Alternative splicing analysis for finding novel molecular signatures in renal carcinomas

Bárbara Sofia Ribeiro Caravela

Integrated Master in Biomedical Engineering 2014/15

Instituto Superior Técnico, University of Lisbon, Portugal

Abstract

The analysis of next-generation sequencing data in oncology has been primarily focused on the genome, even though most molecular oncogenic mechanisms ultimately involve transcriptomic variation. The void in current knowledge about cancer transcriptomes and particularly oncogenic alternative splicing (AS) alterations may be overcome by analyzing RNA sequencing (RNA-Seq) data. AS is an essential mechanism of gene expression regulation, allowing a single gene to give origin to various transcripts and, consequently, different proteins. The deregulation of this process is known to cause disease and there is evidence for the role of splicing regulation in cellular programs altered in oncogenesis.

This project aimed to define potential molecular signatures of prognostic value and to identify possible therapeutic targets for clear cell renal cell carcinoma (ccRCC). Thus, RNA-Seq data for tumor and matched normal renal tissue from The Cancer Genome Atlas were analyzed. The performed biostatistics analysis suggests that AS quantification could be a good measure to distinguish between tumor and normal samples. Furthermore, there are AS events that seem to be associated with survival in oncologic patients. A classification of events according to their AS quantification data distributions is also proposed, from which a loss of AS regulation in ccRCC may be inferred.

Overall, this work shows that AS quantification analysis may have a great impact in clinical practice, ameliorating health care in oncology and driving precision medicine. Besides honing diagnosis and prognosis, it can contribute to the improvement of the efficacy of current genetic therapies or even lead to innovative ones.

Keywords

Clear cell renal cell carcinoma, alternative splicing, molecular signatures, The Cancer Genome Atlas, biostatistics.

1. Introduction

A gene has its effects evidenced due to the synthesis of functional products through a process called gene expression (GE). Given that GE plays a crucial role in the maintenance of homeostasis, it has been increasingly and comprehensively studied. As part of GE, AS produces distinct mRNA sequences through the exclusion of specific exons, or parts of them, or even the retention of introns. The analysis of transcriptomic data, which are obtained from high-throughput RNA sequencing (RNA-Seq), enables the investigation of diseases at a molecular level beyond GE.

Cancer is a group of genetic diseases that may affect almost any part of the body, being characterized by abnormal cell proliferation. Due to cancer's genetic nature, previous studies using next-generation sequencing data in oncology were mostly focused on the genome. However, most molecular oncogenic mechanisms ultimately involve transcriptomic variation, with significant relations between tumor malignancy and AS alterations already been reported [1,2]. The deregulation of AS is known to cause disease and there is evidence for the involvement of splicing regulation in cellular programs modified in oncogenesis [3].

This work aimed to contribute to the understanding of altered splicing patterns in ccRCC, as well as to define novel molecular signatures with prognostic value.

2. Methods

2.1. Data collection and preparation

RNA-Seq and clinical data of patients with ccRCC were downloaded from The Cancer Genome Atlas Data Portal [4]. The cRPKM measure [5] was chosen to quantify the GE of normal and tumor samples. Using the MISO software [6], PSI values were attributed to the AS events for the same samples. PSI stands for 'percentage spliced in' and indicates the proportion of transcripts that include an alternative exon [7]. Ensembl and BioMart [8] were used to get the annotation of genes (with associated events) and transcripts, respectively.

After a filtering and pre-processing step so as to have both clinical and RNA-Seq data for all samples, data from 428 samples (366 tumor and 62 matched normal) from 366 patients were further considered. The used and/or analyzed data files are available online (imm.medicina.ulisboa.pt/group/compbio/Resources/Barbara/).

2.2. Clinical data analysis

A preliminary analysis of the clinical data was performed using the IBM SPSS Statistics software [9] (version 21). After a first descriptive statistics analysis of frequencies, some of the features were excluded from the data set according to the following criteria: presence of the same value for all observations; more than 50% of missing values; repeated or less recent data; irrelevant data for the present study. In addition, it was necessary to recode and convert specific values of the data, as well as compute a new variable (age in years) using an original feature (age in days).

The selected clinical data were also subjected to a survival analysis [10,11] using both IBM SPSS Statistics and R [12] (version 3.1.3). In R, for which RStudio [13] (version 0.98.1103) was used as interface, the package 'survival' was required.

