and acceptor, of the junction, are from the same island. This junction is accepted because this region is highly covered [40].

Wang and colleagues [46] reported that analysis of mappings of sequence reads to exon-exon junctions indicated that 92-94% of human genes undergo AS, approximately 86% with a minor isoform frequency of 15% or more. Based on this conclusion, before reporting splice junctions, TopHat discards those that are estimated to occur at a frequency lower than 15% avoiding false junctions reports [40].

Finally, the algorithm reports all the spliced alignments it finds, and then it builds a set of non-redundant splice junctions using these alignments and the depth of coverage of the exons flanking them.

### 2.4.3. Cufflinks

After obtaining the alignment files from TopHat, Cufflinks is used. This software and its many packages are able to assemble transcripts, estimate their abundances and test for differential expression and regulation in RNA-Seq samples [47].

To quantify GE from RNA-seq data with precision, one needs to identify which isoform of each gene corresponds to each read. This cannot, obviously, be done without knowing all the isoforms of that gene. Relying on previously available transcriptome annotations may lead to inaccurate expression values if those are incomplete. In order to avoid this, Cufflinks assembles individual transcripts from the RNA-Seq reads that have been aligned to the reference genome. However, if the user opts to, the reads can be compared to a previously available annotation [48]. As previously explained, a gene may sometimes have multiple AS events, leading to multiple possible reconstructions of the gene model that explain the sequencing data. Therefore, Cufflinks reports a transcriptome assembly of the data containing the minimum number of full-length transcript fragments (transfrags) needed to justify all the splicing outcomes present in the input data [48].

Once the data is assembled, Cufflinks quantifies the expression level of each transfrag resorting to a rigorous statistical model in which the probability of observing each fragment is a linear function of the abundances of the transcripts from which it could have originated [49]. Based on these estimations Cufflinks will automatically discard transfrags that are significantly less abundant when compared to the others [48].

Another aspect to take into consideration when working with Cufflinks is the use of several replicate RNA-seq samples. Although the first instinct could be to bundle up the samples and analyse them together, this is not an efficient and effective approach. Firstly, because it becomes computationally demanding. Secondly, because the probability of incorrectly assembling the transcripts increases, given the added complexity that comes from analysing a potentially more diverse set of transcripts. Thirdly, merging samples leads to the loss of useful information about the biological/technical variability associated with replication, with implications on the statistical analysis of differential expression. The best strategy is therefore to analyse each RNA-seq sample individually and then merge the resulting assemblies using Cuffmerge [48].

Finally, another useful functionality of Cufflinks is the comparison between the obtained assemblies and a reference annotation using Cuffcompare. This allows users to identify previously unknown isoforms. Obviously, the user should validate experimentally these newly discovered isoforms, since this is a complex process and many errors may occur [48].

### 2.4.4. MISO

Mixture-of-isoforms (MISO) is a probabilistic framework that quantitates the expression level of

alternatively spliced genes from RNA-Seq data and identifies differentially regulated isoforms or exons across samples [50].

To detect AS, MISO uses sequence reads aligned to splice-junction sequences. These sequences can be pre-computed from known or predicted exon-intron boundaries or discovered *de novo* using software such as TopHat [51]. An annotation of the AS events must also be provided. Annotations are available for the major classes of AS and alternative RNA processing events in the human (hg18, hg19), mouse (mm9, mm10) and fruit fly (modENCODE) genomes [52]. These annotations indicate the constitutively and alternatively spliced isoforms associated with each AS event. Based on the provided annotation, MISO estimates the 'percent spliced in' (Ψ) associated with each event. Ψ is defined as the expression of constitutively spliced isoforms as a fraction of the total expression of both alternatively and constitutively spliced isoforms (Eq. 1) [51, 53]:

$$\Psi = \frac{\#\text{ of constitutively spliced isoforms reads}}{\#\text{ of constitutively spliced isoforms reads} + \#\text{ of constitutively alternatively isoforms reads}}, \quad (\text{Eq. } 1)$$

For instance, for a SE event the isoform containing a given cassette exon and the flanking constitutive exons is deemed the constitutively spliced isoform (inclusion reads) and the isoform containing only the flanking exons is the alternatively spliced isoform (exclusion reads) (Figure 7).

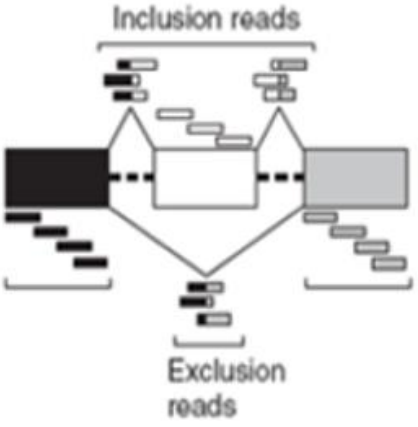



**Figure 7-** SE event. Inclusion reads: reads aligned against the alternative exon or its junctions; exclusion reads: reads aligned against to the junction between constitutive exons; constitutive reads: reads that align to the body of flanking exons.

The estimation algorithm is based on sampling and falls in the family of techniques known as Markov Chain Monte Carlo [50]. This estimation is endowed with several sources of bias in short read counts, including those due to the cDNA fragmentation and primer amplification steps of current RNA-seq protocols. Thus MISO outputs the lower and upper bounds of the 95% confidence interval on the Ψ estimate [53].

## 2.5. Hypothesis testing

A hypothesis test is the testing of an assumption about a population parameter. There are always two statistical hypotheses: the null hypothesis (is usually the hypothesis that any differences between samples result purely from chance, generally denoted by $H_0$) and the alternative hypothesis (the hypothesis that sample observations are influenced by some non-random cause, generally denoted by $H_1$). The outcome of a hypothesis test is *Reject $H_0$ in favour of $H_1$* or *Do not reject $H_0$*. Two types of errors can result from a hypothesis test [54, 55]:

- **Type I error:** the rejection of a null hypothesis when it is true. The probability of committing a Type I error is commonly known as $\alpha$.
- **Type II error:** the failure to reject of a null hypothesis that is false.

### 2.5.1. P-value and $\alpha$

The p-value attests for the robustness of a hypothesis test, being simply the probability of rejecting $H_0$ when that hypothesis is actually true. This value is a conditional probability, in that its calculation is based on an assumption (condition) that $H_0$ is true. This is the most critical concept to keep in mind as it means that one cannot infer from the p-value whether $H_0$ is true or false [56].

Obviously a smaller p-value indicates a more robust result and values smaller or equal to 0.05 are generally regarded as acceptable p-values for the rejection of the null hypothesis. When multiple comparisons are made it is necessary to employ methods that ensure that the accepted p-values are not compromised. These methods are necessary because when dealing with a large number of tests there is a high probability of observing at least one significant result just due to chance. For instance, considering an acceptable p-value of 0.05 and performing 30 tests, the probability of observing at least one significant result purely by chance is expressed in Eq. 2:

$$\begin{aligned} &P(at\ least\ one\ significant\ result) \\ &= 1 - P(no\ significant\ results) \\ &= 1 - (1 - 0.05)^{30} = 78.54\%, \end{aligned} \qquad (Eq.\ 2)$$

so the probability of getting at least one significant result, even if all tests are not actually significant, is 78.54%.

One of the most conservative and simpler multiple comparison correction methods is the Bonferroni correction. This method simply consists in dividing the initially accepted p-value, $\alpha$, by the total of comparisons made, $n$, to obtain the adjusted acceptable p-value, $\alpha_{adj}$ [57]:

$$\alpha_{adj} = \frac{\alpha}{n}. \qquad (\text{Eq. } 3)$$

The Bonferroni correction, however, is a very conservative one. Less conservative and more popular approaches include the Benjamini–Hochberg procedure, developed by Benjamini and Hochberg in

1995. It aims to control the false discovery rate (FDR), *i.e.* the proportion of *discoveries* (significant results) that are actually false positives [58]. This technique is quite simple, considering tests for which the acceptable FDR, $q$, is 5%. The first step is to order the p-values from the smallest to the largest and then rank them, being the smallest value ranked with $i = 1$ and the largest $i = n$. Finally, the largest p-value that respects Eq. 4 and all the smaller p-values are considered significant:

$$p_i < \frac{i}{n} q. \qquad \text{(Eq. 4)}$$

## 2.5.2. Student's t-test

In 1899, the prestigious brewery Guinness hired William Sealy Gosset and tasked him with studying the quality of the ingredients used to produce their beer. He soon realized that the statistical tests for averages available in that day and age were only suitable for large sample sizes and not for the small and cost-efficient samples available [59, 60]. Thus during a sabbatical leave Gosset developed Student's t-test which he described in the paper *The probable error of a mean* [61]. This article was published under the alias Student (hence the name of the method) due to a Guinness policy that prohibited their staff to publish their scientific findings [62].

Student's t-test allows for the comparison of the means of two populations, even small size ones, assuming that these follow a near normal distribution and the compared population have an equal interval scale. A t-test is a statistical hypothesis test, with $H_0$ being the hypothesis that the means of the compared populations are equal and $H_1$ the hypothesis that the means of the compared populations are different [63]. In a t-test we assume $H_0$ and then see if the data are sufficiently at odds with that assumption that we feel justified in rejecting $H_0$ in favour of $H_1$ [58]. The t-test statistic is expressed in Eq. 5:

$$\mathrm{t} = \frac{\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2}{\sqrt{\frac{\mathrm{s}_1}{\mathrm{n}_1} + \frac{\mathrm{s}_2}{\mathrm{n}_2}}}, \qquad \text{(Eq. 5)}$$

where $\bar{x}_1$ and $\bar{x}_2$ are, respectively, the means of the first and second set of values, $s_1$ and $s_2$ the standard deviation, and $n_1$ and $n_2$ the size of each set of values. If one considers paired t-samples (where the compared populations are measurements from the same subjects but in different conditions) the applied formula changes slightly (Eq. 6), since the standard deviation ($s$) and size ($n$) is the same for both populations [64]:

$$t = \frac{\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2}{\frac{\mathrm{s}}{\sqrt{\mathrm{n}}}}. \qquad \text{(Eq. 6)}$$

One rejects $H_0$ if $|t| > t_\alpha$, where $t_\alpha$ is a tabulated value that depends upon the legth of the analysed samples and $\alpha$ [65].

Nowadays, Student's t-test applications go far beyond quality tests for beer ingredients. This popular

test is widely used in the fields of marketing, biology and medicine, among others.

## 2.5.3. Wilcoxon signed-rank test

If the analysed populations do not meet all the assumptions under which the t-test operates, we cannot reasonably apply that test. A non-parametric alternative for paired student's t-test is the Wilcoxon Signed-Rank test. This test is suitable for populations that do not follow a normal distribution. As one would expect, the $H_0$ and $H_1$ hypotheses are the same as those of the t-test [66]. The procedure for this test is the following [67]:

1. For each item in a sample of $n$ items, the difference score $D_i$ is computed, between the two paired values.
2. The + and - signs are neglected and the set of $n$ absolute differences is listed.
3. Any absolute difference score of zero is omitted from further analysis, thereby yielding a set of $n$' nonzero absolute difference scores, where $n' \leq n$. After removing values with absolute difference scores of zero, $n'$ becomes the actual sample size.
4. Ranks from 1 to $n'$ are assigned to each $|D_i|$, the smallest absolute difference score gets rank 1 and the largest gets rank $n'$. If two or more $|D_i|$ are equal, the mean of the ranks they would have been assigned individually, if there were no repeated values, is assigned to each.
5. The + or - signs are reassigned to each of the $n'$ ranks, resulting in $R_i$, depending on whether $D_i$ was originally positive or negative.
6. The Wilcoxon test statistic, $W$, is computed as the sum of the positive ranks (Eq. 7):

$$W = \sum_{i=1}^{n'} R_i^{[+]}. \qquad (Eq.7)$$

For large samples ($n' > 20$) the $Z_{stat}$ test statistic is expressed by Eq. 8:

$$Z_{stat} = \frac{W - \frac{n'(n'+1)}{4}}{\sqrt{\frac{n'(n'+1)(2n'+1)}{24}}}. \qquad (Eq.8)$$

One rejects $H_0$ if $|\mathbf{Z_{stat}}| > Z_\alpha$, where $Z_\alpha$ is a tabulated value that depends upon the size of the analysed samples and $\alpha$ [68].

## 2.5.4. One-sample Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test (K-S test) is a non-parametric test for the equality of continuous, one-dimensional probability distributions. This test can be used to compare two samples (two-sample K–S test) or to compare a sample with a reference probability distribution (one-sample K–S test), thus allowing the user to verify if the analysed sample distributes itself in a similar fashion to the reference probability distribution [69]. The one-sample K-S test is commonly used to test for normality.

The one sample K-S test is used to decide whether or not to reject $H_0$ (the analysed sample comes from the reference probability distribution) in favour of $H_1$ (the analysed sample does not come from the reference probability distribution). Considering a random sample $X = [x_1, x_2, \dots, x_n]$ of size $n$ with unknown distribution denoted by $F(x)$, let $S(x)$ be the empirical distribution function based on the random sample $X$ and $F^*(x)$ a completely specified hypothesized distribution function. The K-S test consists in the comparison of $S(x)$ with $F^*(x)$ to see if there is a good agreement. Let $T$ be the test statistic, the greatest vertical difference between $S(x)$ and $F^*(x)$ [70].

$$T = \sup_x |F^*(x) - S(x)|. \qquad (\text{Eq. } 9)$$

If $T > 1 - \alpha$, $H_0$ is rejected at the level of significance $\alpha$.

## 2.6. Dimension reduction and regression methods

Regression analysis is a statistical method used to estimate the relationship among variables. In particular, the basic model of Multiple Linear Regression (used to find correlation between a set of independent variables and a given response) is expressed in Eq. 10:

$$Y = \hat{\beta} X + \varepsilon, \qquad (Eq. 10)$$

where $Y$ is the response of the model, $X$ is the set of observations that explain the response $Y$, $\hat{\beta}$ is a vector containing the estimated coefficients and $\varepsilon$ is the independent identically distributed normal error which is minimized in order to find $\hat{\beta}$.

Dimension reduction methods are crucial when trying to decrease the number of predictors of a model. These methods are widely used in genomics, for instance. One of the more popular ones is the Tikhonov regularization also known as ridge regression [71]. This method is similar to least squares regression but the estimated coefficients tend towards zero, thus providing a larger decrease of the number of predictors used to explain the outcome of the model when compared to the latter method [72]. Ridge regression is expressed by the following equation:

$$\hat{\beta} = \operatorname{argmin}_{\beta_0, \beta \in \mathbb{R}} \left( \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right), \qquad (Eq. 11)$$

where $\hat{\beta}$ is a vector containing the estimated coefficients, $n$ is the number of observations in the data set, $y_i$ is the response at observation $i$, $\beta_0$ is a scalar that represents the interception of the function

derived by $\beta$ vector, which contains the coefficients attributed to each variable, $x_i$ is the observations registered for each predictor and $\lambda$ is a positive regularization parameter. A bigger $\lambda$ value reduces the number of nonzero components of $\beta$. $\sum_{j=1}^{p}\beta_j^2$ is known as $L^2$ norm. Recently (1996) a similar method to ridge regression has been developed by Robert Tibshirani the lasso regression [73]. This method offers a more radical dimension reduction than the ridge regression. It is expressed by:

$$\hat{\beta} = \text{argmin}_{\beta_0,\beta\in\mathbb{R}}\left(\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right), \qquad (Eq.\,12)$$

$\sum_{j=1}^{p}|\beta_j|$ is known as $L^1$ norm. In 2003, Hui Zou and Trevor Hastie developed the elastic net regression [74]. This method is more flexible than the latter two since it combines the $L^1$ and $L^2$ norms, resulting in the combination of ridge and lasso regression, and it is expressed by the following equation:

$$\hat{\beta} = \text{argmin}_{\beta_0,\beta\in\mathbb{R}}\left(\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\left(\alpha|\beta_j| + \frac{1-\alpha}{2}\beta_j^2\right)\right), \qquad (\text{Eq. } 13)$$

where $\alpha$ is a scalar, with a value between 0 and 1, that mediates the weight given to $L^1$ and $L^2$ norms. When $\alpha = 1$ elastic net regression is the same as lasso regression and when $\alpha = 0$ elastic net regression tends to ridge regression [73].

The aforementioned methods are often used with logistic regression. Logistic regression is a linear model for classification thus, it minimizes a *hit or miss* cost function rather than the sum of square residuals (as in ordinary regression) [74, 75]. It is applied to models with binary outcomes and it is expressed by the following equation:

$$P\left(y = 1|\, x_i, \hat{\beta}\right) = \frac{e^{x_i\hat{\beta}}}{1 + e^{x_i\hat{\beta}}}. \qquad (\text{Eq. } 14)$$

Cross-validation is a *standard* method for evaluating the performance of a model. This method consists in the separation of the data set into training set or sets (which are used to develop the model) and a test set, which is set aside and used to test the performance of the model. There are several kinds of cross validation. The most basic one is the holdout method, which consists simply in randomly dividing the available data into two sets: a training set used to develop the model and a test set used to evaluate the model. A more complex method is K-fold cross-validation. In this kind of cross-validation the data is randomly divided into K subsets, then K-1 subsets are used to compute the model and the subset left out serves as a test set. This procedure is repeated K times, ensuring that each data point is used both as training and testing data. This method is less efficient than the holdout model, yet it provides a more robust approach [75]. The performance of the model is measured using either mean squared error (Eq. 15):

$$mean\ squared\ error = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2, \qquad (Eq.\,15)$$

where $\hat{Y}_i$ is the predicted value for observation $i$ and $Y_i$ its true value. When evaluating models with continuous outcomes or deviance (D, Eq. 16) for models with dichotomous outcomes:

$$D(y) = -2\left(\log\left(p(y|\hat{\theta}_0)\right) - \log\left(p(y|\hat{\theta}_S)\right)\right), \qquad (Eq.16)$$

where $\hat{\theta}_0$ denotes the values of the parameters for the fitted model, while $\hat{\theta}_S$ denotes the parameters for the saturated model.

Once the binary model is developed, it needs to be tuned, *i.e.* we need to estimate the threshold that differentiates 0 and 1 response of the model. To that end *receiver operating characteristic* (ROC) is generally the method of choice. This metric is used to check the quality of classifiers. For each class of a classifier, ROC applies threshold values across the interval [0, 1] to outputs. For each threshold, two values are calculated, the True Positive Ratio (the number of outputs greater or equal to the threshold, divided by the number of one targets), and the False Positive Ratio (the number of outputs less than the threshold, divided by the number of zero targets) [76]. The most commonly used global index of diagnostic accuracy is the area under the ROC curve (AUC). Values of AUC close to 1.0 indicate that the marker has high diagnostic accuracy.

## 2.7. Correlation

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. There are many correlation tests, such as Pearson product-moment and Spearman's rank.

Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two variables [77].

Similarly, Spearman's rank correlation coefficient is used to measure the strength of a monotonic (but not necessarily linear) association between two variables [78].

## 2.8. Survival analysis

Survival analysis consists of time-to-event analysis, *i.e.* the time until an occurrence of interest takes place such as death, infection and mechanical failure, among others [79]. A basic and important concept in survival analysis is censoring. If for a given subject the event of interest does not occur while the study is in progress or the subject abandons the study for some reason, the exact time for the occurrence of the event of interest is unknown and the subject is censored [79].

A very common method used in survival analysis is Cox Regression Model or Proportional Hazard. This model is a statistical method for investigating the effect of several variables (covariates) upon the time a specified event takes to happen [80]. The method does not assume any particular *survival model* but it is not truly non-parametric because it does assume that the effects of the predictor variables upon survival are constant over time and are additive in one scale [81]. The Cox regression model is expressed by the following equation:

$$h(t) = h_0(t)\, exp(\beta' x), \qquad \text{(Eq. 17)}$$

where $h_0(t)$ is a baseline, $\beta$ is a p-vector of regression coefficients (which give the proportional change that can be expected in the hazard related to changes in the corresponding covariate) and $x$ is a vector of p covariates [81].

In survival analysis, another useful method is the Kaplan-Meier estimator. This method allows for the estimation of the proportion of the population that would survive a given length of time under the same circumstances [80]. This estimator follows this simple expression:

$$S(t) = \frac{number\ of\ subjects\ at\ risk\ at\ time\ t}{total\ number\ of\ subjects}. \qquad \text{(Eq. 18)}$$

This expression allows for the construction of a survival curve, which is a step function. The comparison of survival curves corresponding to two or more groups of subjects is done resorting to logrank test [80].

This is a non-parametric hypothesis test widely used in clinical trials. For instance, it allows for the comparison of the survival of two different groups subjected to different conditions (for example one group taking drug A vs. one taking drug B).

## 2.9. Knowledge-based tools for biological interpretation of results

Several knowledge-based bioinformatics tools (*i.e.,* systems based on stored complex information, structured or not) can be used for the biological interpretation of data analysis results, namely in order to get a sense of which biological processes are most prominent to the biological phenomena under study or to identify associated putative changes in critical pathways. There are many online tools developed for that purpose, such as DAVID and GSEA described below.

### 2.9.1. DAVID bioinformatics resources

The Database for Annotation, Visualization and Integration Discovery (DAVID) consists of an online (*http://david.abcc.ncifcrf.gov)* integrated biological knowledge base and analytic tools aiming at systematically extracting biological meaning from large gene/protein lists [82].

High-throughput genomic, proteomic and bioinformatics scanning approaches allow for the study of a significantly large number of biological mechanism which often results in an extensive list of *interesting* genes. Biological analysis and interpretation of these results is a complex matter for which DAVID proves itself useful. The procedure starts by simply uploading a list of genes of interest. DAVID currently collects and integrates over 40 publicly available annotation categories, including GO terms, protein–

protein interactions, protein functional domains, bio-pathways, gene functional summaries, gene tissue expression and literature, among others [82].

To take advantage of the wide variety of annotation categories that DAVID integrates, the user has at his disposal useful tools such as: Gene Name Batch Viewer (searches functionally related genes within and out the uploaded list), Gene Functional Classification (reduces large lists of genes into functionally related groups of genes, thus unravelling the biological content captured by high throughput technologies) and Functional Annotation (gene-annotation enrichment analysis, functional annotation clustering (Figure 8), BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more) [83].

To calculate the degree of enrichment, DAVID takes into account the background of genes. If the user does not define a background DAVID will, by default, set the human genome as background. If the user does decide to define a background, it can do so simply by uploading the background gene list. Although there is no gold standard to define the background, a general guideline is to use the list of genes that have a chance of being selected in the study [82]. The measure of the magnitude of enrichment is quite simple. As an example, let's assume 10% of user's genes are kinases, knowing that 1% of genes in the human genome (background population) are kinases. Thus, the fold enrichment is tenfold. Fold enrichment 1.5 and above are generally regarded as interesting [84]. Associated to the fold enrichment value is the EASE score, a modified and more conservative Fischer's exact p-value. Let's consider a hypothetical example: in the human genome background (30000 genes in total), 40 genes are involved in the p53 signalling pathway (Table 2). In a given list of 300 genes, 3 are found to belong to the p53 signalling pathway. Then we ask the question if 3/300 is more than random chance when compared to the human background of 40/30000 [85].

| | User Genes | Genome |
|---|---|---|
| In Pathway | 3-1 | 40 |
| Not In Pathway | 297 | 29960 |

**Table 2-** Hypothetical example of Fischer exact p-value and EASE p-value calculation. EASE p-value calculation is a more conservative approach which considers 2 (3-1) instead of 3 in t [85].

Applying a multivariate generalization of the hypergeometric probability function, the Fischer's exact p-value is:

$$\mathrm{p} = \frac{\binom{40+3}{3}\binom{29\,960+297}{297}}{\binom{29\,960+3+40+297}{297+3}} = 0.008, \qquad \text{(Eq. 19)}$$

whereas the EASE score is:

$$\mathrm{p} = \frac{\binom{40+2}{2}\binom{29\,960+297}{297}}{\binom{29\,960+2+40+297}{297+2}} = 0.06, \qquad \text{(Eq. 20)}$$

Considering the Fischer's exact p-value, the user could consider that the analysed gene list was specifically associated (enriched) in p53 signalling pathway ($p<0.05$). However, considering the EASE score this user gene list is specifically associated (enriched) in p53 signalling pathway no more than random chance ($p>0.05$) [85].

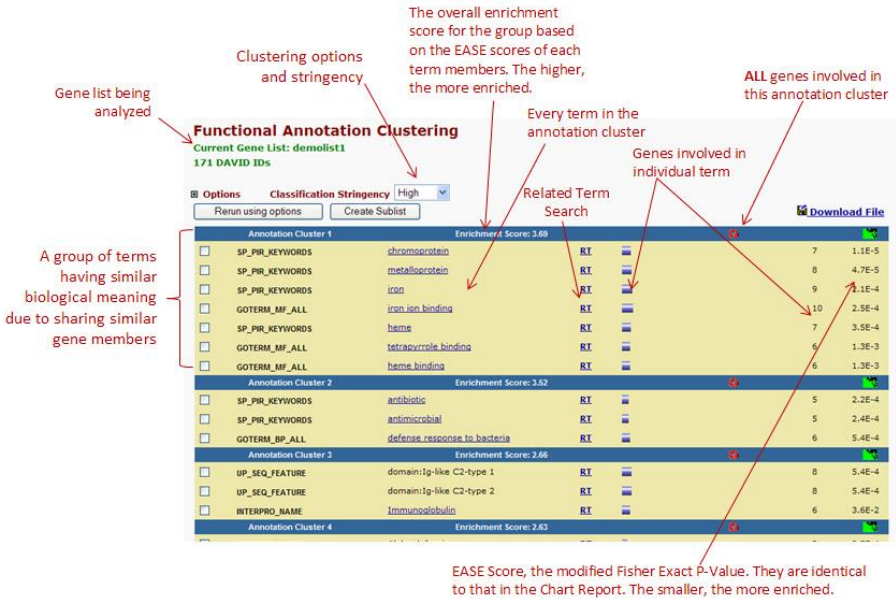



**Figure 8-** Functional Annotation Clustering Interface with captions [86].

## 2.9.2. Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether a previously defined set of genes shows statistically significant, concordant differences between two biological states (classes) [87].

The user must provide a GE dataset. The genes of that dataset will be ranked based on the correlation between their expression and the class distinction by using any suitable metric producing an ordered list $L$. Given an *a priori* defined set of genes $S$ (prepared by the user or one of the multiple gene sets from reliable databases that GSEA integrates), the goal of GSEA is to determine whether the members of $S$ are randomly distributed throughout $L$ or primarily found at the top or bottom. It is expect that sets related to the phenotypic distinction will tend to show the latter distribution. There are three fundamental steps in the GSEA method [87].

**Step 1:** Calculation of the enrichment score (ES) that reflects the degree to which a set $S$ is overrepresented at the extremes $L$. The score is calculated by walking down the list $L$, increasing a running-sum statistic when a gene that is represented in $S$ is found and decreasing it if the gene does not belong. The magnitude of the increment depends on the correlation of the gene with the phenotype. The ES is the maximum deviation from zero as shown in Figure 9.

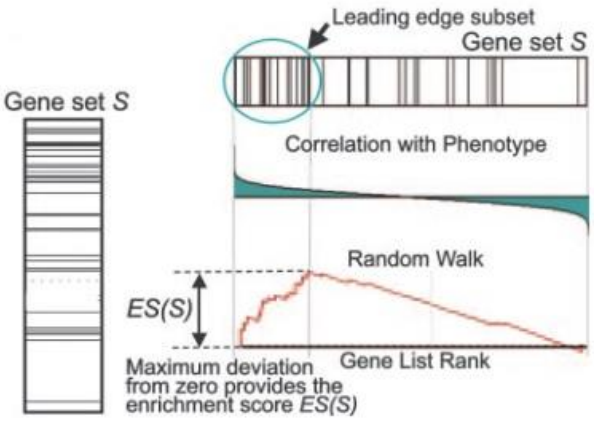



**Figure 9-** The enrichment thought the analysed gene list is represented by a plot that always starts and finish at zero. The ES is the maximum deviation from zero [87].

**Step 2:** Estimation of Significance Level of ES (p-value). The estimate is done resorting to an empirical phenotype-based permutation test procedure that pre-serves the complex correlation structure of the GE data (the user can set the number of permutations made, being that a higher number of permutations will be more reliable). The phenotype labels are permuted and the ES of the gene set for the permuted data is recomputed, thus generating a null distribution for the ES. The empirical, nominal p-value of the observed ES is then calculated relative to this null distribution.

**Step 3:** Adjustment for Multiple Hypothesis Testing. When an entire database of gene sets is evaluated, the estimated significance level is adjusted to account for multiple hypothesis testing. First, the ES is normalized for each gene set accounting for its size, yielding a normalized enrichment score (NES). Then the proportion of false positives is estimated by calculating the FDR corresponding to each NES. The FDR is the estimated probability that a set with a given NES represents a false positive finding. It is computed by comparing the tails of the observed and null distributions for the NES.

# 3

# Methods and Materials

## 3.1. Data description

This section describes the methods applied in the pre-processing of the dataset for analysis, in the identification of cancer-specific AS patterns, in the identification of AS patterns that serve as cancer stage classifiers, in the identification of AS prognostic factors and in the gene enrichment analysis.

The analyses were conducted on RNA-seq data for 62 ccRCC and their matched normal kidney samples and 76 ccRCC available from the TCGA project. The analysed cohort is made up by 67 stage I (48.75%), 14 stage II (10.14%), 29 stage III (21.01%) and 28 stage IV (20.29%) ccRCC patients (a more detailed description is available in Appendix B). Information on the AS events analysed is available in Appendix C.

## 3.2. Dataset preparation

The first step in preparing the RNA-Seq dataset for AS analysis is the alignment of the RNA-seq reads to the reference genome (hg19) using TopHat. Resorting to the set of non-redundant splice junctions thereby obtained, one can quantify the expression level of alternative spliced genes using MISO. The generated MISO output files are divided by patient, AS mechanism and tissue status (tumour or normal). Among other information, those files contain the AS event ID, the estimated mean Ψ for each AS event, as well as the lower and upper bounds of the 95% confidence interval of the Ψ estimate. To simplify the proposed analysis all observations were compiled into Excel book files, divided only by AS mechanism and tissue status. These files contain the ID of the patient associated with each observation, as well as the patient's ccRCC stage, vital status, number of follow-up days, and whether metastases were found on that patient or not. Additionally, each file contains a sheet with information on each analysed AS event, namely its associated spliced gene and genomic coordinates. The files were generated using a MATLAB script [88]. A scheme of the implemented procedure to obtain the estimated Ψ values is available in Figure 10.

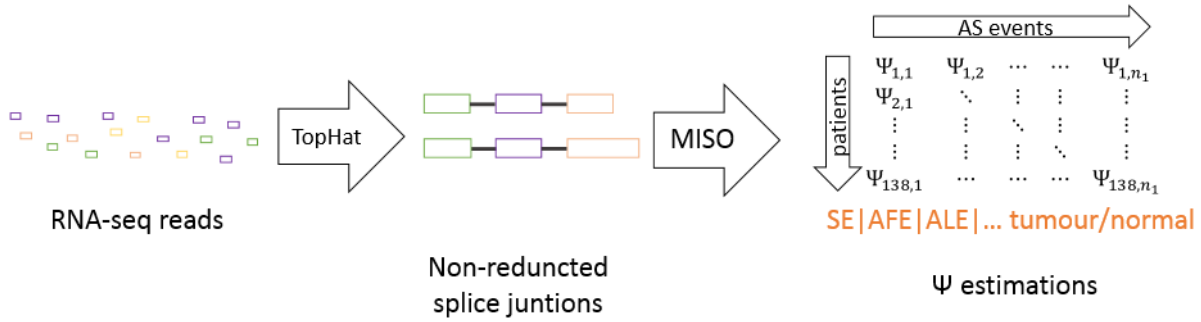



**Figure 10 –** General scheme of dataset preparation. For organizational proposes Ψ estimates were divided by AS mechanism and tissue status. These estimates can be seen as a matrix where each column represents an AS events and each row a patient. The Ψ estimates

Cufflinks was used to quantify GE, based on the TopHat alignments. The GE is measured in fragments per kilobase of transcript per million mapped reads (FPKM) [89].

## 3.3. Identification of cancer-specific AS patterns

To identify cancer-specific AS patterns, the MISO output files were used. Valid observations for both tumour and normal tissue of the same patient were required for an AS event in that patient to be considered in the analysis.

The identification of cancer-specific AS patterns was made by analysing the difference between the median Ψ of tumour and normal observations for each AS event ($\Delta\tilde{\Psi}$, expressed in Eq. 21):

$$\Delta\tilde{\Psi} = \tilde{\Psi}_{tumor} - \tilde{\Psi}_{normal}, \qquad (Eq. 21)$$

where $\tilde{\Psi}_{tumor}$ and $\tilde{\Psi}_{normal}$ are the median of the set of tumour and normal observations, respectively. This difference was analysed resorting to a suitable hypothesis test for the difference between two datasets, associating a p-value to that difference.

To select the most suitable method to test the difference between the Ψ registered for normal and tumour tissues, the K-S test was used (resorting to a MATLAB function). This test allows for the checking of the normality of the distributions of the tumour and normal sets of observations for each event. It was concluded that the majority of observation sets do not follow a normal distribution thus one cannot, in good conscious, apply the parametric paired t-test to test for differences between tumour and normal observation sets. A non-parametric test should be used instead. Wilcoxon Signed Rank test was the selected method.

The Wilcoxon Signed Rank function was applied resorting to MATLAB. As a considerable number of tests were done, multiple testing corrections were necessary. To ensure maximum confidence in the selected results, Bonferroni correction was used (with $\alpha = 0.01$). Additionally, only events that registered a $|\Delta\tilde{\Psi}| > 0.2$ were initially accepted. A general scheme of the procedure used to identify significant differences between the median Ψ of normal and tumour samples is available in Figure 11.

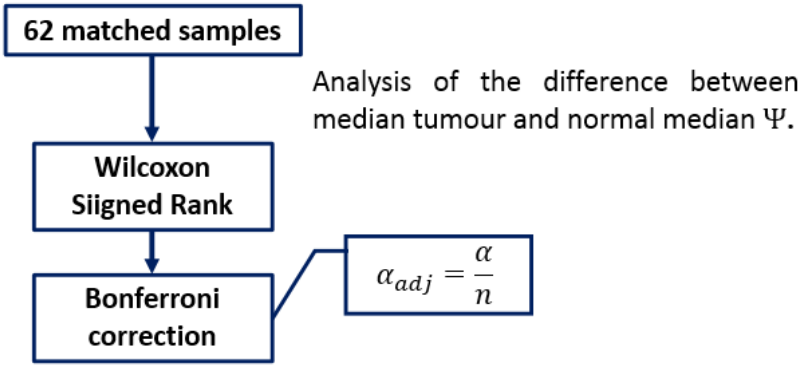



**Figure 11 -** General scheme of the identification of cancer-specific AS patterns.

## 3.4. Binary tumour stage classifier

Cancer stage classification based on RNA-seq data is a complex matter, in an attempt to simplify the proposed problem we chose to develop a binary classifier. To that end MISO Ψ estimates of 138 ccRCCs were used. To prepare the dataset, we first selected AS events that had MISO Ψ estimates registered for all 138 patient's tumour tissues we were analysing. From the initial 106206 events, only 18291 were selected. The MISO Ψ estimates were arranged in a matrix where each row corresponded to a patient and each column to an AS event. Afterwards we classified each patient with 1 or 0 according to their tumour stage. To understand which stage separations provided better results, different classification systems and combinations were used.

- Patients with stage I cancer were classified as 0 whereas patients with stages II, III and IV cancer were classified as 1.
- Patients with stages I and II cancer were classified as 0 whereas patients with stages III and IV cancer were classified as 1.
- Patients with stages I, II and III cancer were classified as 0 whereas patients with stage IV cancer were classified as 1.

Using the *lassoglm* Matlab function, logistic regression, using multiple $\alpha$ values, was done on the data of 80 randomly selected patients (the data of the remaining 58 patients was set aside to be used as test data). Note that when $\alpha$<1 the employed method is elastic net regularization whereas when $\alpha$=1 the employed method is lasso. Regression was used to conduct a dimensional reduction of the AS events that classify cancer stages. Different combination of parameters and classification systems were tested to identify which combination provided more reliable results. For each classification system described above, regression was done using various $\alpha$ values (0.1 to 1, with a 0.1 increment) and $\lambda$ values (0.01 to 1, with a 0.01 increment). To select the classification system and parameters that provided more reliable results, the estimated $D$ for each estimated model was analysed. $D$ was estimated with the 10-fold cross-validation method. The $\hat{\beta}$ which had the minimum $D$ associated to it was selected. Using a ROC curve we chose the optimum threshold (the one maximized the module $\|True\ Positie\ Rate - False\ Positive\ Rate\|$) that separates between the classification of 0 or 1. A general scheme of the steps involved in the development of this classifier is available in Figure 12.

