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

Pedro Nuno Brazão Faria
Under supervision of Prof. Susana Vinga Martins and Dr. Nuno Luís Barbosa Morais
Dep. Bioengineering, IST, Lisbon, Portugal
October, 2014

Abstract

In this work, we have analysed gene expression (GE), alternative splicing (AS) and associated patient survival using RNA-seq data from 138 clear cell renal cell carcinomas (ccRCC) 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. 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.

1. Introduction

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]. In recent years, the extensive analysis conducted at the 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 gene activity without alteration of genetic structure), and other genetic aberrations can drive human malignancies [2,3,4]. Specifically, gene expression (GE) alterations at the transcriptional level are being increasingly associated to oncogenesis and tumour progression. Quantitative studies of transcriptomes are therefore deemed as one of the next major tools in the understanding of cancer biology [5].

The recent development of next-generation sequencing (NGS) technologies largely improved our means to study transcriptomes. By using RNA-seq (the use of NGS to sequence 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 [5]. Alternative Splicing is a regulated process during GE that results in a single gene coding for multiple proteins, through various combination of of exons, to produce multiple mRNAs. The different AS mechanisms and respective acronyms are indicated in Appendix A.

2. Methods

2.1. AS and GE quantification

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 *software* [6]. Resorting to the set of non-redundant splice junctions thereby obtained, one can quantify the expression level of alternative spliced genes using MISO *software* [7]. The generated MISO output files are divided by patient, AS mechanism and tissue status (tumour or normal). A general scheme of this procedure is available in Appendix B. Each AS event has a 'percent spliced in' (Ψ) associated to it. Ψ is defined as the expression of constitutively spliced isoforms as a fraction of the total expression of both alternatively and constitutively spliced isoforms. For instance, for skipped exon (SE) events the associated Ψ fraction of mRNAs that represent the isoform including the exon (Figure 1).

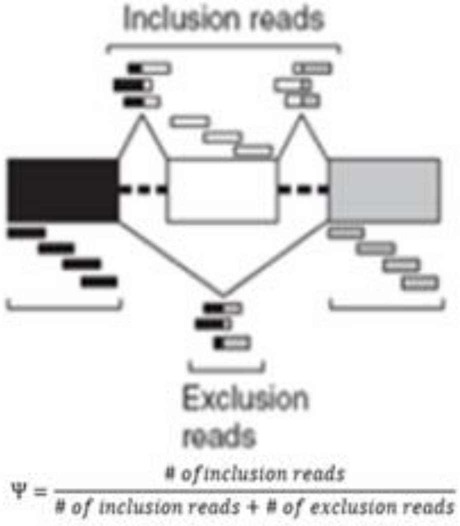

Inclusion reads
Exclusion reads
$\Psi = \frac{\#\ of\ inclusion\ reads}{\#\ of\ inclusion\ reads + \#\ of\ exclusion\ reads}$

Figure 1 – Calculation of the Ψ associated to a SE event.

Cufflinks *software* [8] 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) [9].

2.2 Identification of cancer-specific AS patterns

The identification of cancer-specific AS patterns was done using the Wilcoxon Signed Rank test, which was applied to analyse the difference between the median Ψ registered for matched tumour and normal observations for each AS event ($\Delta\widetilde{\Psi}$, expressed in Eq. 1):

$$\Delta\widetilde{\Psi} = \widetilde{\Psi}_{tumor} - \widetilde{\Psi}_{normal}, (Eq. 1)$$

where $\widetilde{\Psi}_{tumor}$ and $\widetilde{\Psi}_{normal}$are the median Ψs of the sets of tumour and normal observations, respectively. Bonferroni correction (with $\alpha = 0.01$) was applied to the results. Additionally, only events that registered a $|\Delta\widetilde{\Psi}| > 0.2$ were initially considered.

2.3. Binary tumour stage classifier

To prepare the dataset, we first selected AS events that had MISO Ψ estimates registered for all 138 patient's tumour tissues available. 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.