2.3. RNA-Seq data analysis

The PSI and cRPKM high-dimensional data (with about tens to hundreds of megabytes) were analyzed using a server physically located at IDMEC. The analysis of the data was mainly performed in the server's R application, which has RStudio 0.98.1103 as interface, making use of various R packages: car, dplyr, ggbiplot, ggplot2, plyr, scales, stringr, survival, scatterplot3d, datasets, graphics, grDevices, grid, methods, stats, and utils (code also available online). Besides these, the GSEA v2.2.0 application [14,15] and the Functional Annotation Tool of DAVID 6.7 [16] were used for specific analyses regarding biological processes and gene function. Several biological databases accessible from Ensembl's site were used to collect information about the genes which were associated with events and transcripts.

Since the huge data dimensions disable the analysis of all their individual elements (events/transcripts), the applied methods aimed to reduce the set of features considering reasonable proposals of filtering and ranking.

The merged PSI matrices (samples in rows, events in columns) with just one type of samples (normal, independent tumor, or paired tumor) were subjected to a filtering step based on the percentage of missing values. The samples that lacked more than 25% of the data were excluded from the sets. Subsequently, the AS events which still presented missing values were also removed, as well as those having a null variance across tumor and normal samples. The PSI data sets were reduced from 107,336 to 43,178 AS events. Since there were no missing values in the cRPKM matrix (samples in rows, transcripts in columns), a preliminary correlation PCA [17] was performed to verify data quality. Consequently, an outlier sample exhibiting discrepant values was excluded from PSI and cRPKM data sets. The number of transcripts was modified from 51,193 to 51,122 after the exclusion of those with null variance across tumor and normal samples. Moreover, the data sets containing cRPKM values were reduced by keeping only the samples that were considered suitable for the PSI analysis. After this filtering step, 357 tumor and 55 normal samples were further considered.

Covariance and correlation PCA were performed, respectively, over various PSI and cRPKM data sets with tumor and normal samples. In this way, it was possible to scrutinize the differences between the plots of principal components in terms of grouping of samples.

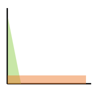
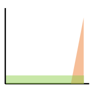
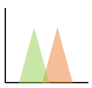
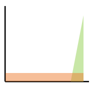
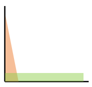
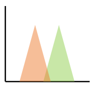
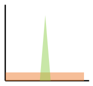
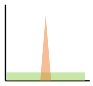
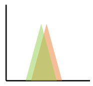
The rank product [18,19] was applied to the first 6 principal components, which explained a large part of the data's variance (approximately 53% and 24% for PSI and cRPKM, respectively). This method allowed the sorting of AS events and transcripts according to their contribution to separate normal and tumor samples.

Kolmogorov-Smirnov tests, Levene's tests, Fligner-Killeen tests, Student's t-tests, and Wilcoxon Rank Sum and Signed Rank tests [20-24] were performed over PSI and cRPKM data. The resultant p-values from the multiple testing (of all AS events/transcripts) were adjusted using the FDR correction [25]. When the corrected p-value of each test was lower than 0.05, which was the chosen significance level, the null hypothesis was rejected. The hypothesis testing enabled the identification of the AS events and transcripts with the most distinct PSI/cRPKM distributions between normal and tumor samples in terms of variance and median.

The GSEA application was used to analyze ranked lists of genes' symbols (associated with AS events/transcripts), whose scores were determined based on the p-values from Levene's tests, Wilcoxon Signed Rank tests, or rank product. The gene set databases of hallmarks, KEGG pathways, gene ontologies, and oncogenic signatures were evaluated in order to associate a biological meaning to specific differences found in PSI and cRPKM data between the types of samples.

The AS events/transcripts were classified according to the comparison of the median and variance values of PSI/cRPKM between normal and tumor samples. An enumeration of all possible combinations of how the medians and variances differ between the types of samples led to the proposed 9 classes shown in Table 1. In practice, the classification groups were formed based on the statistical significance of the p-values from Levene's tests and Wilcoxon tests for independent and paired samples.

Table 1 Proposed classification of AS events/transcripts based on PSI/cRPKM distribution for normal (n) and tumor (t) samples. The letters V and M correspond, respectively, to 'variance' and 'median'.