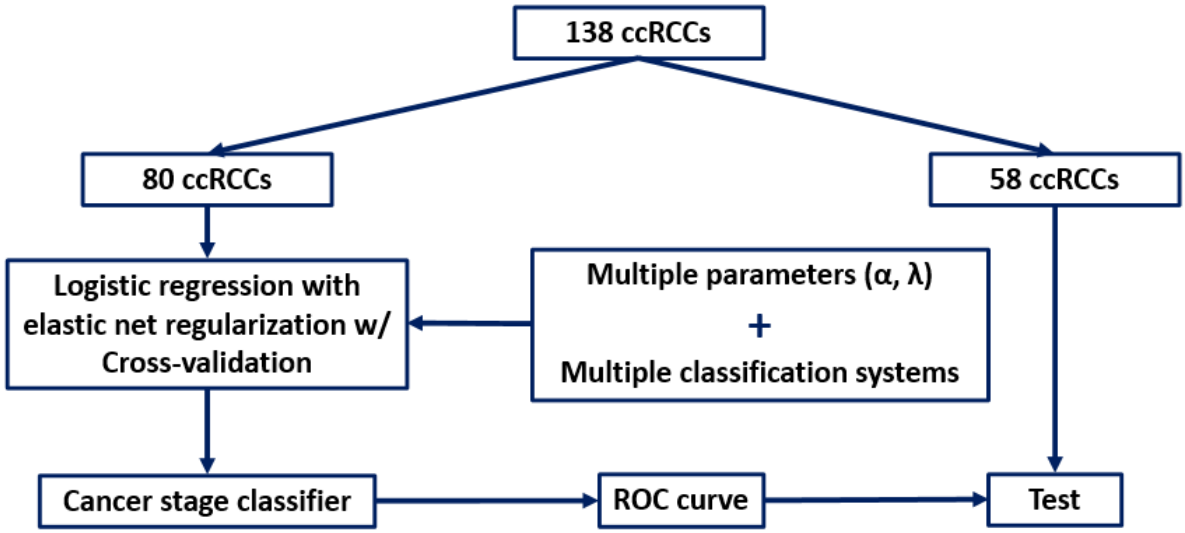



**Figure 12 –** General development scheme of the binary cancer stage classifier.

## 3.5. Identification of independent AS prognostic factors

MISO Ψ estimates were also used to identify AS prognostic factors. Taking into account the available patient info, namely the number of follow up days and the vital status, survival analysis was carried out in both normal and tumour samples for each AS event. Normal tissue samples from 61 patients and tumour tissue samples from 137 patients were used. Note that the samples of 1 patient were exclude from these analysis since the number of follow up days was not available for this patient.

The identification of AS prognostic factors was done with the following approach. For each event, the observations were initially sorted according to their Ψ value, from smallest to the largest. Next, they were divided into 2 initial groups: Low PSI (composed by the 15 observations with the smallest Ψ, in normal tissue analysis and 35 observations in tumour tissue analysis) and High PSI (composed by the remaining observations in both tumour and normal tissue analysis). Note that, to ensure that both groups had a significant number of observation, only events with at least 30 observations in normal tissue and 70 in tumour tissue were analysed (thus guarantying that neither group had less than 15 and 35 observations in normal and tumour tissue analysis, respectively). A logrank test was then applied to analyse the difference between the survival estimates of the two initial groups and the p-value, as well as the number of observations that made up each group, was recorded. Then a redistribution of the observations was done: the observation with the smallest value in the High PSI group was excluded from that group and added to the Low PSI group, the logrank test being then applied again. The resulting p-value was compared to the smallest p-value recorded. If the resulting p-value was smaller it was recorded, as well as the number of observations that made up each group, otherwise it was discarded. Note that each distribution of observations was only considered valid if Low and High PSI groups did not share any equal Ψ values. If there were any common Ψ values between the two groups, a redistribution was done: any patients from the High PSI group that had the shared Ψ value was integrated into the Low PSI group. The process goes on recursively until either group has a number of

observations smaller than 35 in tumour tissue and 15 in normal tissue analysis. This method ensures that the distribution that maximizes the survival separation between PSI groups is considered for each event. A general scheme of this procedure is available in Figure 13.

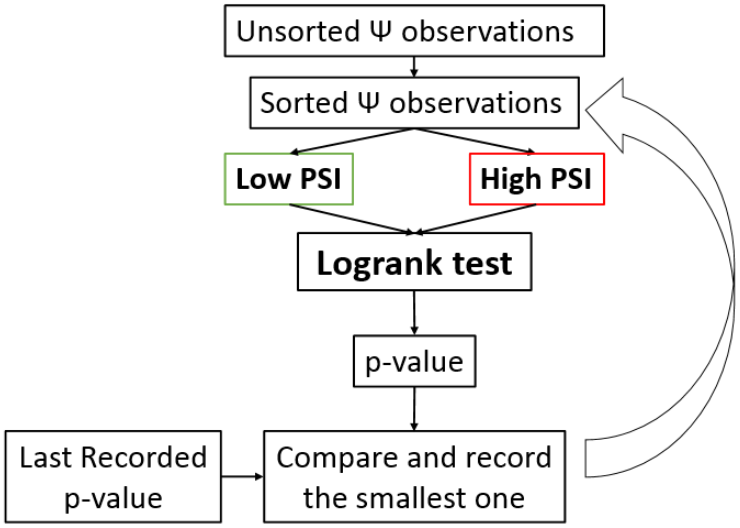



**Figure 13 –** Schematic of the procedure implemented for identifying AS prognostic factors.

Multiple testing correction of the resulting p-values was done with the Benjamini–Hochberg procedure (with α=0.05) to select the AS events that significantly associate with survival. As an additional selection parameter, only AS events that register a difference equal or larger than 0.3 between the smallest and largest Ψ values were considered. With this selection step we guarantee that the segregation between High and Low PSI groups is more effective, avoiding a concentration around a small range of Ψ values. We ultimately selected, from each of both normal and tumour sample groups, the 2 events with the smallest p-value associated to the Logrank test for further analysis.

## 3.6. GSEA

Resorting to Cufflinks, 46533 GE values were measured for the 62 ccRCCs and their matched normal tissues.l GSEA was conducted comparing High and Low PSI groups, with GE signatures from any ultimately selected AS prognostic factor. Molecular Signatures available in the GSEA site (http://www.broadinstitute.org/gsea/msigdb/index.jsp), specifically the c2 (curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts) and c6 (oncogenic signatures defined directly from microarray GE data from cancer gene perturbations) collections, were used.

# 4

# Results

## 4.1. Cancer-specific AS patterns

Using the aforementioned methods and parameters (Section 3.3.), 692 AS events (undergoing in a total of 457 genes) evidenced a difference in Ψ between normal and tumour tissue, the majority of which were SE and AFE events. Naturally, the large number of selected events makes it difficult to biologically interpret the results. A gene enrichment analysis was therefore conducted using DAVID. Using DAVID's Gene Functional Classification, it is possible to cluster the 457 genes associated to the primarily selected AS events into smaller functional related clusters. A total of 35 genes (in which 61 AS events took place) associated with functional clusters that have any functional relation with the oncogenic process (*e.g.* proliferation, angiogenesis, etc.) were selected. To further increase the robustness of the AS event selection, only the 14 AS events (belonging to 10 genes) with the most dramatic changes in Ψ value ($|\Delta\widetilde{\Psi}| > 0.4$) were chosen for biological interpretation. The analysis of the genomic coordinates of the exons involved in the events, as well as the mRNA and protein isoforms produced by them, was carried out resorting to the UCSC genome browser (http://genome.ucsc.edu/) and SMART (http://smart.embl-heidelberg.de/), an online resource for the identification and annotation of protein domains and the analysis of protein domain architectures. The biological interpretation of relevant cancer-specific AS pattern alterations is described in the next sections, and well as summarized in Appendix D.

Using the method described in Section 3.5, we also conducted a survival analysis of the identified cancer-specific AS events. Survival analyses yielding $p<0.05$, making those events putative prognostic factors, are highlighted in the next sections.

### 4.1.1. Fibroblast Growth Factor Receptor 2 (FGFR2)

The *FGFR2* gene is involved in important processes such as cell division, regulation of cell growth and maturation, formation of blood vessels, wound healing, and embryonic development [90].

Our analysis points to an increased inclusion of *FGFR2* exon 9 in tumour tissue. Conversely, the exclusion of exon 8 is more frequent in tumour tissue. This is expected, given that these exons are spliced in a mutually exclusive manner. In addition, the inclusion level of exon 9 is higher than exon 8 for all tumours, except for one Stage I sample. The opposite situation is generally observed in normal tissue.

These exons are key in the synthesis of two of the best documented protein isoforms of this gene. The inclusion of exon 9 gives origin to FGFR2 IIIc protein isoform, whereas the inclusion of exon 8 originates FGFR2 IIIb protein isoform. FGFR2 IIIb and FGFR2 IIIc are both composed by three Ig-like domains, a transmembrane domain and a cytoplasmic tyrosine kinase domain (Figure 14). These protein isoforms are almost identical, except for the latter half of the third Ig-like domain. FGFR2 IIIb is reported to be predominantly expressed in epithelial cells, whereas FGFR2 IIIc is preferentially expressed in mesenchymal cells [91].

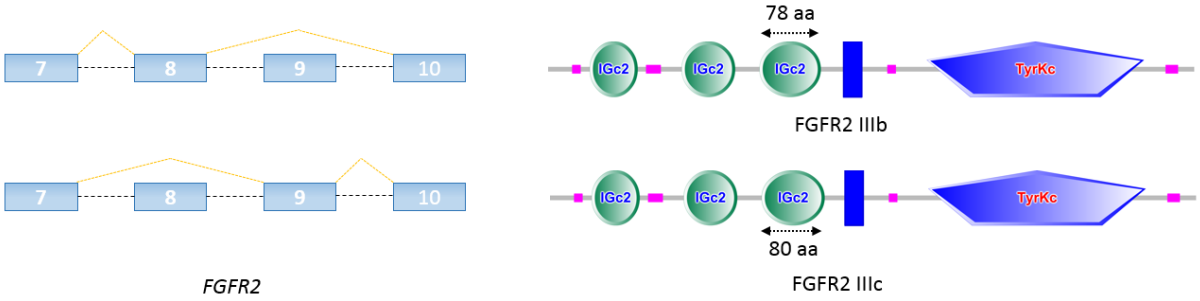



**Figure 14-** AS events in *FGFR2* and protein isoforms originated from those events.

This result is in concordance with recent reports in the literature, with 90% of the ccRCC analysed showing a larger percentage of the FGFR2 IIIc isoform than FGFR2 IIIb isoform, being this AS pattern associated to a worst clinical outcome [92]. This tendency points to an EMT. This is a biological process by which cells lose epithelial characteristics and acquire mesenchymal phenotype. Epithelia are highly ordered monolayers of cells that have apical-basal polarity and adhere tightly to each other via adherens and tight junctions. In contrast, mesenchymal cells differ in shape and display an increased capacity for migration and invasion, thus facilitating tumour metastization (Figure 15) [93].

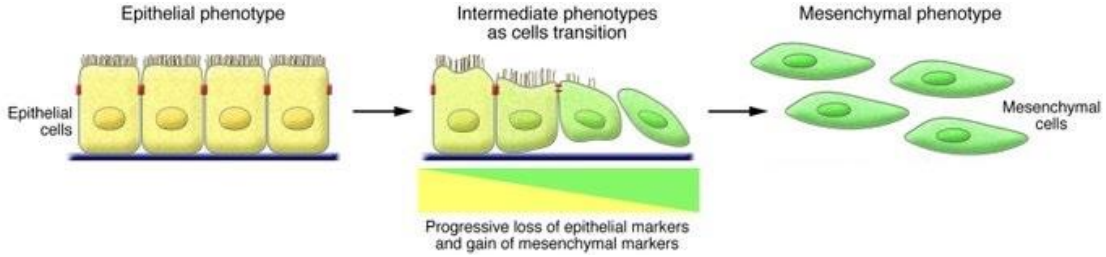



**Figure 15 –** EMT illustration. Epithelial cells are tightly adhered to each other whereas mesenchymal cells are characterized by a migratory capability [94].

In addition, this switch seems to be kidney-specific and it is rarely observed in other cancers. In fact, this tendency is actually opposite to the one reported in some cancers such as prostate cancer, where more advanced tumours may show an increase in the FGFR2 IIIb isoform (which could point to a mesenchymal-epithelial transition associated with the formation of metastases), while less advanced tumours show a decrease in the IIIb isoform and an increase in FGFR2 IIIc isoform [92].

## 4.1.2. Ras-Related C3 Botulinum Toxin Substrate 1 (RAC1)

The protein encoded by *RAC1* is a GTPase belonging to the RAS superfamily of small GTP-binding proteins. These proteins function as molecular switches that cycle between an *ON* state when bound to GTP and an *OFF* state when bound to GDP. The RAC proteins are tightly regulated by various groups of proteins such as Rho-GEFs (Guanine Exchange Factors), which promote binding to GTP, and Rho-GAPs (GTPase activating proteins) that promote the hydrolysis of GTP to GDP by the RAC proteins. The RAC proteins are master regulators of diverse signalling pathways that control the shape, motility and growth of cells [95, 96, 97].

A high decrease in the median Ψ value of isoforms containing exon 3b in tumour tissue was detected in relation to non-tumour tissue. In fact, 95% of the analysed patients showed a decrease in the inclusion level of exon 3b in tumour samples. Interestingly, a mean Ψ value of 0.91 is associated to isoforms containing this exon in non-tumour tissue, thus pointing to clearly high abundance of these isoforms.

Resorting to the UCSC Genome Browser, it was possible to associate the inclusion of exon 3b to the production of the transcript that gives origin to Rac1b protein isoform. This is a splice variant of *RAC1* containing a 19 amino acid insertion next to the switch II region (Figure 16).

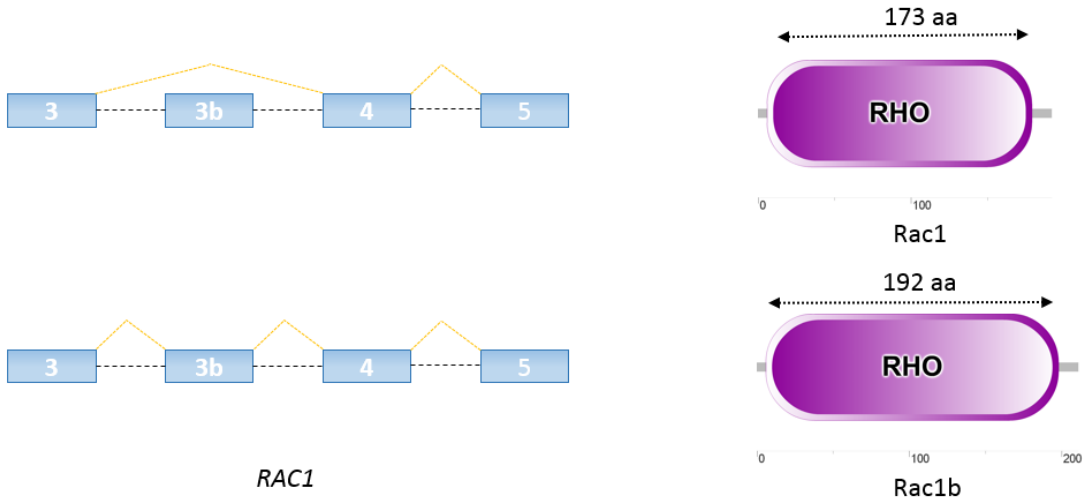



**Figure 16-** AS events in *RAC1* and protein isoforms originated from those events.

Rac1b is characterized by having an accelerated GEF-independent GDP/GTP exchange and an impaired GTP hydrolysis [97]. The Rac1b accelerated and independent fashion in which this protein isoform conducts its activity favours tumour progression, with previous reports pointing to an overexpression of this isoform in breast cancer, colon cancer and lung adenocarcinoma [98, 99, 100]. Being that the expression of the *RAC1* gene remains roughly the same in the cancer/normal switch (according to the GE analysis conducted), one can conclude that the increased proportion of isoforms not containing exon 3b actually leads to an increased number of normal Rac1 isoforms in ccRCC. Although one would expect an increase in the proportion of Rac1b protein isoform in most cancers, due to its aforementioned characteristics, overexpression of normal Rac1 has also been associated with cancer, namely testicular cancer [101]. In addition, both Rac1b and Rac1 have been reported to stimulate NF-κB-mediated (a protein complex extensively linked to tumourigenesis) transcription in colorectal cancer. However, only Rac1 has been shown to induce RelB-mediated gene transcription, which further stimulates NF-κB [102].

### 4.1.3. Spleen Tyrosine Kinase (SYK)

The *SYK* gene encodes a member of the family of non-receptor type Tyr protein kinases that contains two adjacent Src homology 2 (SH2) domains and a kinase domain. This protein is widely expressed in

hematopoietic cells and is involved in coupling activated immunoreceptors to downstream signalling events that mediate diverse cellular responses, including proliferation, differentiation, and phagocytosis [103].

The results of our analysis point towards an increase of the median Ψ value associated to isoforms that include exon 8 in tumour cells. Approximately 98% of the analysed tumours registered this increase. The inclusion of exon 8 gives origin to the *normal* Syk protein isoform, whereas its skipping gives origin to a shorter Syk protein isoform known as SykB (Figure 17). In comparison with the alternatively spliced isoform SykB, Syk contains an insertion of 23 aa within the interdomain B region. Within that sequence of 23 aa there is a nuclear localization signal required for nuclear translocation [104].

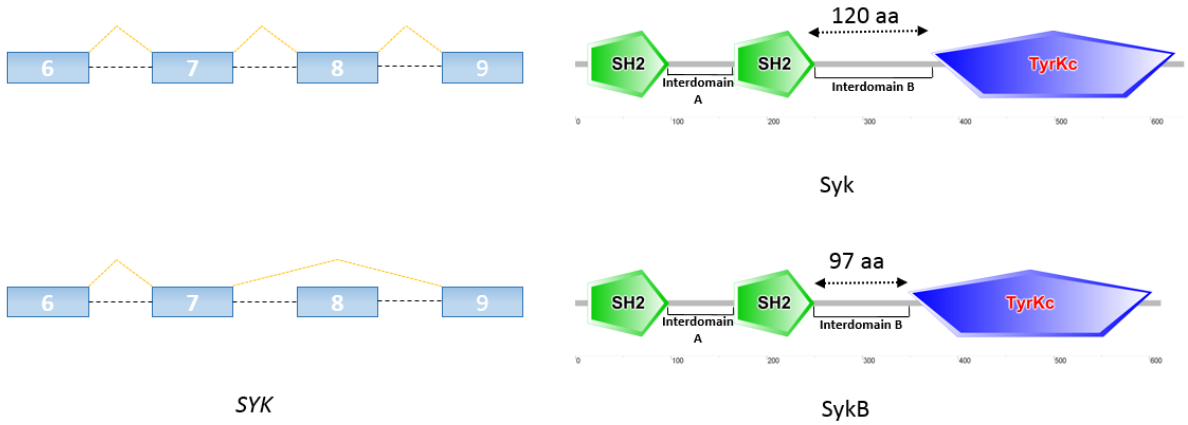



**Figure 17-** AS events in *SYK* and protein isoforms originated from those events.

A previous report indicates that the transfection of the full-length Syk isoform serves as a negative regulator in tumour growth and progression in breast cancer, although the mechanics of this suppression remain unknown [104]. Transfection of SykB, on the other hand, does not. In addition, the study found that the SykB isoform was not expressed in matched normal mammary tissues. This is an inverse scenario to the one observed for ccRCC. To our knowledge, this AS event has not been associated with any other cancer type, suggesting we may be observing a ccRCC-specific AS pattern switch.