For each classification system described above, logistic regression with elastic net regularization (Appendix C) was run on the data of 80 randomly selected patients (the data of the remaining 58 patients was set aside to be used as test data). Various α values (0.1 to 1, with a 0.1 increment) and λ values (0.01 to 1, with a 0.01 increment) were tested. To select the classification system and parameters that provided more reliable results, the estimated deviance for each estimated model was analysed. Deviance was estimated with the 10-fold cross-validation method. The $\hat{\beta}$ which had the minimum deviance associated to it was selected. Using a receiver operater curve (ROC) curve we chose the optimum threshold. 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) [10]. The optimum threshold is the one that offers the best compromise between a lower False Positive Ratio and a higher True Positive Ratio.

2.4. Identification of independent AS prognostic factors and Gene set enrichment analysis

Survival analysis was conducted in both tumour and normal tissue estimates.

For each event, Ψ values were sorted and divided into 2 initial groups: Low PSI (composed by the 15 and 35 smallest Ψs in normal and in tumour tissue analyses respectively, the minimum number of observations considered admissible) and High PSI (composed by the remaining observations in both tumour and normal tissue analyses). 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 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. Multiple testing correction of the resulting p-values was done with the Benjamini–Hochberg procedure (with $\alpha = 0.05$) 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.

Gene set enrichment analysis (GSEA) [11] was conducted comparing the GE of the High and Low PSI groups which gave the optimum logrank p-value for any ultimately selected AS prognostic factor. Molecular Signatures available in the GSEA website (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.

3. Results

3.1. Identification of cancer-specific AS patterns

Using the aforementioned methods and parameters (section 2.2.), 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. Gene enrichment analysis was therefore conducted using DAVID's Gene Functional Classification (http://david.abcc.ncifcrf.gov/). Using this tool it was 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 (proliferation, angiogenesis, etc) were selected. To further increase the robustness of the AS event selection, only the AS events 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/), the latter being 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.

3.1.1. Fibroblast Growth Factor Receptor 2 (FGFR2)

The *FGFR2* gene is involved in important processes such as regulation of cell growth and maturation, cell division and formation of blood vessels [12].

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 2). 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 [13].

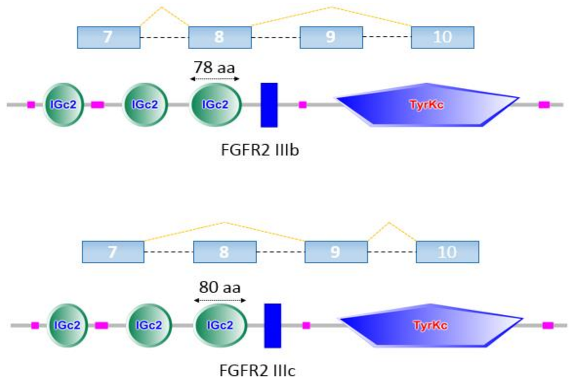


Figure 2 - AS events in *FGFR2* and protein isoforms originated from those events.

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 [14]. This tendency points to an epithelial-mesenchymal transition (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 3) [15].

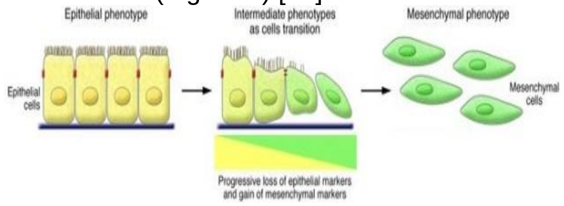

Epithelial phenotype
Intermediate phenotypes as cells transition
Mesenchymal phenotype
Epithelial cells
Mesenchymal cells
Progressive loss of epithelial markers and gain of mesenchymal markers

Figure 3 - EMT illustration. Epithelial cells tightly adhered to each other whereas mesenchymal cells are characterized by a migratory capability [16].

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 [12].

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

MCF2L codes for the guanine nucleotide exchange factor. Diseases associated with *MCF2L* include hypoparathyroidism, and spasticity. This gene is also involved in 1-phosphatidylinositol binding [17].

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 4).