	V(n) < V(t)	V(n) > V(t)	V(n) = V(t)
M(n) < M(t)	LON 	HIT 	SWIN 
M(n) > M(t)	HIN 	LOT 	SWIT 
M(n) = M(t)	SIN 	SIT 	NOC 

DAVID was used in order to analyze if the AS events/transcripts (classified in the same way for independent and paired samples) within a certain group could be related to some specific gene ontology or biological pathway. The motivation for this is to better understand if there are common features among the AS events/transcripts in each class in terms of their biological functions. The analysis was performed over lists of the associated genes' Ensembl_Gene_ID considering the annotation from 4 databases (the 3 default gene ontologies and KEGG PATHWAY). The used background lists contained the Ensembl_Gene_ID of all genes related to some AS event or transcript.

In order to access the association between AS and GE, the events were mapped to the respective transcripts using the annotation tables. The AS events (and respective transcripts) with a PSI range equal to or greater than 0.5 were considered for the performance of Spearman correlation tests [26] across all samples.

To perform a survival analysis, the patients were separated according to whether their tumor samples exhibited a low or high PSI/cRPKM value for each AS event/transcript. This was done by determining, for each case, the PSI/cRPKM boundary that most significantly separated those groups of patients. In addition, Spearman correlation tests were performed over all pairs event/transcript, being the resultant p-values corrected using the FDR estimation. The survival analysis was conducted over the 1,000 AS events that were most uncorrelated with the associated transcripts, due to their potential as prognostic factors. After those transcripts were subjected to a similar survival analysis, only the AS events which revealed a significant difference between survival curves (log-rank test p-value < 0.05 after FDR correction), contrarily to their associated transcripts, were further considered. This filtering step enabled to kept just the cases for which the survival rate is influenced by AS and not GE.

3. Results and Discussion

3.1. Clinical data analysis

The age of the 366 patients, most of whom are Caucasian, varies between 27 and 89 years old. Several features were found to be associated with the survival probability (log-rank test p-value < 0.05): neoplasm histologic grade; ethnicity; tumor pathologic classification; patient's qualitative analysis results; person neoplasm status; and patient's age group. Figure 1 shows a Kaplan-Meier plot, obtained in R, depicting the survival function by tumor pathologic stage. Only the curves referring to stages I and II are not significantly different from each other.

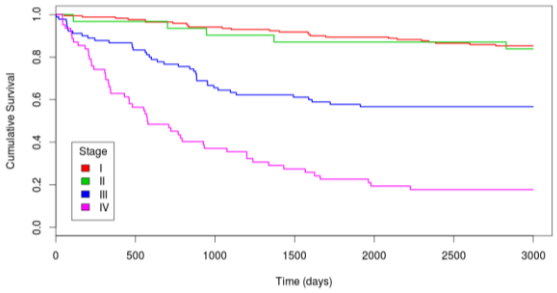


Figure 1 Kaplan-Meier plot of the survival function by tumor pathologic stage.

3.2. RNA-Seq data analysis

The main results from the PCA that was performed over the filtered PSI and cRPKM data can be seen in Figures 2 and 3. The scree plots of Figure 2 indicate that an important percentage of the variance of the PSI/cRPKM data is explained by the first 3/6 principal components. The PCA plots of Figure 3 show the grouping of normal and tumor samples according to the PSI and cRPKM data. The unexpected separation of tumor samples for PSI values was due to a bias originated by technical source. Nevertheless, it seems to exist few overlapping of groups (tumor and normal samples) in both plots.

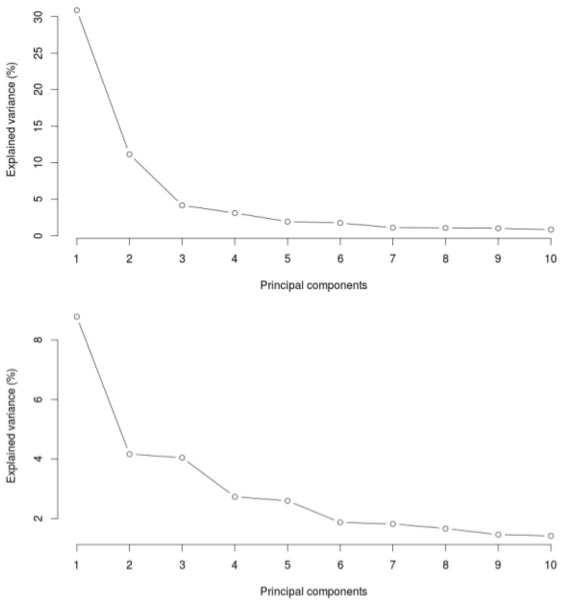


Figure 2 Scree plots for the first 10 principal components of PSI (top) and cRPKM (bottom) values.