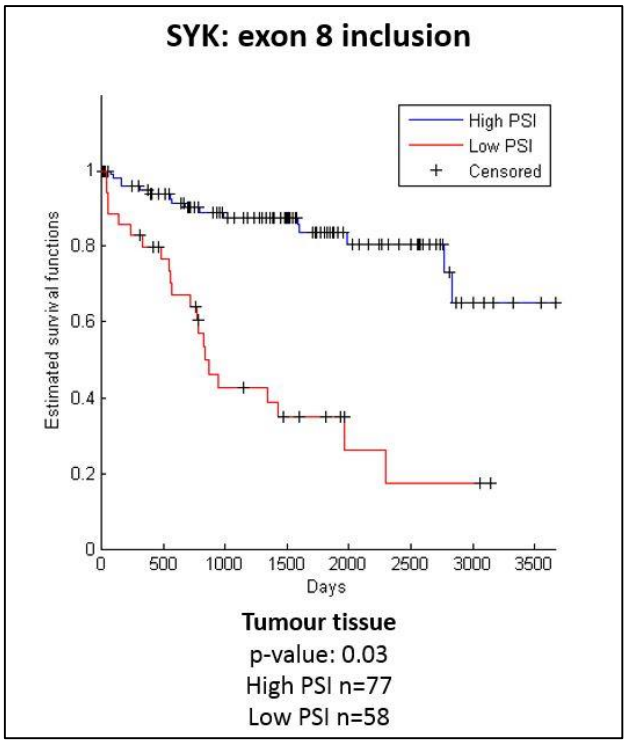



**Figure 18 -** Estimated survival functions for patients with high and low Ψ associated to the inclusion of *SYK*'s exon 8, in tumour tissue.

Interestingly, a higher inclusion (High PSI, Ψ≥0.89) seems to yield a significantly better clinical outcome in tumour tissue (Figure 18). This might be due the characteristics of full-length Syk, known to serve as a negative regulator in tumour growth and progression [105].

## 4.1.4. Kalirin, RhoGEF Kinase (KALRN)

The *KALRN* gene is responsible for promoting the exchange of GDP by GTP. It also activates specific Rho GTPase family members, thereby inducing various signalling mechanisms that regulate neuronal shape, growth, and plasticity, through their effects on the actin cytoskeleton. Diseases associated with *KALRN* include Huntington's disease and coronary heart disease 5 [106].

Our analysis points to an increase of the use of exon 62 as ALE exon instead of exon 51, in tumour tissue when compared to normal tissue. This tendency is present in 98% of the analysed tumours. The usage of exon 62 as ALE gives origin to the Trad protein isoform, whereas the use of exon 51 gives origin to Kalirin isoform.

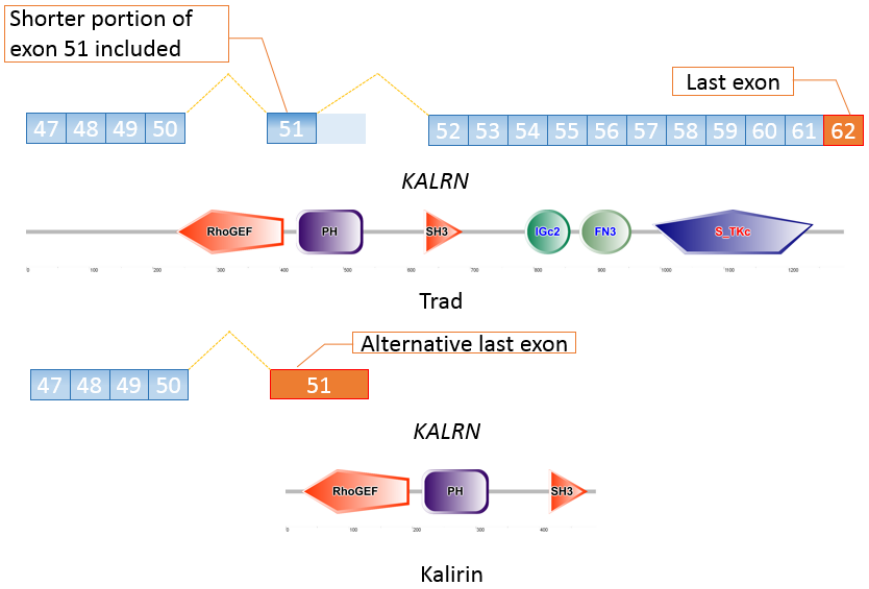



**Figure 19-** AS events in *KALRN* and protein isoforms originated from those events.

As indicated in Figure 19, the Trad isoform is larger than Kalirin and has three additional domains when compared to the latter, including a Serine/Threonine protein kinase catalytic domain (S_TKc). Protein kinases play a role in a multiples important cellular processes, including division, proliferation, apoptosis, and differentiation. Phosphorylation, a process mediated by protein kinases, usually results in a functional change of the target protein by changing its enzyme activity, cellular location, or association with other proteins [107]. The catalytic subunits of protein kinases are highly conserved and have been used to develop kinase-specific inhibitors for the treatments of a number of diseases [108].

Although no previous reports that relate this specific gene or its products to cancer were found, the expression of multiple Serin/Threonine kinases seems to be altered in several cancers [109].

## 4.1.5. MCF.2 Cell Line Derived Transforming Sequence-Like (MCF2L)

*MCF2L* codes for the guanine nucleotide exchange factor. This protein potentially links pathways that signal through *RAC1*, *RHOA* and *CDC42* by catalysing guanine nucleotide exchange on *RHOA* and *CDC42* and interacting specifically with the GTP-bound form of *RAC1*, suggesting that it functions as an effector of *RAC1*. Diseases associated with *MCF2L* include hypoparathyroidism and spasticity. This gene is also involved in 1-phosphatidylinositol binding. An important paralogue of this gene is *KALRN* [110].

In cancer, the median Ψ value associated to isoforms originated through the usage of exon 1 as AFE is decreased in relation to the median Ψ value in normal tissue, where Ψ values are generally superior to 0.8 in the analysed samples. The alternative event is the usage of exon 5 as an AFE, which gives origin to a shorter guanine nucleotide exchange factor that has its N-terminal truncated (Figure 20).

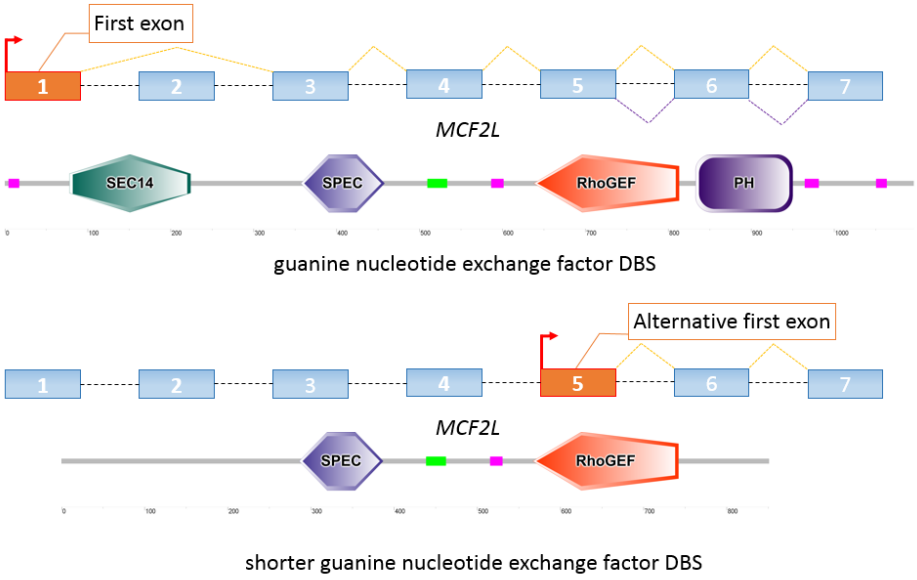



**Figure 20-** AS events in MCF2L and protein isoforms originated from those events.

The truncation of this terminal confers tumourigenic properties to this isoform, which is concordant with our analysis' results and other reports of a higher abundance of this isoform in ccRCC [110].

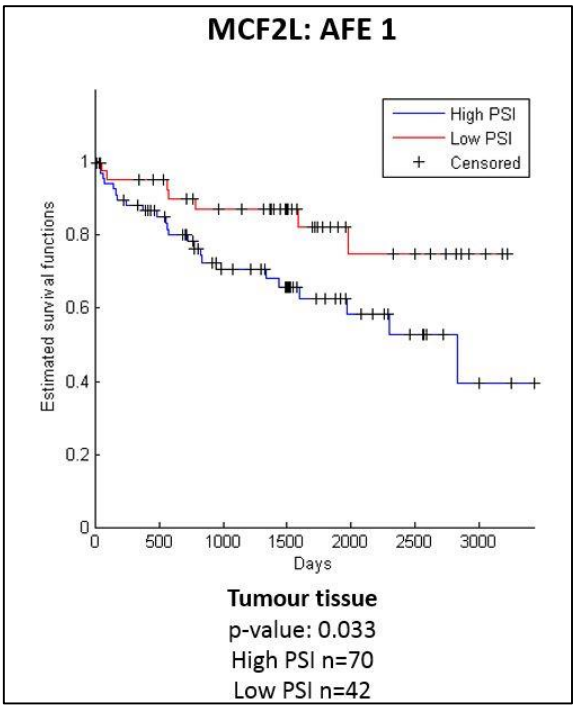



**Figure 21 -** Estimated survival functions for patients with high and low Ψ associated to the use of *MCF2L*'s exon 1 as AFE, in tumour tissue.

Higher Ψ values associated to the use of *MCF2L*'s exon 1 as alternative first exon seem to provide a worst clinical outcome (Figure 21).

As previously referred, *MCF2L* acts as an effector of *RAC1* and it is a paralogue of *KALRN.* Being that a switch in the Ψ values associated to AS events that occur in both *RAC1* and *KALRN* genes was

identified, the correlation between the Ψ values associated to the use of *MCF2L*'s exon 1 as alternative AFE and the Ψ values associated to the indicated events of both genes was analysed. Both Pearson product-moment correlation coefficient and Spearman's rank correlation coefficient were computed. No correlation was found between the Ψ values associated to the use of *MCF2L*'s exon 1 and either AS events.

## 4.1.6. Protein Tyrosine Phosphatase, Non-Receptor Type 6 (PTPN6)

*PTPN6* encodes a protein member of the protein tyrosine phosphatase (PTP) family. PTPs are known to be signalling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation. The N-terminal part of this PTP contains two SH2 domains, which act as protein phospho-tyrosine binding domains, and mediate the interaction of this PTP with its substrates. This PTP is expressed primarily, and functions as an important regulator of multiple signalling pathways, in hematopoietic cells [111]. It is also overexpressed in epithelial ovarian cancer and appears to be associated with breast adenocarcinoma [111, 112]. This gene has two alternative promoters, immediately upstream of exons 1 and 2 (Figure 22). These exons are included in a mutually exclusive manner [112].

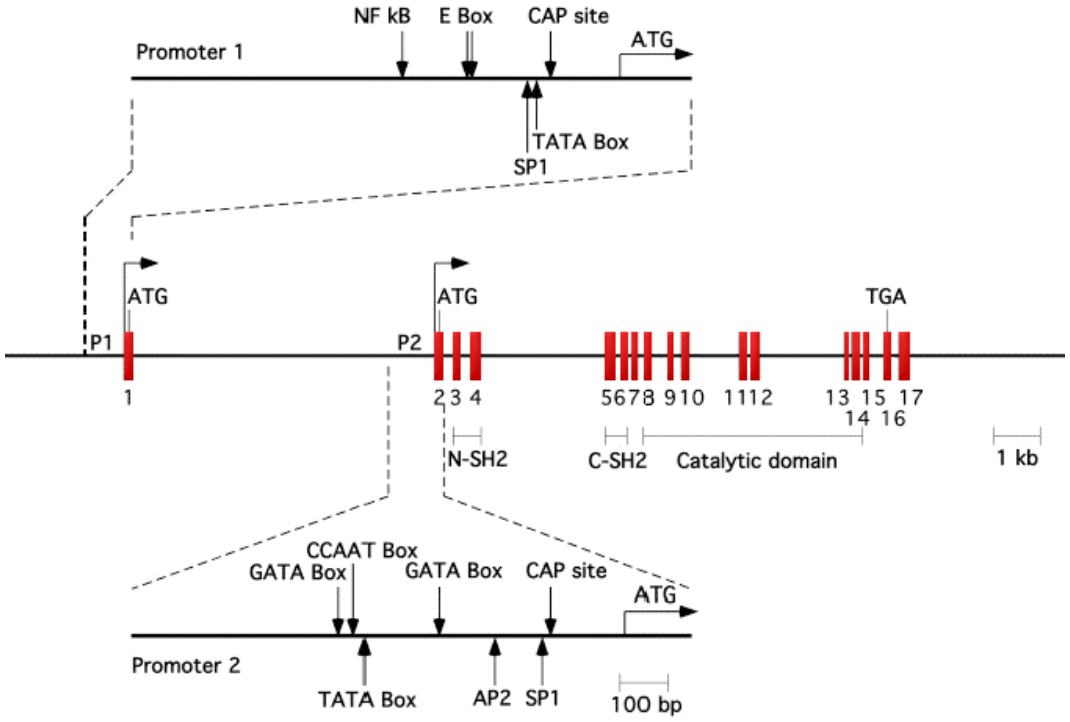



**Figure 22-** The different alterative first exons of PTNP6 gene [113].

Our analysis points to an increase of the usage of exon 2, instead of exon 1, as AFE in tumour cells. The usage of exon 1 as AFE is associated to epithelial cells, while the usage of exon 2 is associated to hematopoietic cells. The functional significance of alternative promoter usage remains to be established.

Also, whether the minor differences found at the N-terminal of these isoforms affect their overall properties is not yet known [114].

## 4.1.7. CD44 Molecule (Indian Blood Group) (CD44)

The *CD44* gene encodes for a cell-surface glycoprotein involved in cell-cell interactions, cell adhesion and migration. This protein participates in a wide variety of cellular functions including lymphocyte activation, recirculation and homing, haematopoiesis, and tumour metastasis [115]. Transcripts for this gene undergo complex AS that results in many functionally distinct isoforms, as shown in Figure 23.

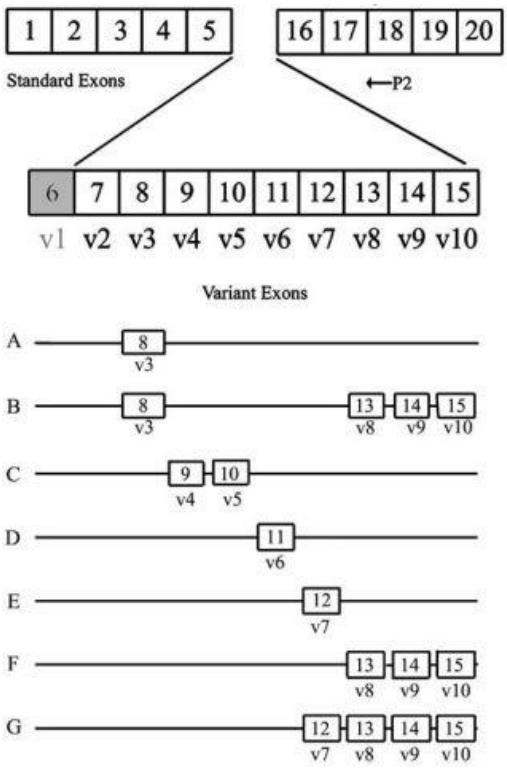



**Figure 23-** Different isoforms of *CD44* gene. Exons v1 through v10 are alternative exons [116].

Our analysis' results point to an increase of the exclusion of exons 12 (v7) and 13 (v8) supported by a decrease in the median Ψ value associated to these exons in tumour tissue. In the literature, it is reported that the CD44E isoform (F, in Figure 23), which is associated to epithelial cells and includes exon v8, is not expressed in ccRCC lower grade tumours. However it is expressed in higher grade tumours [117]. Our results are not in concordance to the ones reported in the literature. A decrease of Ψ value associated to the inclusion of exon v8 in 86.67% of the cases was detected. In fact, of the patients that experienced an increase in the Ψ value associated to this event in tumour tissue in relation to normal tissue, only one had a tumour stage IV tumour (other patients that experienced this increase had stage I or II tumours). This stage IV patient had a significant increase of the Ψ value associated to this event from 0.1 to 0.95. In addition, the average of the Ψ value associated to the inclusion of exon v8 is higher in stage I or II tumours than in stage III or IV tumours (0.19 *vs.* 0.12) as well as the median

(0.11 *vs*. 0.08). Higher levels of exclusion of exon v8 translate into lower production levels of CD44E isoform, which may suggest EMT and the facilitation of metastization.

## 4.1.8. Other cancer-specific AS events

Other AS events met the defined criteria but little information about the genes where these occur and protein isoforms originated by them was found. Therefore, no conclusions about the functional impact of the observed AS shifts were reached.

In tumours, there is an increase in the median Ψ value associated to the retention of intron 6 of Deoxyribonuclease I (*DNASE1)* gene, which seems to be involved in cell death by apoptosis and DNA double strand cleavage. Alternative transcriptional splice variants of this gene have been observed but not thoroughly characterized [118]. Additionally, the inclusion of its exon 3 decreases in tumours. The exclusion of exon 3 of this gene gives origin to a shorter DNASE1 (Figure 24). The production of this protein isoform seems to be favoured in tumour cells. However, we found no information about the functional consequences of this shrinkage.