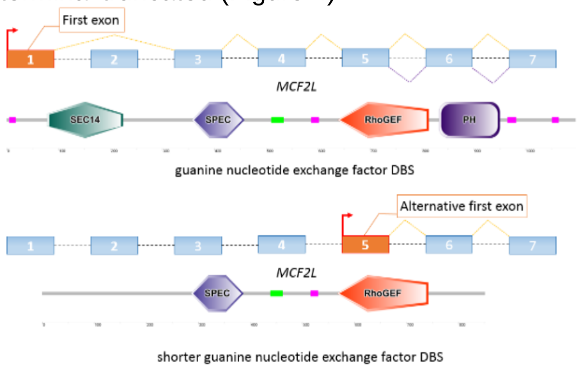

First exon
1
2
3
4
5
6
7
MCF2L
SEC14
SPEC
RhoGEF
PH
guanine nucleotide exchange factor DBS
Alternative first exon
1
2
3
4
5
6
7
MCF2L
SPEC
RhoGEF
shorter guanine nucleotide exchange factor DBS

Figure 4 - AS events in MCF2L and protein isoforms originated from those events.

The truncation of this terminal confers tumorigenic properties to this isoform, which is concordant with our analysis' results and other reports of a higher abundance of this isoform in ccRCC [17].

3.1.3. 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 [18]. Transcripts for this gene undergo complex alternative splicing that results in many functionally distinct isoforms, as shown in Figure 5.

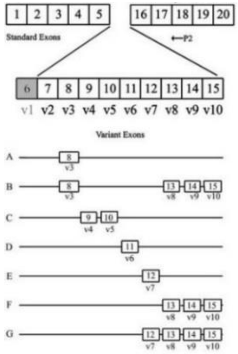


Figure 5 - Different isoforms of *CD44* gene. Exons v1 through v10 are alternative exons [19].

Our analysis points to an increase of the exclusion of exons 12 (v7) and 13 (v8), supported by a decrease in the respective median Ψ values, in tumour tissue. In the literature, it is reported that the CD44E isoform (F, in Figure 5), 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 [20]. 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 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 translates into lower production levels of CD44E isoform may suggest EMT, thus facilitating tumour metastization.

3.2. 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 value ($D = 102.96$, Figure 6.a)) 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 from the initial 18291.

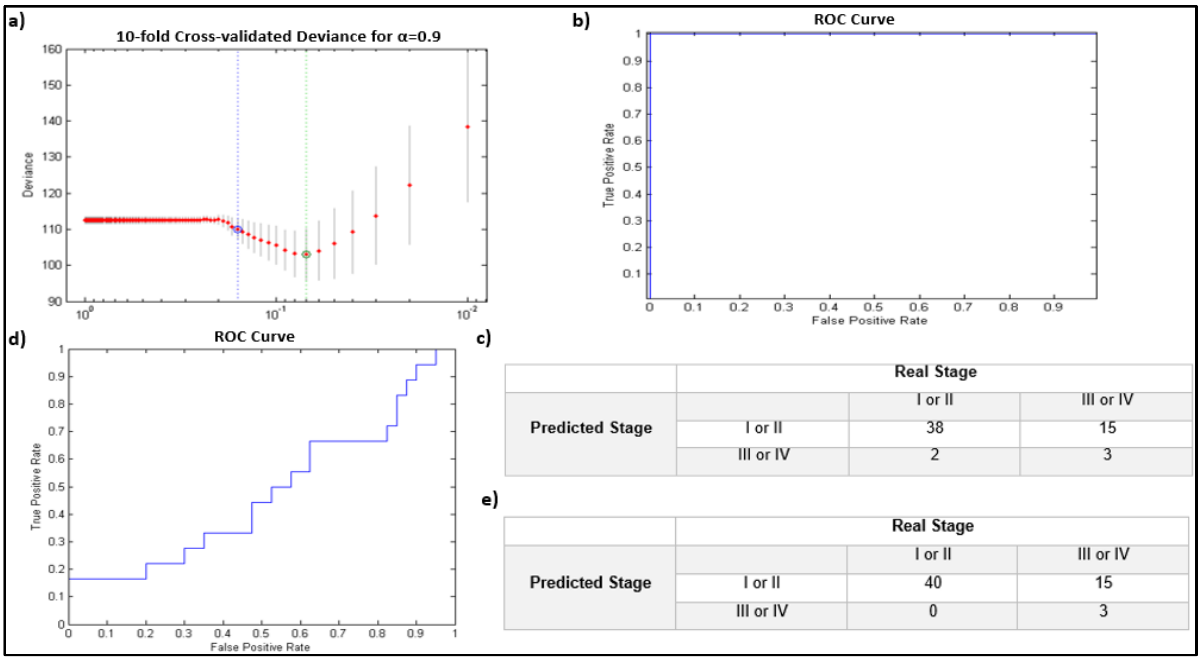
When applying ROC, an optimum threshold of 0.56 was obtained (with True Positive Rate (sensitivity) =1 and False Positive Rate (1-specificity)=0) (Figure 6.b)). This threshold provides a 100% accurate separation. The results of testing this classifier with the 58 patients that were not used in the regression are indicated in Figure 6.c). With the data gathered in Figure 6.c) one can calculate the traditional ratios that are used to access the quality of a classifier: sensitivity and specificity [21]. 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:

$$Sensitivity = \frac{TP}{TP + FN} = \frac{3}{3 + 15} = 16.7\%, (Eq. 2)$$

where TP (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 FN (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:

$$Specificity = \frac{TN}{TN + FP} = \frac{38}{38 + 2} = 95\%, (Eq. 3)$$

where TN (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 FP (False positives) is the number of patients that have a stage I or II cancer and the classifier misclassified their cancer as 1.



c)

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

e)

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

Figure 6 - a) 10-fold cross-validation plot for α=0.9. Deviance estimated for each lambda with error bars for each estimate. The traced green line indicates the lambda at which the minimum deviance is obtained; **b)** ROC curve for the model obtained using α=0.9 and λ=0.07; **c)** Table with estimated stages obtained with the classifier *vs*. real stages; **d)** ROC curve taking into account the predicted and real stages of the test subjects; **e)** Table with estimated stages obtained with the classifier *vs*. real stages using new threshold.

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 and the real cancer stage group of the test subjects. To that end a ROC curve was used (Figure 6.d)). The threshold which maximized $\|True\ Positie\ Rate - False\ Positive\ Rate\|$ was selected as the new threshold value. The new threshold value was 0.62, with results indicated in Figure 6.e). 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. 4):

$$Specificity = \frac{TN}{TN + FP} = \frac{40}{40 + 0} = 100\%. (Eq. 4)$$

3.3. Independent AS prognostic factors

After applying the methods described in section 2.4. A total of 67 AS events (from 30 genes) in normal tissue and 39 AS events (from 30 genes) in tumour tissue met the defined requirements. 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.

3.3.1. Identification of independent AS prognostic factors and GSEA 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. The PXDN protein is an extracellular matrix-associated peroxidase, thought to function in extracellular matrix consolidation, phagocytosis, and defence [22]. This gene seems to play a crucial part in Heme Oxygenase-1 tumour adhesion-promoting effects [23]

In normal tissue, the Ψ values associated to use of exon 16 as ALE seem to yield a worst clinical