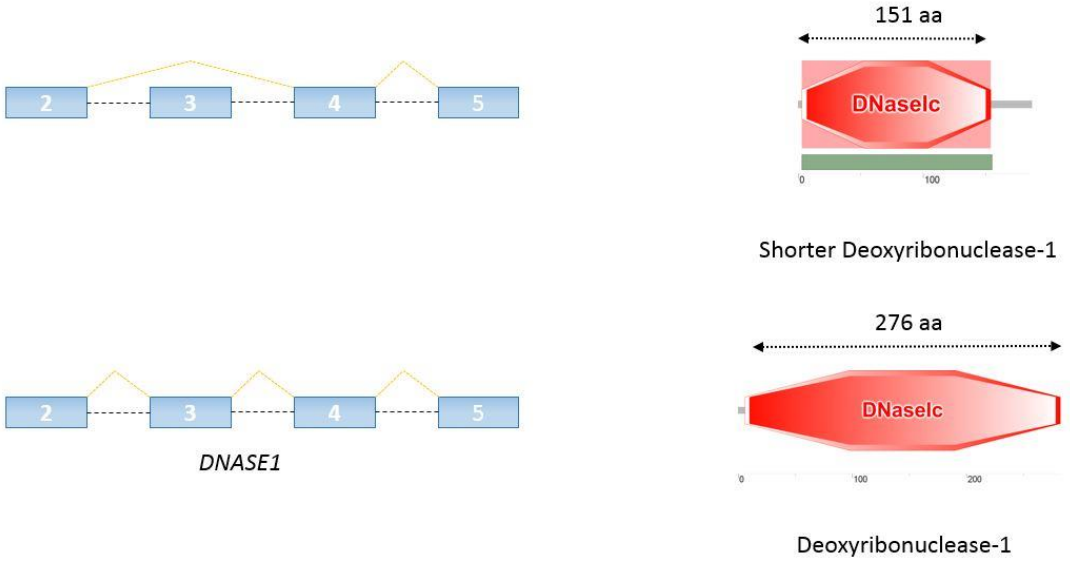



**Figure 24 -** AS events in *DNASE1* and protein isoforms originated from those events.

No information about the protein isoform originated by the retention of intron 6 was found either.

For the G Protein-Coupled Receptor 132 (*GPR132*) gene, known to be involved in apoptosis [119], there is a decrease in inclusion of its exon 4 in tumours. The exclusion of exon 4 of this gene gives origin to a shorter GPR132, the variant 2 protein, whereas its inclusion gives origin to variant 1 (Figure 25).

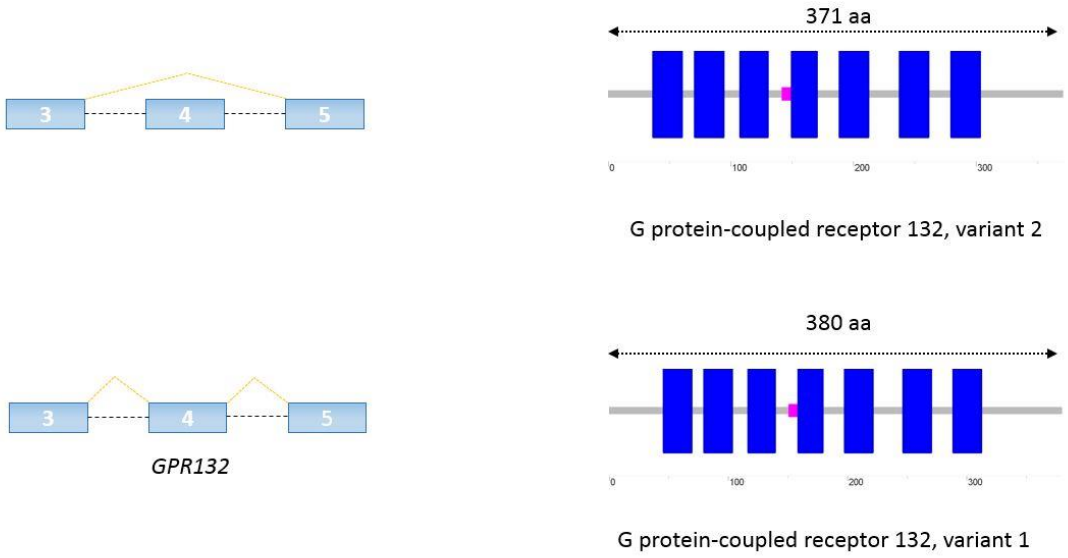



**Figure 25 -** AS events in *GPR132* and protein isoforms originated from those events.

In tumours, a larger abundance of variant 2 isoform has been observed. No information about the functional differences between these protein isoforms could be found in the literature.

Finally, the median Ψ value associated to the inclusion of exon 5 of Tumour Necrosis Factor, Alpha-Induced Protein 8 (TNFAI8) gene decreases in tumour tissue. This gene acts a negative mediator of apoptosis and may play a role in tumour progression [120]. The inclusion of exon 5 gives origin to a TNFAIP8 isoform whereas its exclusion gives origin to a slightly shorter isoform (Figure 26).

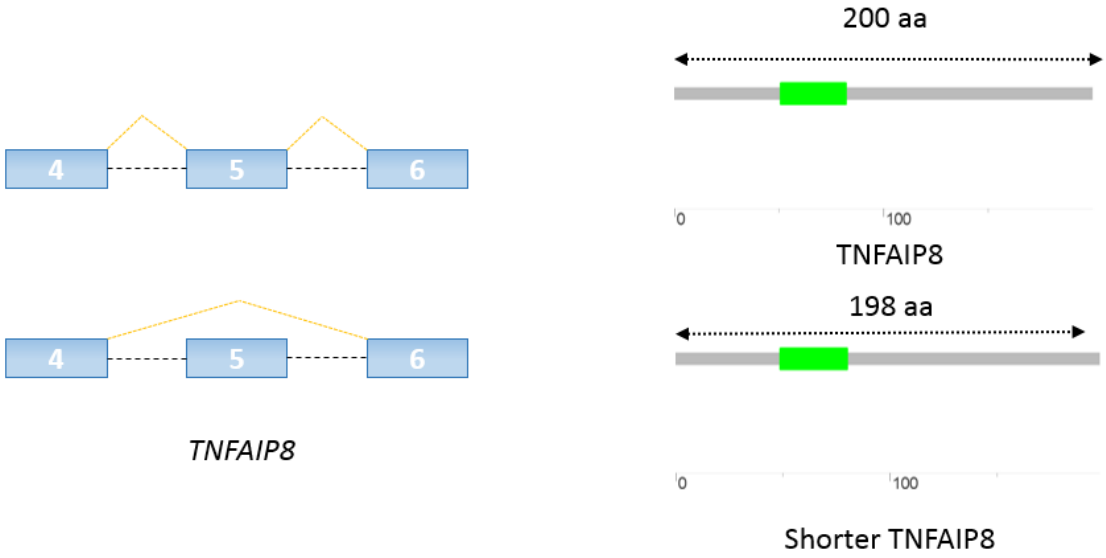



**Figure 26 -** AS events in *TNFAIP8* and protein isoforms originated from those events.

Once again, no information about the functional differences of these isoforms were found in the literature.

## 4.2. Binary tumour stage classifier

The classification system giving better overall results in our analyses was the one where patients with stages I and II cancers were classified as 0 and patients with stages III and IV cancers were classified as 1. Specifically, the lowest deviance D value ($D = 102.96$) was obtained with $\alpha = 0.9$ and $\lambda = 0.07$. D was estimated using 10-fold cross-validation. The obtained regression used 41 AS events as parameters for classification (indicated on Appendix E).

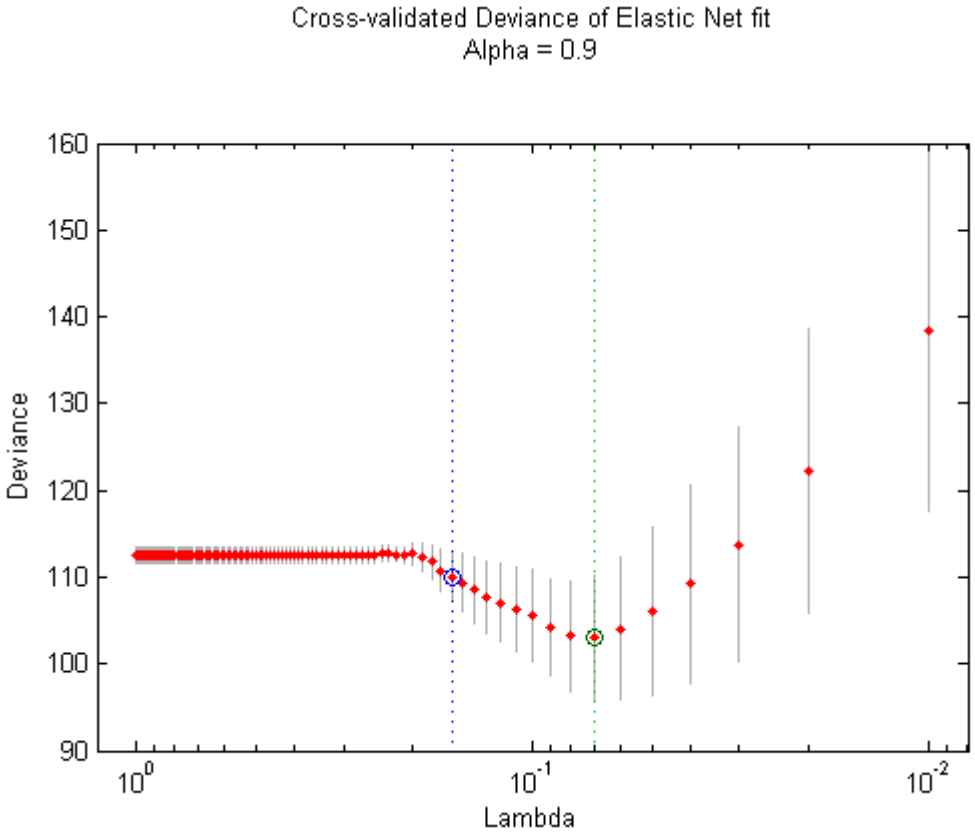



**Figure 27 -**10-fold cross-validation plot for α=0.9. $\boldsymbol{D}$ estimated for each lambda with error bars for each estimate. The traced green line indicates the lambda at which the minimum $\boldsymbol{D}$ is obtained.

Applying ROC an optimum threshold of 0.56 was obtained (with True Positive Rate=1 and False Positive Rate=0) (Figure 28). This threshold provides a 100% accurate separation.

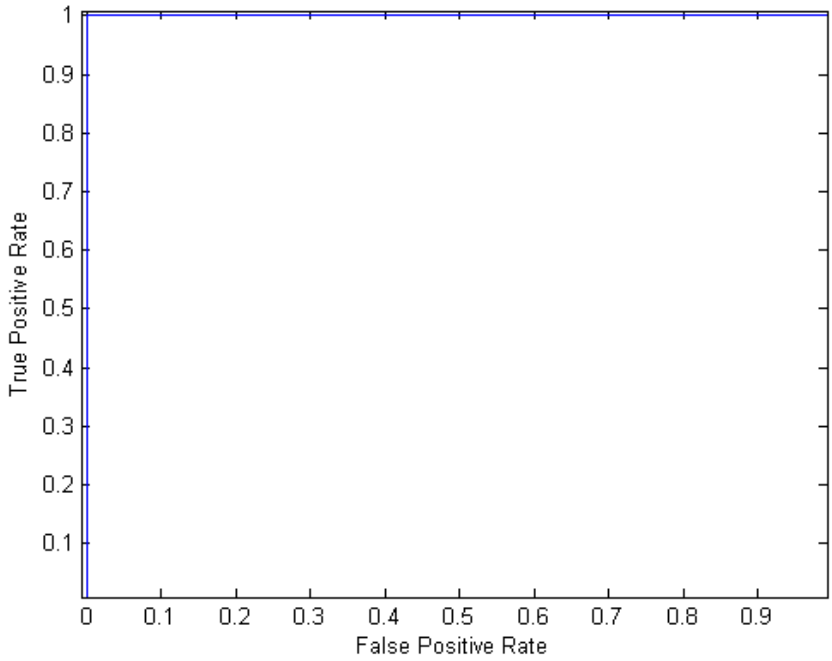



**Figure 28-** ROC curve for the model obtained using α=0.9 and λ=0.07.

The results of testing this classifier with the 58 patients that were not used in the regression are indicated in Table 3.

| | | **Real Stage** | |
|---|---|---|---|
| **Predicted Stage** | | I or II | III or IV |
| | I or II | 38 | 15 |
| | III or IV | 2 | 3 |

**Table 3 –** Estimated stages obtained with the classifier *vs*. real stages.

With the data gathered in Table 3 one can calculate the traditional ratios that are used to access the quality of a classifier: sensitivity and specificity [121]. In this context the sensitivity of the classifier refers to the ability of the classifier to correctly identify patients who have a stage III or IV cancer:

$$Sensitivity = \frac{True\ positives}{True\ positives + False\ negatives} = \frac{3}{3+15} = 16.67\%, \qquad (Eq. 22)$$

where *True positives* is the number of patients that have a stage III or IV cancer and the classifier correctly classified their cancer as 1 and *False negatives* is the number of patients that have a stage III or IV cancer and the classifier misclassified their cancer as 0.

The specificity of this classifier refers to its ability to correctly identify those patients with stage I or II cancer:

$$Specificity = \frac{True\ negatives}{True\ negatives + False\ positives} = \frac{38}{38+2} = 95\%, \qquad (Eq. 23)$$

where *True negatives* is the number of patients that have a stage I or II cancer and the classifier correctly classified their cancer as 0 and *False positives* is the number of patients that have a stage I or II cancer and the classifier misclassified their cancer as 1.

Even though the application of this method provides a significant dimension reduction (from 18291 to 41) the results are not very satisfactory. The specificity of this classifier is very high but the sensitivity is extremely low. Nevertheless we believe that the results unveil some potential associated to the cancer stage classification through the use of MISO Ψ estimates. In an effort to try and optimize the results, a new threshold was calculated taking into account the estimated values and real values of the test subjects. To that end a ROC curve was used (Figure 29).

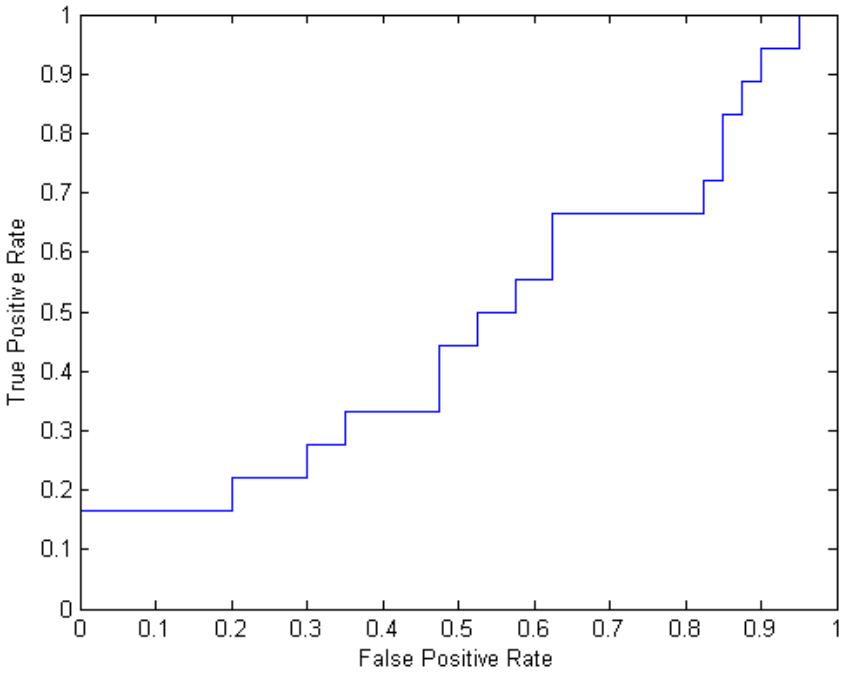



**Figure 29 –** ROC curve taking into account the predicted and real stages of the test subjects.

The threshold which maximized the module $\|True\ Positie\ Rate - False\ Positive\ Rate\|$ was selected as the new threshold value. The new threshold value was 0.62, with results indicated in Table 4.

| | | **Real Stage** | |
|---|---|---|---|
| **Predicted Stage** | | I or II | III or IV |
| | I or II | 40 | 15 |
| | III or IV | 0 | 3 |

**Table 4 –** Estimated stages obtained with the classifier *vs*. real stages using new threshold.

The improvement was not significant. The sensitivity of the classifier using the new threshold remained the same. The only improvement was verified in the specificity of the classifier (Eq. 24):

$$Specificity = \frac{True\ negatives}{True\ negatives + False\ positives} = \frac{40}{40 + 0} = 100\%. \qquad (Eq.\,24)$$

## 4.3. Independent AS prognostic factors

A total of 67 AS events (from 30 genes) in normal tissue and 39 AS events (from 30 genes) in tumour tissue were considered statistically significant by Benjamini–Hochberg multiple test correction and registered a difference equal or larger than 0.3 between. From these AS events, the analysis of the 2 events with the smallest p-value, from each of both normal and tumour sample groups, is described in the following sections.

## 4.3.1. Independent AS prognostic factors in normal tissue

The Ψ value associated to the use of *PXDN* (Peroxidasin Homolog (Drosophila)) exon 62 as ALE seems to be a good prognostic factor in normal tissue (Figure 30). The PXDN protein is an extracellular matrix-associated peroxidase, thought to function in extracellular matrix consolidation, phagocytosis, and defence [122]. This gene seems to play a crucial part in Heme Oxygenase-1 tumour adhesion-promoting effects [123].

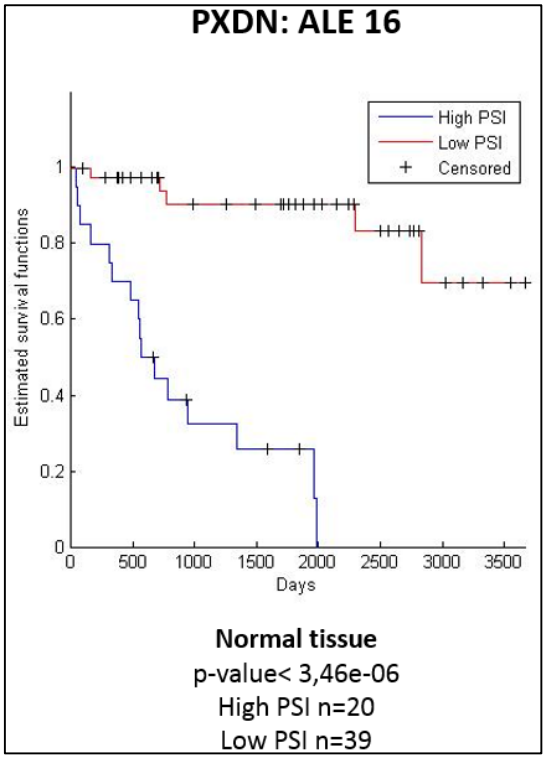



**Figure 30-** Estimated survival functions for patients with high and low Ψ associated to the use of *PXDN*'s exon 16 as ALE, in normal tissue.

In normal tissue, the Ψ values associated to use of exon 16 as ALE seem to yield a worst clinical outcome to the High PSI group (Ψ≥0.8). The Ψ values associated to this AS event in normal tissue range from 0.53 to 0.87, with a median value of 0.77. To our knowledge there are no previous reports relating survival with this AS event or gene.

Also in normal tissue, the inclusion of exon v5 of the *CD44* gene also seems to serve as a prognostic factor (Figure 31). As previously referred in Section 4.1.7., the protein coded by this gene takes part on a wide variety of cellular functions including lymphocyte activation, recirculation and homing, haematopoiesis, and tumour metastasis.

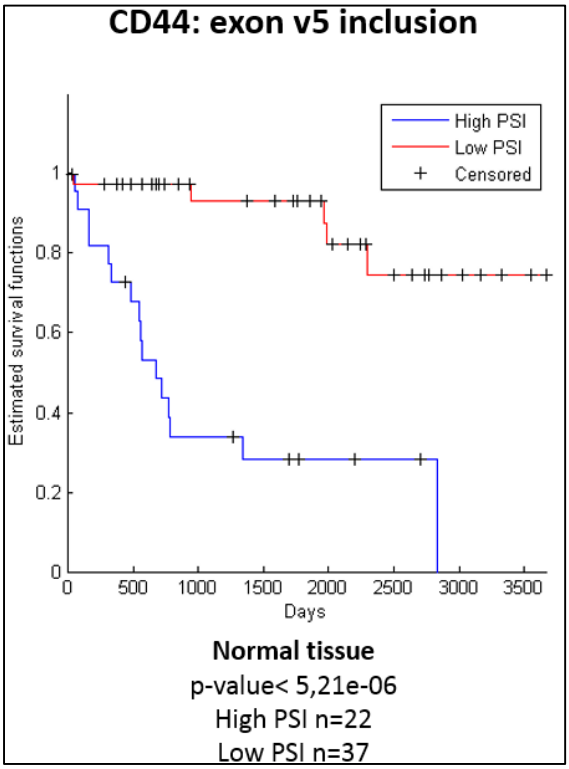



**Figure 31 -** Estimated survival functions for patients with high and low Ψ associated to the inclusion of *CD44*'s exon v5, in normal tissue.

Higher levels of inclusion of exon v5 (High PSI group, Ψ≥0.16) seem to be associated to a significantly worst outcome when compared to lower levels of inclusion of this exon (Low PSI group). The Ψ values associated to this AS event in normal tissue range from 0.03 to 0.68, with a median value of 0.12. In the literature there are various reports relating high inclusion of exon v5 with tumour progression and worst clinical outcome. Increased levels of exon v5 have been associated to more advanced stages of colorectal tumour progression (advanced polyps and invasive carcinomas) [124]. Also, higher inclusion of exon v5-containing CD44 isoforms has been associated to poor overall survival in breast cancer [125]. Finally, reports point to higher levels of exon v5-containing CD44 isoforms as cancer staging progresses in human thymic epithelial neoplasms, relating these isoforms to invasiveness. Interestingly, in the same article the authors found that even though higher levels of these isoforms were related to more aggressive thymic epithelial neoplasms, better survival cancer was associated to higher levels of expression of these isoforms [126].

## 4.3.2. Independent AS prognostic factors in tumour tissue

In tumour tissue, the usage of an A3SS in exon 4 of Baculoviral IAP Repeat Containing 5 gene (*BIRC5*) seems to be a prognostic factor (Figure 32). The protein encoded by this gene, known as survivin, has dual roles in promoting cell proliferation and preventing apoptosis [127]. Survivin expression is turned off during fetal development and not found in non-neoplastic tissues, however it is found in most human cancers [128].

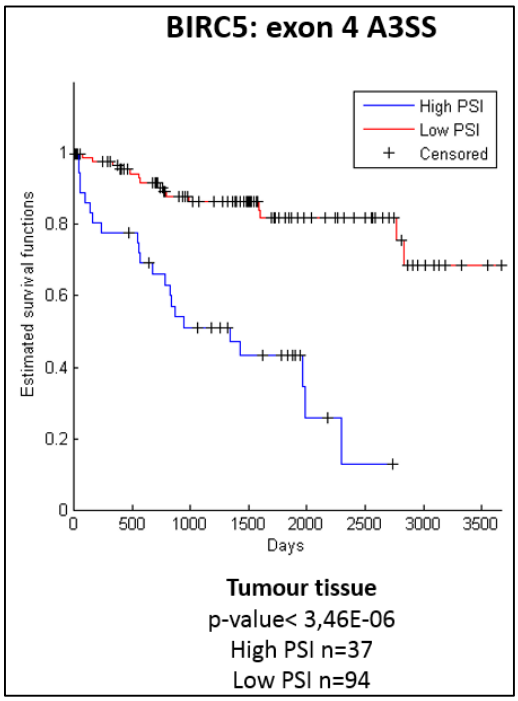



**Figure 32 -** Estimated survival functions for patients with high and low Ψ associated to the use of chromosome 17 coordinates 76210870 and 76212745 as donor and acceptor sites in *BIRC5*'s exon 4, in tumour tissue. The alternative acceptor site is coordinate 76212747.

Higher levels of usage of the constitutive acceptor site (High PSI group, Ψ≥0.96) seem to be associated to a worst clinical outcome when compared to lower levels of its usage (Low PSI group). The Ψ values associated to this AS event in tumour tissue range from 0.47 to 1, with a median value of 0.91. According to UCSC Genome Browser this AS event affects the survivin 3B isoform. Survivin 3B has been reported to promote the escape of malignant cells from immune recognition by blocking the cytotoxicity of natural killer cells. It also inhibits the activation of caspase-6, thus increasing the resistance of neoplastic cells to various chemotherapeutics [129]. The usage of the alternative acceptor originates an mRNA isoform containing a premature stop codon. This premature stop codon will very probably drive the mRNA isoform to degradation through the nonsense-mediated mRNA decay (NMD) pathway (a translation-coupled quality control system that recognizes and degrades aberrant mRNAs with truncated open reading frames due to the presence of a premature termination codon) or simply produce a truncated protein [130]. Thus, higher usage levels of the alternative acceptor site will translate into lower levels of functional protein. This suggests that this acceptor site may be part of a mechanism to prevent the production of the oncogenic protein isoform.

Finally, in tumour tissue the Ψ value associated to the inclusion of exon 3 of Forkhead Box M1 gene (*FOXM1*) seems to be related with survival (Figure 33). This gene encodes for a transcriptional factor that regulates expression of cell cycle genes essential for DNA replication and mitosis. It also plays a role in DNA breaks repair, participating in the DNA damage checkpoint response, and in cell proliferation control [131].

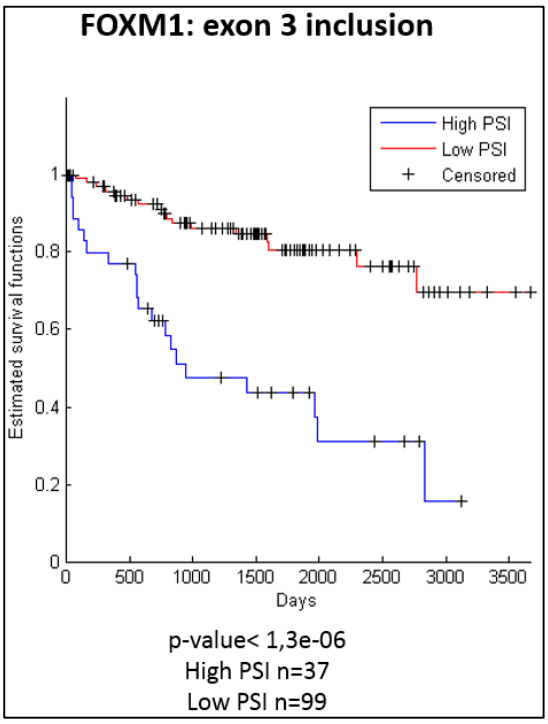



**Figure 33 -** Estimated survival functions for patients with high and low Ψ associated to the inclusion of *FOXM1*'s exon 3, in tumour tissue.

The Ψ values associated to this AS event in tumour tissue range from 0.42 to 0.98, with a median value of 0.96. A worst clinical outcome is associated to higher levels of inclusion of exon 3 (High PSI Ψ≥0.96). In the literature, *FOXM1* is described as only having 2 alternative exons Va and VIIa (Figure 34). Exon 3 is therefore reported to be constitutive [132].

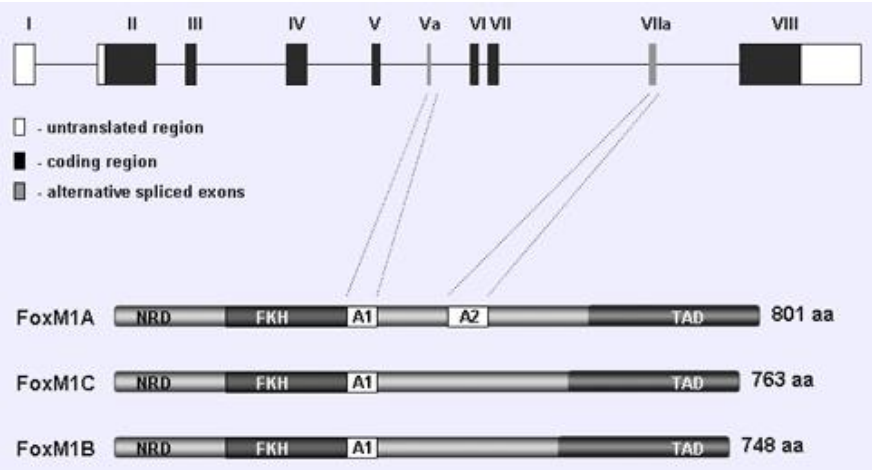



**Figure 34 -** AS and corresponding mRNA isoforms of *FOXM1* [132].

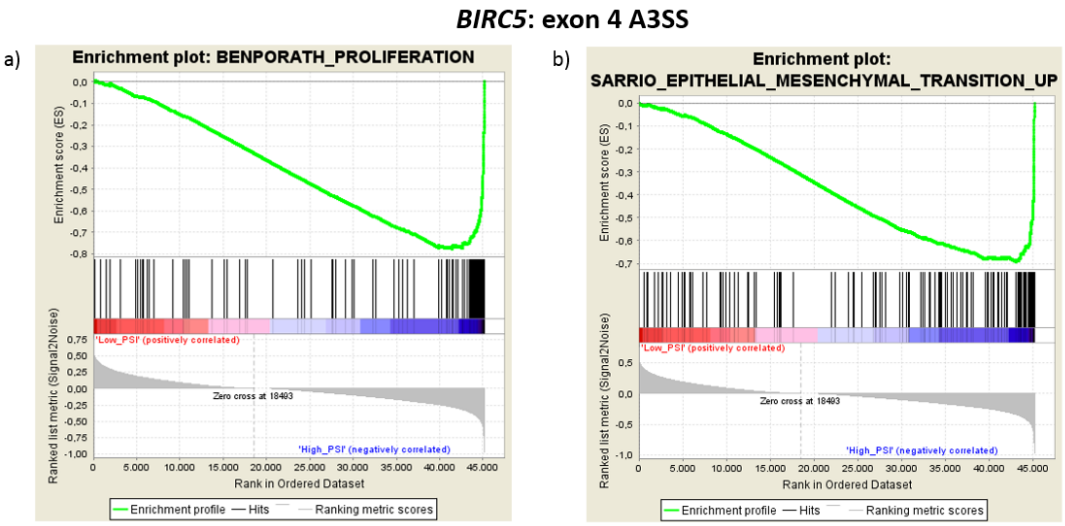
According to Ensembl, the exclusion of exon 3 originates an isoform that is degraded by NMD. This might indicate that lower levels of functional FOXM1 protein may be associated to a better prognostic. A similar scenario is observed in gastric cancer in which overexpression of *FOXM1* has been associated to a worst prognostic [133]. In addition, *FOXM1* overexpression has also been associated to EMT in pancreatic cancer [134].

## 4.4. GSEA

GSEA was conducted for each relevant AS event, comparing the High and Low PSI groups. The patients division between these two phenotypes was strictly the same as the one used for the survival analysis conducted for the respective AS event. The GE values used were either from normal or tumour tissue, depending upon the tissue type for which the AS event evidenced putative prognostic properties. Interpretations of the relevant outputs from these analyses are described in the following sections.

### 4.4.1. *BIRC5*: exon 4 A3SS

GSEA was conducted on tumour tissue GE. We have analysed the 39 patients from the Low PSI group and the 19 belonging to the High PSI group for which GE data was available. As indicated by the survival analysis, a higher Ψ value is associated with a worse prognosis.




**Enrichment High PSI**
**ES:** -0.78
**p-value:** 0.022
**n**=133

**Enrichment High PSI**
**ES:** -0.69
**p-value:** 0.023
**n**=167

**Figure 35 -** Gene enrichment analysis of High and Low PSI phenotypes, associated to A3SS event in *BIRC5* in tumour tissue. a) Upregulation of BENPORATH_PROLIFERATION gene set; b) Upregulation of SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP gene set in High PSI group.

In the High PSI phenotype, an upregulation of genes belonging to the BENPORATH_PROLIFERATION gene set was identified (Figure 35.a). BENPORATH_PROLIFERATION is a set of genes defined in human breast tumour expression data that are associated with embryonic stem cell identity in the expression profiles of various human tumour types [135]. Cancer cells possess traits reminiscent of those ascribed to normal stem cells. These cells are characterized by high proliferation potential. Patients with a higher Ψ value associated to this AS event evidence a GE signature that favours cell proliferation when compared to patients with lower Ψ associated to the same event. In addition, there seems to be an upregulation of genes belonging to the SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP gene set in the High PSI phenotype (Figure 35.b). This set corresponds to genes whose overexpression correlate with EMT in breast cancer. EMT is highly associated with tumour metastases. This might indicate that the tumour of the patients of the High PSI group may have a bigger predisposition to metastasize than those in the Low PSI group [136]. These results are in concordance to the lower survival rate associated to High PSI group patients.

## 4.4.1. *FOXM1*: exon 3 inclusion

Again gene enrichment analysis was conducted on tumour tissue GE. We have analysed the 43 patients from the Low PSI group and the 18 belonging to the High PSI group for which GE data was available. As indicated by the survival analysis, a higher Ψ value is associated with a worse prognosis.

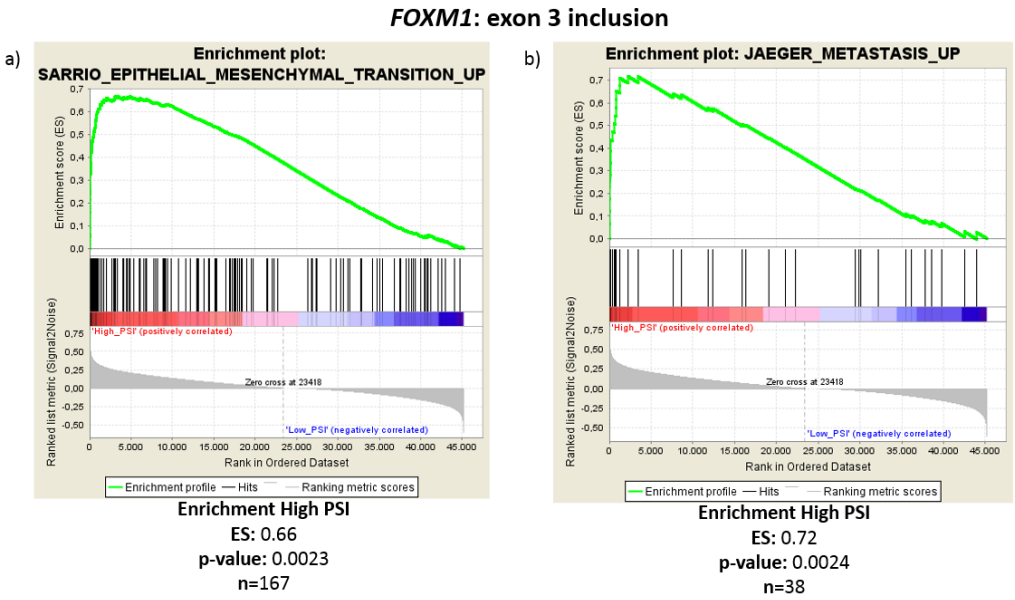



**Figure 36 -** Gene enrichment analysis of High and Low PSI phenotypes, associated to the inclusion of *FOXM1's* exon 3 in tumour tissue. a) Upregulation of BENPORATH_PROLIFERATION gene set; b) Upregulation of SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP gene set in High PSI group.

Once again, an upregulation of the genes of SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP gene set was found in patients who

registered a higher Ψ value associated to the inclusion of exon 3 of *FOXM1* gene (Figure 36.a). A similar conclusion to the one presented in the previous section can be drawn.

An upregulation of the genes of the JAEGER_METASTASIS_UP gene set was also found in the High PSI phenotype associated to this AS event (Figure 36.b). This set is defined by up-regulated genes in metastases from malignant melanoma compared to the primary tumours [137]. This GE pattern might indicate that the patients from the High PSI group might have a bigger incidence of metastases. In fact, this association is significant, with 10 of the 18 patients (55.6%) that made up the High PSI group having metastic ccRCC, whereas metastases were only detected in 7 of the 43 patients (16%) that made up the Low PSI (p-value of 0.0038 for the corresponding Fisher's exact test). Once again, those results were expected since a lower survival rate associated to High PSI group patients.

# 5

# Conclusion

## 5.1. Discussion

In this thesis, we discuss the analyses of AS, GE and survival data aiming to identify cancer-specific AS patterns as well as AS events that serve as prognostic factors in ccRCC. In addition, we describe the application of dimension reduction and regression methods in order to develop a cancer stage classifier based on AS patterns.

Our analyses identified a large number of cancer-specific AS events, thus suggesting that, similarly to GE, AS patterns primarily separate normal from tumour samples. Specifically, the identification of a normal/tumour *switch* pattern in the inclusion levels of *FGFR2*'s exons 8 and 9 serves as a proof-of-principle to our approach, since these events are among the few reported in the literature. Interestingly, some identified cancer-specific AS events easily-interpretable possible biological implications. This is the case for the decreased expression, in tumour tissue, of isoforms originated through the usage of *MCF2L*'s exon 1 as first exon. In this case, the cancer-specific AFE is exon 5, whose usage gives origin to a shorter and highly tumourigenic guanine nucleotide exchange factor that that has its N-terminal truncated. In addition a great number of cancer-specific AS events suggest EMT.

The developed classification methodology was not effective in the use of AS event to predict cancer stage. The high number of parameters considered (18291 AS events) and the limited number of observations available (138 patients) suggest that the used model may be overfitting the data. Overffiting occurs whenever a statistical model describes a random error or noise instead of the underlying relationship between the parameters and the studied outcome [138].

The conducted survival analysis did return a considerable number of statistically significant AS events. These results suggest that there is great potential in the use of AS patterns as independent prognostic factors.