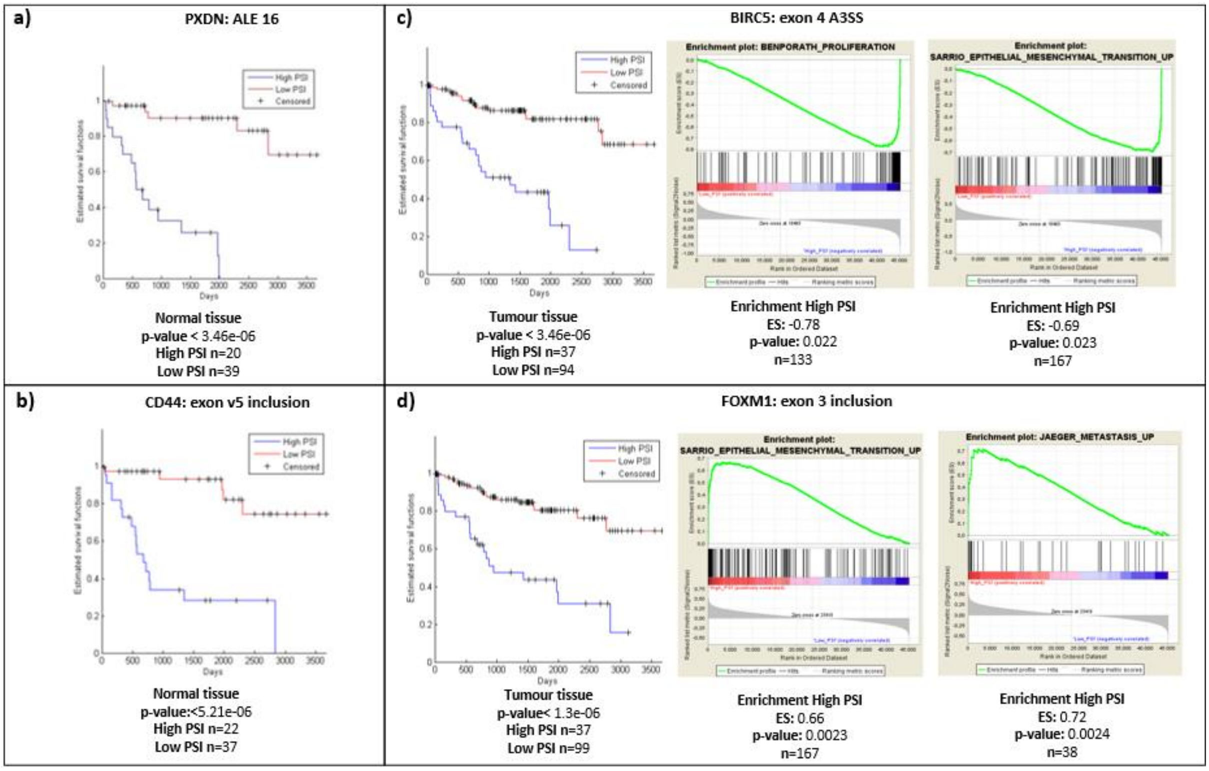


Figure 7 - a) Estimated survival functions for patients with High and Low Ψ associated to the use of PXDN's exon 16 as ALE, in normal tissue; **b)** Estimated survival functions for patients with High and Low Ψ associated to the inclusion of CD44's exon v5, in normal tissue; **c)** Estimated survival 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. Gene enrichment analysis of High and Low PSI phenotypes, associated to A3SS event in BIRC5 in tumour tissue. There is an upregulation of BENPORATH_PROLIFERATION gene set and SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP gene set in High PSI group; **d)** Estimated survival functions for patients with High and Low Ψ associated to the inclusion of FOXM1's exon 3, in tumour tissue. Gene enrichment analysis of High and Low PSI phenotypes, associated to the inclusion of FOXM1's exon 3 in tumour tissue. There is an up regulation of BENPORATH_PROLIFERATION gene set and SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP gene set in High PSI group.

outcome to the High PSI group (Ψ≥0.8) as indicated in Figure 7.a). 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. GSEA did not return relevant results.

Also in normal tissue, the inclusion of exon v5 of the *CD44* gene also seems to serve as a prognostic factor. As previously referred in section 3.1.3., 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. 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) as indicated in Figure 7.b). 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) [24]. Also, higher inclusion of exon v5-containing CD44 isoforms has been associated to poor overall survival in breast cancer [25]. 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 [26]. GSEA did not return relevant results.

3.3.2. Identification of independent AS prognostic factors and GSEA 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. The protein encoded by this gene, known as survivin, has dual roles in promoting cell proliferation and preventing apoptosis [29]. Survivin expression is turned off during fetal development and not found in non-neoplastic tissues, however it is found in most human cancers [30].

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) as indicated in Figure 7.c). 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 [30]. 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 [30]. 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. We have analysed the 39 patients from the Low PSI group and the 19 belonging to the High PSI group for which GE data were available. The GSEA indicated an upregulation of genes belonging to the BENPORATH_PROLIFERATION gene set in the High PSI group. 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 [31]. 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. This set corresponds to genes whose overexpression correlate with EMT in breast cancer [32]. 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. These results are in concordance to the lower survival rate associated to High PSI group patients.

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. 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 [33]. 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 $\Psi \geq 0.96$) as indicated in Figure 7.d). In the literature, *FOXM1* is described as only having 2 alternative exons Va and VIIa. Exon 3 is therefore reported to be constitutive [34]. 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 [35]. In addition, *FOXM1* overexpression has also been associated to EMT in pancreatic cancer [36]. We have analysed the 43 patients from the Low PSI group and the 18 belonging to the High PSI group for which GE data were available. Once again, GSEA indicated an upregulation of the genes of SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP gene set in patients who registered a higher Ψ value associated to the inclusion of exon 3 of *FOXM1* gene. A similar conclusion to the one presented in the previous paragraph 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. This set is defined by up-regulated genes in metastases from malignant melanoma compared to the primary tumours [37]. 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.

4. Discussion and conclusions

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

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

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

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

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

[7]- 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

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

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

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

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

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

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

[14]- 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

[15]- 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

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

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

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

[19]- 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

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

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

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

[23]- 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

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

[25]- 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

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

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

[28]- 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

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

[30]- 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

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

[32]- 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

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

[33]- 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

[35]- 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

[36]- 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

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

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

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

Appendix A

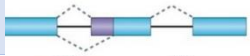
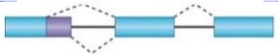
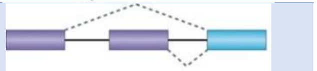
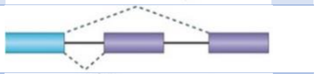
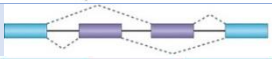
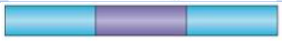
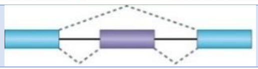
AS mechanism	Acronym	Schematic representation
Alternative 3´ splice-site selection	A3SS	
Alternative 5´ splice-site selection	A5SS	
Alternative first exon	AFE	
Alternative last exon	ALE	
Muttually exclusive exons	MXE	
Intron Retention	RI	
Skipping exon	SE	

Table 1 - Schematics and acronyms of the different types of AS [38].

Appendix B

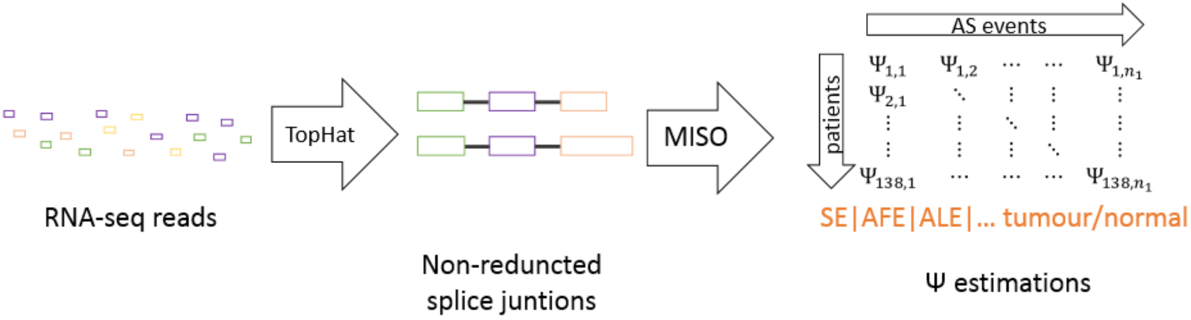

RNA-seq reads
TopHat
Non-reduncted splice juntions
MISO
AS events
patients
$\Psi_{1,1}$ $\Psi_{1,2}$ … … Ψ_{1,n_1}
$\Psi_{2,1}$
$\Psi_{138,1}$ … … … Ψ_{138,n_1}
SE|AFE|ALE|... tumour/normal
Ψ estimations

Figure 7 – 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 event and each row a patient.

Appendix C

The elastic net regression is more flexible than Lasso and Ridge regressions, since it combines the norms used by these two methods, respectively L^1 and L^2. Elastic net regression is expressed by the following equation:

$$\hat{\beta} = argmin_{\beta_0,\beta\in\square}\left(\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\left(\alpha|\beta_j| + \frac{1-\alpha}{2}\beta_j^2\right)\right), (Eq.\ 5)$$

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, β_0 is a scalar that represents the interception of the function derived by β vector, that contains the coefficients attributed to each variable, x_i is the observations registered for each predictor and λ is a positive regularization parameter, a bigger λ value reduces the number of nonzero components of β, $\sum_{j=1}^{p}|\beta_j|$ is known as L^1 norm, $\sum_{j=1}^{p}\beta_j^2$ is known as L^2 norm and α is a scalar, with a value between 0 and 1, that mediates the weight given to L^1 and L^2 norms [39].