Finally, gene enrichment analysis of survival data gives biological sustenance to these potential clinical tools. Specifically, the upregulation of gene sets related to high proliferative potential, EMT and metastasis is reassuringly observed in patients with poorer survival expectancy.

These results suggest a great potential of AS signatures derived from tumour transcriptomes in providing etiological leads for cancer progression and as a clinical tool. A deeper understanding of the contribution of splicing alterations to oncogenesis could lead to improved cancer prognosis and contribute to the development of RNA-based anticancer therapeutics, namely splicing-modulating small molecule compounds.

# References

**[1]-** General cancer classification, staging, and grouping, retrieved on June 24, 2014, from http://stedmansonline.com/webFiles/Dict-Stedmans28/APP21.pdf

**[2]-** IARC, Latest world cancer statistics Global cancer burden rises to 14.1 million new cases in 2012: Marked increase in breast cancers must be addressed, 2013

**[3]-** Tran B., et al., Cancer Genomics: Technology, Discovery, and Translation, *Journal of Clinical Oncology 2012;* 30:647-660

**[4]-** Somatic mutation, staging, and grouping, retrieved on June 24, 2014, from http://ghr.nlm.nih.gov/glossary=somaticmutation

**[5]-** Epigenetic, retrieved on June 24, 2014, from http://ghr.nlm.nih.gov/glossary=epigenetic

**[6]-** Hanahan D, Weinberg R, Hallmarks of cancer: the next generation, *Cell* 2011; 144: 646-674

**[7]**- Oltean S, Bates DO, Hallmarks of alternative splicing in cancer, *Oncogene* 2013; 1-8

**[8]-** Feng H, *et al.*, Opportunities and methods for studying alternative splicing in cancer with RNA-Seq, *Cancer Letters* 2012

**[9]-** Alberts B., *et al., Essential Cell Biology,* Garland Science, 2013, pp.223-257

**[10]-** Transcription: Advanced Look, retrieved on February 16, 2014, from http://vcell.ndsu.edu/animations/transcription/advanced.htm

**[11]-** mRNA Splicing: Advanced Look, retrieved on February 16, 2014, from http://vcell.ndsu.edu/animations/mrnasplicing/advanced.htm

**[12]-** mRNA Processing: Advanced Look, retrieved on February 16, 2014, from http://vcell.ndsu.edu/animations/mrnaprocessing/advanced.htm

**[13]-** Press Release: The Nobel Prize in Physiology or Medicine 1993, retrieved on February 16, 2014, from http://www.nobelprize.org/nobel_prizes/medicine/laureates/1993/press.html

**[14]-** Mueller W., Hertel K., *Alternative pre-mRNA Splicing: Theory and Protocols,* Wiley-Blackwell, 2012, pp. 21-28

**[15]-** Eukaryotic mRNA Processing, retrieved on February 16, 2014, from http://oregonstate.edu/dept/biochem/hhmi/hhmiclasses/bb451/lectnotesgdp/EukaryoticmRNA.html

**[16]-** Early P., et al., Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways, *Cell* 1980; 20: 313–319

**[17]-** Qun P., et al., Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing, *Nature* 2008; 40: 1413–1415

**[18]-** Nielsen T., Brenton G., Expansion of the eukaryotic proteome by alternative splicing, *Nature* 2010; 463: 457–463

**[19]-** Matlin A., *et al*., Understanding alternative splicing: towards a cellular code, *Nature Reviews* 2005; 6: 386-398

**[20]-** Keren H., *et al*., Alternative splicing and evolution: diversification, exon definition and function, *Nature Reviews* 2010; 11: 345-355

**[21]-** Different types of alternative splicing, retrieved on June 1, 2014, http://www.nature.com/nrg/journal/v11/n5/box/nrg2776_BX1.html

**[22]-** Liu Y, *et al.,* Quantification of alternative splicing variants of human telomerase reverse transcriptase and correlations with telomerase activity in lung cancer, *PLoS One* 2012; 7: e38868

**[23]-** Reeve JG, *et al.,* Expression of apoptosis-regulatory genes in lung tumour cell lines: relationship to p53 expression and relevance to acquired drug resistance, *Br J Cancer* 1996; 73: 1193-1200

**[24]-** Krajewska M, *et al*., Immunohistochemical analysis of bcl-2, bax, bcl-X, and mcl-1 expression in prostate cancers, *Am J Pathol* 1996; 148: 1567-1576

**[25]-** Krajewska M, *et al.,* Elevated expression of Bcl-X and reduced Bak in primary colorectal adenocarcinomas. Cancer Res 1996; 56: 2422-2427

**[27]-** Shultz JC, *et al.,* Alternative splicing of caspase 9 is modulated by the phosphoinositide 3-kinase/Akt pathway via phosphorylation of SRp30a, *Cancer Res* 2010; 70: 9185-9196

**[28]-** Shkreta L, *et al.*, Anticancer drugs affect the alternative splicing of Bcl-x and other human apoptotic genes*, Mol Cancer Ther* 2008; 7: 1398-1409

**[29]-** Valastyan S, Weinberg R, Tumor metastasis: molecular insights and evolving Paradigms, *Cell* 2011; 147: 275-292

**[30]-** Warzecha CC, *et al*., An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. *EMBO J* 2010; 29: 3286-3300.

**[31]-** Konrad L, *et al*., Alternative splicing of TGF-betas and their high-affinity receptors T beta RI, T beta RII and T beta RIII (betaglycan) reveal new variants in human prostatic cells, *BMC Genomics* 2007; 8: 318

**[32]-** Zhang K, *et al*., Patterns of missplicing caused by RB1 gene mutations in patients with retinoblastoma and association with phenotypic expression, *Hum Mutat* 2008; 29: 475-484

**[33]-** Lohmann DR, RB1 gene mutations in retinoblastoma, *Hum Mutat* 1999; 14: 283-288

**[34]-** Flaherty KT, *et al*., Inhibition of mutated, activated BRAF in metastatic melanoma, *N Engl J Med* 2010; 363: 809-819

**[35]-** Diaz R, *et al.*, p73 isoforms affect VEGF, VEGF165b and PEDF expression in human colorectal tumours, VEGF165b downregulation as a marker for poor prognosis, *Int J Cancer* 2008; 123: 1060-1067

**[36]-** The Cancer Genome Atlas Research Network, Comprehensive molecular characterization of clear cell renal cell carcinoma, *Nature* 2013; 499: 43-49

**[37]-** Kidney cancer, retrieved on February 18, 2014, http://www.wcrf.org/cancer_statistics/data_specific_cancers/kidney_cancer_statistics.php

**[38]-** Renal Cell Carcinoma (RCC): Fact sheet, retrieved on February 18, 2014, http://www.pfizer.com/files/news/esmo/renal_cell_carcinoma_fact_sheet.pdf

**[39]-** Wang Z., *et al.*, RNA-Seq: a revolutionary tool for transcriptomics, *Nat Rev Genet.* 2009; 10(1): 57–63

**[40]-** Trapnell C., *et al.*, TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics* 2009; 25: 1105-1111

**[41]-** Bowtie: an ultrafast memory-efficient short read aligner, retrieved on February 3, 2014, from http://bowtie-bio.sourceforge.net/index.shtml

**[42]-** Maq: Mapping and Assembly with Qualities, retrieved on February 3, 2014, from http://maq.sourceforge.net/

**[43]-** SEQAN, retrieved on February 3, 2014, from http://www.seqan.de/

**[44]-** Phred - Quality Base Calling, retrieved on February 17, 2014, from http://www.phrap.com/phred/

**[45]-** Bowtie Manual, retrieved on February 17, 2014, http://www.animalgenome.org/bioinfo/resources/manuals/rna_bowtie.txt

**[46]-** Wang E.T., *et al.*, Alternative Isoform Regulation in Human Tissue Transcriptomes, *Nature* 2008; 456(7221): 470–476

**[47]-** Cufflinks: Transcript assembly, differential expression, and differential regulation for RNA-Seq, retrieved on February 4, 2014, from http://cufflinks.cbcb.umd.edu

**[48]-** Trapnell C., *et al.*, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nat Protoc.* 2012; 7(3): 562–578

**[49]-** Trapnell C., *et al.*, Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation, *Nature Biotechnology.* 2010; 28: 511–515

**[50]-** MISO: Probabilistic analysis and design of RNA-Seq experiments for identifying isoform regulation, retrieved on February 5, 2014, from http://genes.mit.edu/burgelab/miso/index.html

**[51]-** Yardenl K., *et al.*, Analysis and design of RNA sequencing experiments for identifying isoform regulation, *Nat Methods*, 2010; 7(12): 1009–1015

**[52]-** MISO documentation, retrieved on March 28, 2014, from http://genes.mit.edu/burgelab/miso/docs/#summary-format

**[53]-** Kakaradov B., *et al.*, Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data, *BMC Bioinformatics*, 2012; 13(Suppl 6): S11

**[54]-** Hypothesis testing, retrieved on March 29, 2014, http://www.math.uah.edu/stat/hypothesis/Introduction.html

**[55]-** What is Hypothesis Testing?, retrieved on March 29, 2014, http://stattrek.com/hypothesis-test/hypothesis-testing.aspx

**[56]-** Dorey Frederick, P-value, *Clinical Orthopaedics and Related Research,* 2010; 468: 2297–2298

**[57]-** Bonferroni Correction, retrieved on March 29, 2014, from http://mathworld.wolfram.com/BonferroniCorrection.html

**[58]-** A tutorial on False Discovery Rate, retrieved on March 17, 2014, fromControl http://www.stat.cmu.edu/~genovese/talks/hannover1-04.pdf

**[59]-** Zabell S., On Student's 1908 Article "The Probable Error of a Mean", *Journal of the American Statistical Association*, 2008; 103

**[61]-** 'Student' and Small-Sample Theory by E. L. Lehmann, retrieved on March 29, 2014, from http://statistics.berkeley.edu/sites/default/files/tech-reports/541.pdf

**[61]-** Guinness, t-Tests & Proving a Pint Really Does Taste Better in Ireland, retrieved on March 29, 2014, from http://blog.minitab.com/blog/michelle-paret/guinness-t-tests-and-proving-a-pint-really-does-taste-better-in-ireland

**[62]-** Student, The Probable Error of a Mean, *Biometrika,* 1908; 6: 1-25

**[63]-** Hypothesis testing, retrieved on March 29, 2014, http://www.stats.gla.ac.uk/steps/glossary/hypothesis_testing.html#twoside

**[64]-** Paired Difference t-test**,** retrieved on March 29, 2014, from http://www.cliffsnotes.com/math/statistics/univariate-inferential-tests/paired-difference-t-test

**[65]-** Paired Difference t-test**,** retrieved on March 29, 2014, from http://math.tutorvista.com/statistics/paired-t-test.html

**[66]-** The Wilcoxon Signed-Rank Test, retrieved on May 17, 2014, from http://vassarstats.net/textbook/ch12a.html

**[67]-** Wilcoxon Signed Ranks Test: Nonparametric Analysis for Two Related Populations, retrieved on May 17, 2014, from http://wps.prenhall.com/wps/media/objects/11886/12171343/OnlineTopics/bbs12e_onlinetopic_ch12-8.pdf

**[68]-** A Level Maths Notes: Wilcoxon Signed Rank Tables, retrieved on May 17, 2014, from http://astarmathsandphysics.com/a-level-maths-notes/a-level-maths-notes-wilcoxon-signed-rank-tables.html

**[69]-** Kolmogorov-Smirnov test, retrieved on May 15, 2014, from http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Kolmogorov-Smirnov_test.html

**[70]-** The Kolmogorov Goodness-of-Fit Test (Kolmogorov-Smirnov one-sample test), retrieved on May 15, 2014, from http://www.math.nsysu.edu.tw/~lomn/homepage/class/92/kstest/kolmogorov.pdf

**[71]-** Tibshirani R., Regression Shrinkage and selection via Lasso, *Journal of the Royal Statistical Society 1996,* 58:267-288

**[72]-** ON THE STABILITY OF INVERSE PROBLEMS, retrieved on June 14, 2014, from http://a-server.math.nsc.ru/IPP/BASE_WORK/tihon_en.html

**[73]-** Lasso and elastic net, retrieved on June 14, 2014, from http://www.mathworks.com/help/stats/lasso-and-elastic-net.html

**[74]-** Zou H., Hastie T., Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society 2005,* 67:301-320

**[75]-** Cross Validation, retrieved one June 14, 2014, http://www.cs.cmu.edu/~schneide/tut5/node42.html

**[76]-** Receiver operating characteristic, retrieved on June 30, 2014, from http://www.mathworks.com/help/nnet/ref/roc.html

**[74]-** Zou H., Hastie T., Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society 2005,* 67:301-320

**[75]-** Cross Validation, retrieved one June 14, 2014, http://www.cs.cmu.edu/~schneide/tut5/node42.html

**[76]-** Receiver operating characteristic, retrieved on June 30, 2014, from http://www.mathworks.com/help/nnet/ref/roc.html

**[77]-** Pearson Product-Moment Correlation, retrieved on August 2, 2014, from https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

**[78]-** Spearman's Rank-Order Correlation, retrieved on August 2, 2014, from https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

**[79]-** What Is Survival Analysis?, retrieved on March 29, 2014, http://www.mathworks.com/help/stats/survival-analysis.html#btnxirj-1

**[80]-** Walter Stephen, What is a Cox Model?, 2nd edition, Hayward Medical Communications, 2009

**[81]-** Cox (Proportional Hazards) Regression, retrieved on March 29, 2014, http://www.statsdirect.com/help/default.htm#survival_analysis/cox_regression.htm

**[82]-** Huang W, *et al.,* Systematic and Integrative Analysis of Large Gene Lists Using DAVID, *Nature Protocols* 2008; 4: 44 - 57

**[83]-** Supplementary Data 4. Summaries of gene identifier types and annotation categories supported in the DAVID system, *Nature Protocols* 2008

**[84]-** Table 2. Major statistical methods and associated parameters used in DAVID, retrieved on May 17, 2014, from http://www.nature.com/nprot/journal/v4/n1/fig_tab/nprot.2008.211_T2.html

**[85]-** EASE Score, a modified Fisher Exact P-Value, retrieved on May 17, 2014, from http://david.abcc.ncifcrf.gov/helps/functional_annotation.html#fisher

**[86]-** Functional Annotation Tool, retrieved on May 17, 2014, from http://david.abcc.ncifcrf.gov/helps/functional_annotation.html

**[87]-** Gene Set Enrichment Analysis, retrieved on May 17, 2014, from http://www.broadinstitute.org/gsea/index.jsp

**[88]-** MATLAB The Language of Technical Computing, retrieved on June 1, 2014, from http://www.mathworks.com/products/matlab/

**[89]-** Cufflinks: Frequently Asked Questions, retrieved on June 1, 2014, from http://cufflinks.cbcb.umd.edu/faq.html#fpkm

**[90]-** FGFR2 gene, retrieved on June 13, 2014, from http://ghr.nlm.nih.gov/gene/FGFR2

**[91]-** Oncology Genes: FGFR2 gene, retrieved on June 13, 2014, http://atlasgeneticsoncology.org/Genes/FGFR2ID40570ch10q26.html

**[92]-** Zhau Qui, *et al.,* Tumor-Specific Isoform Switch of the Fibroblast Growth Factor Receptor 2 Underlies the Mesenchymal and Malignant Phenotypes of Clear Cell Renal Cell Carcinomas, *Clinical Cancer Research 2013*; 10:2460-2472

**[93]-** Epithelial to Mesenchymal Transition, retrieved on September 23, 2014, from http://www.rndsystems.com/molecule_group.aspx?g=3568&r=7&utm_source=poster&utm_medium=goURL&utm_term=EMT&utm_campaign=Epithelial%2Bto%2BMesenchymal%2BTransition

**[94]-** The basics of epithelial-mesenchymal transition, retrieved on September 23, 2014, from http://www.jci.org/articles/view/39104//1

**[95]-** RAC1 gene, retrieved on June 13, 2014, http://www.genecards.org/cgi-bin/carddisp.pl?gene=RAC1

**[96]-** Chun Z.., *et al.,* The Rac1 splice form Rac1b promotes K-ras-induced lung tumorigenesis, *Oncogene* 2013; 32: 903-909

**[97]-** Dennis F., *et al.,* Alternative Splicing of Rac1 Generates Rac1b, a Self-activating GTPase, *Journal of Biological Chemestry 2004;* 279: 4743-4749

**[98]-** Matos P, Peter J., Increased Rac1b Expression Sustains Colorectal Tumor Cell Survival, *Molecular Cancer Research 2008*; 6:1178-1184

**[99]-** Derek C., *et al.,* Rac1b and reactive oxygen species mediate MMP-3-induced EMT and genomic instability, *Nature 2005*, 436:123-127

**[100]-** Alternative splicing of tumour-related rac1b is regulated by upstream signalling pathways, retrieved on June 13, 2014, http://repositorio.insa.pt/bitstream/10400.18/1024/1/posterRac1b_Sinal%202012.pdf

**[101]-** Kamai T., *et al.,* Overexpression of RhoA, Rac1, and Cdc42 GTPases Is Associated with Progression in Testicular Cancer, *Clinical Cancer Research 2004,* 10:4799- 4805

**[102]-** Matos P., Jordan P.*,* Rac1, but Not Rac1B, Stimulates RelB-mediated Gene Transcription in Colorectal Cancer Cells, *The Journal of Biological Chemistry 2006,* 281:13724- 13732

**[103]-** Wang L., *et al*.*,* Alternative Splicing Disrupts a Nuclear Localization Signal in Spleen Tyrosine Kinase That Is Required for Invasion Suppression in Breast Cancer, *Cancer Research 2003,* 63:4724-4730

**[104]-** SYK gene, retrieved on June 13, 2014, http://www.genecards.org/cgi-bin/carddisp.pl?gene=SYK

**[105]-** Coopman C., *et al.,* The Syk tyrosine kinase: A new negative regulator in tumor growth and progression, *Cancer Letters 2006,* 241:159- 173

**[106]-** KALRN gene, retrieved on June 13, 2014, http://www.genecards.org/cgi-bin/carddisp.pl?gene=KALRN

**[107]-** SMART: S_tkc domain annotation, retrieved on June 13, 2014, http://smart.embl.de/smart/do_annotation.pl?DOMAIN=s_tkc

**[108]-** Hanks S., *et al,* Protein Kinase Family: Conserved Features and Deduced Phylogeny of the Catalytic Domains, *Science 2013;* 241

**[109]-** Capra M., *et al.,* Frequent Alterations in the Expression of Serine/Threonine Kinases in Human Cancers, *Cancer Research 2006*; 66:8147-8157

**[110]-** MCF2L gene, retrieved on June 13, 2014, http://www.genecards.org/cgi-bin/carddisp.pl?gene=MCF2L

**[111]-** PTPN6 gene, retrieved on June 14, 2014, http://www.genecards.org/cgi-bin/carddisp.pl?gene=PTPN6

**[112]-** Mok A., *et al.*, Overexpression of the Protein Tyrosine Phosphatase, Nonreceptor Type 6 (PTPN6), in Human Epithelial Ovarian Cancer, Gynecologic Oncology1995; 57:209-303

**[113]-** Oncology Genes: PTPN6 gene, retrieved on June 14, 2014, http://atlasgeneticsoncology.org/Genes/GC_PTPN6.html

**[114]-** Banville D., *et al.,* Human protein tyrosine phosphatase 1C (PTPN6) gene structure: alternate promoter usage and exon skipping generate multiple transcripts, Genomic 1995; 27:165-173

**[115]-** CD44 gene, retrieved on June 14, 2014, http://www.genecards.org/cgi-bin/carddisp.pl?gene=CD44

**[116]-** Omara-Opyene A., *et al.,* Prostate cancer invasion is influenced more by expression of a CD44 isoform including variant 9 than by Muc18, *Laboratory Investigation 2004*, 84:894-907

**[117]-** Terpe H., *et al.,* Expression of CD44 Isoforms in Renal Cell Tumors, *American journal of Pathology 1996, Vol. 148;* 69:3501-3509

**[118]-** SMART: S_tkc domain annotation, retrieved on June 14, 2014, http://smart.embl-heidelberg.de/smart/do_annotation.pl?DOMAIN=CH

**[119]-** GPR132 gene, retrieved on June 15, 2014,http://www.genecards.org/cgi-bin/carddisp.pl?gene=GPR132

**[120]-** TNFAIP8 gene, retrieved on June 15, 2014, http://www.genecards.org/cgi-bin/carddisp.pl?gene=TNFAIP8

**[121]-** Clinical tests: sensitivity and specificity, retrieved on June 28, 2014, from http://ceaccp.oxfordjournals.org/content/8/6/221.full

**[122]-** PXDN gene, retrieved on June 25, 2014, from http://www.genecards.org/cgi-bin/carddisp.pl?gene=PXDN

**[123]-** Tauber Stefanie., *et al.,* Transcriptome analysis of human cancer reveals a functional role of Heme Oxygenase-1 in tumor cell adhesion, *Molecular Cancer* 2010; 9:200

**[124]-** Wielenga V., *et.* al, Expression of CD44 Variant Proteins in Human Colorectal Cancer Is Related to Tumor Progression, Cancer Research 1993; 53:4754-4756.

**[125]-** Tempfer C., *et. al,* Prognostic Value of Immunohistochemically Detected CD44 Isoforms CD44v5, CD44v6 and CD44v7-8 in Human Breast Cancer, *European Journal of Cancer 1996*, 32A: 2023-2025

**[126]-** Lee S., *et. al,* Prognostic Significance of CD44v5 Expression in Human Thymic Epithelial Neoplasms, *The Society of Thoracic Surgeons 2003*, 76:213–218

**[127]-** Baculoviral IAP Repeat Containing 5, retrieved on June 29, 2014, from http://www.genecards.org/cgi-bin/carddisp.pl?gene=BIRC5,

**[128]-** Mahotka C., *et. al*, Survivin- D Ex3 and Survivin-2B: Two Novel Splice Variants of the Apoptosis Inhibitor Survivin with Different Antiapoptotic Properties 1, *Cancer Research 1999*; 59:6097-6102

**[129]-** Végran F., Boidot R., Survivin-3B promotes chemoresistance and immune escape by inhibiting caspase-8 and -6 in cancer cells, *OncoImmunology 2013*, Vol. 2

**[130]-** Nonsense-mediated mRNA decay — Mechanisms of substrate mRNA recognition and degradation in mammalian cells, retrieved on June 29, 2014, from http://www.sciencedirect.com/science/article/pii/S1874939913000278

**[131]-** FOXM1 (forkhead box M1), retrieved on June 29, 2014, from http://atlasgeneticsoncology.org/Genes/GC_FOXM1.html

**[132]-** Supplementary material Raf/MEK/MAPK signaling stimulates the nuclear translocation and transactivating activity of FOXM1c retrieved on June 31, 2014, from http://jcs.biologists.org/content/118/4/795/suppl/DC1

**[133]-** Li Q., *et al.,* Critical Role and Regulation of Transcription Factor FoxM1 in Human Gastric Cancer Angiogenesis and Progression, *Cancer Research 2009*; 69:3501-3509

**[134]-** Bao B., Over-Expression of FoxM1 Leads to Epithelial–Mesenchymal Transition and Cancer Stem Cell Phenotype in Pancreatic Cancer Cells, *Journal of Cell Biochemistry 2011,* 112:2296–2306

**[135]-** BENPORATH_PROLIFERATION, retrieved on June 31, 2014, from http://www.broadinstitute.org/gsea/msigdb/cards/BENPORATH_PROLIFERATION.html

**[136]-** SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP, retrieved on June 31, 2014, from http://www.broadinstitute.org/gsea/msigdb/geneset_page.jsp?geneSetName=SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP

**[137]-** Gene Set: JAEGER_METASTASIS_UP, retrieved on June 31, 2014, from http://www.broadinstitute.org/gsea/msigdb/geneset_page.jsp?geneSetName=JAEGER_METASTASIS_UP

**[138]-** Overfitting, retrieved on November 31, 2014, from http://www.ma.utexas.edu/users/mks/statmistakes/ovefitting.html

[139]-

# A

# Tools brief description

## A.1. Tools brief description

| Tool | Description |
|---|---|
| **RNA-Seq** | RNA-seq is a recently developed technology for transcriptome profiling that uses NGS to sequence DNA molecules reversely transcribed from RNAs. By using RNA-seq one can not only measure gene expression levels with an unmatched precision but also discover and quantify previously unknown transcripts and splicing isoforms [8, 39]. |
| **TopHat** | TopHat is a fast splice junction mapper for RNA-Seq reads. This software does not rely on known splice junctions [40]. This particularity makes TopHat able to identify previously unknown splice variants of genes. |
| **Cufflinks** | This software and its many packages are able to assemble transcripts, estimate their abundances and test for differential expression and regulation in RNA-Seq samples [47]. |
| **MISO** | This software provides a probabilistic framework that quantitates the expression level of alternatively spliced genes from RNA-Seq data and identifies differentially regulated isoforms or exons across samples [50]. |

**Table 5 –** Brief description of the tools used to analyse the available data: RNA-seq (experimental method), TopHat, Cufflinks and MISO (software).

# B

# Patient statistics

## B.1. Patient statistics

| | | # | % |
|---|---|---|---|
| **Gender** | Male | 97 | 70.29% |
| | Female | 41 | 29.71% |
| **Ethnicity** | White | 130 | 94.21% |
| | Black or African American | 5 | 3.62% |
| | No info available | 3 | 2.17% |
| **Vital Status** | Alive | 99 | 71.74% |
| | Deceased | 37 | 28.26% |
| **Follow-up days** | <1000 | 60 | 43.48% |
| | 1000-2000 | 48 | 34.78% |
| | 2000-3000 | 23 | 16.67% |
| | ≥3000 | 6 | 4.35% |
| | unknown | 1 | 0.72% |
| **Tumour Stage** | Stage I | 67 | 48.55% |
| | Stage II | 14 | 10.14% |
| | Stage III | 29 | 21.01% |
| | Stage IV | 28 | 20.29% |
| **Age (years)** | 38-44 | 13 | 9.42% |
| | 45-54 | 27 | 19.57% |
| | 55-64 | 47 | 34.06% |
| | 65-75 | 36 | 26.09% |
| | Over 75 | 14 | 10.14% |
| **Metastases** | Present | 26 | 18.84% |
| | Absent | 112 | 81.16% |
| **Total number of patients** | | 138 | |

**Table 6 -** Patient related information. This info is available on the TCGA Data Portal.

# C

# Event Statistics

## C.1. Event statistics

| Alternative spliced patterns | Tissue status | # of observations |
|---|---|---|
| UTR | tumour | 348333 |
| | normal | 151267 |
| Total # of UTR observations | 499600 | |
| Total # of UTR events | 2642 | |
| Paired UTR observations # (%) | 295832 (59.21%) | |
| RI | tumor | 723135 |
| | normal | 318227 |
| Total # of RI observations | 1041362 | |
| Total # of RI events | 5847 | |
| Paired RI observations # (%) | 598778 (57.5%) | |
| MXE | tumor | 339792 |
| | normal | 149916 |
| Total # of MXE observations | 489708 | |
| Total # of MXE events | 2687 | |
| Paired MXE observations # (%) | 285494 (58.3%) | |
| A3SS | tumor | 1780843 |
| | normal | 781500 |
| Total # of A3SS observations | 2562343 | |
| Total # of A3SS events | 14463 | |
| Paired A3SS observations # (%) | 1495516 (58.37%) | |
| A5SS | tumor | 1603749 |
| | normal | 702510 |
| Total # of A5SS observations | 2306259 | |
| Total # of A5SS events | 12591 | |
| Paired A5SS observations # (%) | 1364956 (59.18%) | |
| ALE | tumor | 1282601 |
| | normal | 565421 |
| Total # of ALE observations | 1848022 | |
| Total # of ALE events | 10131 | |
| Paired ALE observations # (%) | 1075250 (58.18%) | |
| AFE | tumor | 2394617 |
| | normal | 1082272 |
| Total # of AFE observations | 3476889 | |
| Total # of AFE events | 19488 | |
| Paired AFE observations # (%) | 1955578 (56.25%) | |
| SE | tumor | 4749102 |
| | normal | 2100021 |
| Total # of SE observations | 6849123 | |
| Total # of SE events | 38357 | |
| Paired SE observations # (%) | 3962428 (57.85%) | |
| Total | tumor | 13222172 |
| | normal | 5851134 |
| Total # of observations | 19073306 | |
| Total # of events | 106206 | |
| Paired observations # (%) | 11038832 (57.88%) | |

**Table 7 -** Event related information. Divided by AS mechanism (# of events analysed, # of observations).

# D

# Identification of cancer-specific AS patterns results summary

# D.1. Identification of cancer-specific AS patterns results summary

| Gene | Results (normal *vs.* tumour) | Observations and Biological interpretation |
|---|---|---|
| **FGFR2** | • Lower levels of inclusion of exon 8 in tumour cells<br>• Higher levels of inclusion of exon 9 in tumour cells | • EMT<br>• ccRCC-specific AS pattern switch |
| **RAC1** | • Lower levels of inclusion of exon 3b in tumour cells | • Favours production of normal RAC1 protein isoform which:<br>○ has been associated to testicular cancer<br>○ has been reported to stimulate NF-κB-mediated (associated to tumourigenesis) |
| **SYK** | • Higher levels of inclusion of exon 8 in tumour cells | • Favours normal Syk protein. This scenario has not, to our knowledge, been observed in other cancers. This may be a ccRCC-specific AS pattern switch |
| **KALRN** | • Higher usage levels of exon 62 as ALE | • Favours Trad isoform production which contains an additional Serine/Threonine protein kinase catalytic when compared to Kalirin isoform. The expression of multiple Serin/Threonine kinases seems to be altered in several cancers. |
| **MCF2L** | • Lower usage levels of exon 1 as AFE | • Favour production of a truncated guanine nucleotide exchange factor DBS which is highly tumourigenic |
| **PTPN6** | • Higher usage levels of exon 2 as AFE in alternative to exon 1 | • The usage of exon 1 as AFE is associated to epithelial cells, while the usage of exon 2 is associated to hematopoietic cells. |
| **CD44** | • Lower levels of inclusion of exon v7 in tumour cells<br>• Lower levels of inclusion of exon v8 in tumour cells | • Lower levels of inclusion of exon v8 indicate possible EMT |

**Table 8 –** Results summary, with observations and biological interpertations associated to each result.

# E

# Classifier Events

## E.1. Classifier Events

| MISO AS event ID | AS pattern | Gene | chromossome | Start | end |
|---|---|---|---|---|---|
| chr19:44572549:44572641:-@chr19:44572026\|44572241:44571960:- | A3SS | AL159977.1 | chr19 | 39880120 | 39880801 |
| chr2:24153232:24153343:+@chr2:24157250\|24157253:24157329:+ | A3SS | RP11-665N17.4 | chr2 | 24299728 | 24303825 |
| chr3:49134381:49134522:-@chr3:49134291\|49134297:49134121:- | A3SS | AC095064.1 | chr3 | 49159117 | 49159518 |
| chr3:9379554:9380016:+@chr3:9381700\|9381705:9382004:+ | A3SS | RNU6-123P | chr3 | 9404554 | 9407004 |
| chr10:3145311:3145384\|3145380:+@chr10:3145564:3145710:+ | A5SS | Z83851.4 | chr10 | 3155311 | 3155710 |
| chr22:28409009:28409068\|28409053:+@chr22:28420741:28424593:+ | A5SS | RNU6-139P | chr22 | 30079009 | 30094593 |
| chr2:223229:223101\|223097:-@chr2:219966:221191:- | A5SS | PCAT7 | chr2 | 229966 | 233229 |
| chr3:132202754:132202878\|132202848:+@chr3:132217686:132218253:+ | A5SS | PHACTR3 | chr3 | 130720064 | 130735563 |
| chr9:70869674:70869779\|70869771:+@chr9:70877348:70879628:+ | A5SS | AP001595.1 | chr9 | 71679854 | 71689808 |
| 116841@uc001hrg.1@uc001hqz.1 | ALE | TMEM39A | chr1 | 227946696 | 227958412 |
| 1374@uc009ysj.1@uc001oof.2 | ALE | ZNF154 | chr11 | 68522088 | 68522943 |
| 148398@uc001abv.1@uc001abw.1uc001abx.1 | ALE | CTA-407F11.8 | chr1 | 871152 | 879961 |
| 29922@uc001gfv.1@uc001gft.1uc001gfu.1 | ALE | AQP7P4 | chr1 | 169101769 | 169256646 |
| 4277@uc010kpu.1uc003ryc.2@uc003ntn.2uc003nto.2 | ALE | AC055876.2 | chr6 | 2780212 | 31478901 |
| 57669@uc010flk.1uc002tmh.2@uc002tmg.1uc010fll.1uc010flm.1 | ALE | AP000859.4 | chr2 | 120861636 | 120936695 |
| 728340@uc003jww.1@uc010iyk.1uc003kav.2uc003kau.2uc003kaw.2 | ALE | ZNF770 | chr5 | 68874539 | 70331623 |
| 7982@uc003vix.1@uc003vin.1 | ALE | AC096643.1 | chr7 | 116838374 | 116863961 |
| chr10:88920577:88920711:+@chr10:88925626:88925832:+@chr10:88929812:88930040:+ | SE | RP11-266E16.1 | chr10 | 88930597 | 88940060 |
| chr12:107485095:107485173:+@chr12:107485563:107485648:+@chr12:107486736:107487289:+ | SE | TOMM22P3 | chr12 | 108960966 | 108963160 |

| | | | | | |
|---|---|---|---|---|---|
| chr12:51975499:51975690:+@chr12:51977901:51977975:+@chr12:51978096:51978201:+ | SE | RN7SL443P | chr12 | 53689232 | 53691934 |
| chr13:47705276:47705614:+@chr13:47711363:47711469:+@chr13:47725945:47726248:+ | SE | AC005197.2 | chr13 | 48807275 | 48828247 |
| chr15:53349655:53349883:-@chr15:53317893:53317981:-@chr15:53314272:53314446:- | SE | RP11-375O18.2 | chr15 | 55526980 | 55562591 |
| chr19:487951:488147:+@chr19:492184:492252:+@chr19:492339:493091:+ | SE | FAM60BP | chr19 | 536951 | 542091 |
| chr19:8181567:8181746:+@chr19:8221999:8222133:+@chr19:8225383:8225500:+ | SE | RP11-367F23.2 | chr19 | 8275567 | 8319500 |
| chr1:154256428:154256582:-@chr1:154254547:154254783:-@chr1:154247996:154248302:- | SE | MIR33A | chr1 | 155981372 | 155989958 |
| chr1:206006981:206007163:+@chr1:206007747:206007791:+@chr1:206010289:206010330:+ | SE | RP1-249H1.3 | chr1 | 207940358 | 207943707 |
| chr1:206006981:206007163:+@chr1:206010289:206010334:+@chr1:206023260:206023983:+ | SE | RP1-249H1.3 | chr1 | 207940358 | 207957360 |
| chr1:6807739:6807857:+@chr1:6854404:6854487:+@chr1:6870301:6870848:+ | SE | FAM83H-AS1 | chr1 | 6885152 | 6948261 |
| chr2:230972781:230973201:+@chr2:230973293:230973372:+@chr2:230973914:230974022:+ | SE | FBXL18 | chr2 | 231264537 | 231265778 |
| chr3:52532483:52532632:+@chr3:52532708:52532812:+@chr3:52532927:52533073:+ | SE | CAP2P1 | chr3 | 52557443 | 52558033 |
| chr6:106870669:106870834:-@chr6:106862932:106863059:-@chr6:106847596:106847674:- | SE | CTD-3216D2.4 | chr6 | 106740903 | 106764141 |
| chrX:100759149:100759230:-@chrX:100758938:100759016:-@chrX:100756764:100758412:- | SE | RP11-137J7.2 | chrX | 100870108 | 100872574 |

| | | | | | |
|---|---|---|---|---|---|
| **chr11:27473014:27476770:-@chr11:27472697:27473013:-** | UTR | Y_RNA | chr11 | 27516121 | 27520194 |
| **chr17:17909282:17909362:+@chr17:17909363:17912444:+** | UTR | MIR4752 | chr17 | 17968557 | 17971719 |
| **chr17:42188356:42188532:+@chr17:42188533:42189997:+** | UTR | RP11-220I1.4 | chr17 | 44833189 | 44834830 |
| **chr19:54686884:54686910:+@chr19:54686911:54687377:+** | UTR | RP11-38G5.4 | chr19 | 49995072 | 49995565 |
| **chr8:1716140:1716325:+@chr8:1716326:1718315:+** | UTR | ANKRD30B | chr8 | 1728733 | 1730908 |
| **chr1:159442821:159443011:+@chr1:159445607:159445683:+@chr1:159445897:159446009:+@chr1:159446271:159446782:+** | MXE | BX088702.2 | chr1 | 161176197 | 161180158 |
| **chr11:118393878:118393973:-@chr11:118393282:118393353:-** | A3SS | RP11-69L16.6 | chr11 | 118888072 | 118888763 |
| **chr17:4803778:4803934:-@chr17:4803244:4803661:-** | A3SS | AF207550.1 | chr17 | 4862521 | 4863211 |
| **chr3:151768379:151768406:+@chr3:151768494:151768523:+** | A3SS | ZNF519 | chr3 | 150285689 | 150285833 |

**Table 9 -** Classifier events and respective genomic coordinates (according to MISO hg19 genome annotation).