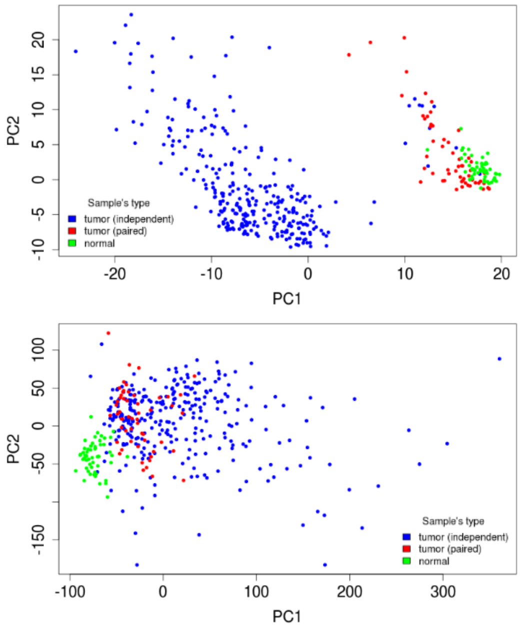


Figure 3 PCA plots for PSI (top) and cRPKM (bottom) values. Each colored dot represents a sample. PC1 and PC2 correspond to the first and second principal components, respectively.

The PSI/cRPKM distributions of the AS events/transcripts with the most significant rank product p-values are shown, respectively, in Figures 4 and 5.

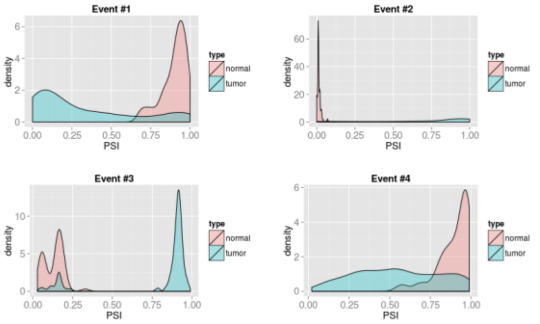


Figure 4 Distributions of PSI values for the 4 events with the most significant rank product p-values. The number presented on the title of each plot is the rank of the event.

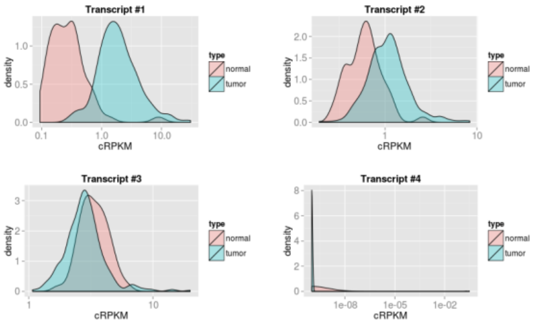


Figure 5 Distributions of cRPKM values for the 4 transcripts with the most significant rank product p-values. The number presented on the title of each plot is the rank of the transcript. The horizontal axis of each plot is in logarithmic scale.

The analysis of the PSI/cRPKM distributions enables the identification of molecular alterations caused by ccRCC. The plot of event #3 in Figure 4, for instance, appears to be the case of an isoform switch. Furthermore, the plot of transcript #1 in Figure 5 suggests an over expression of the associated gene in tumors. According to the observable differences between the types of samples, the rank product seems to be a good method to find modified AS events and relevant transcripts.

The PSI distributions of the events with the most significant p-values from Levene's and Wilcoxon Signed Rank tests are shown, respectively, in Figures 6 and 7. In other words, those figures depict the events with the greatest differences of PSI variances and medians, respectively, between normal and tumor samples. All the events of Figure 6 present a small variance for tumor samples comparatively to the normal ones, suggesting that there are specific isoforms in tumors. Curiously, the examples of Figure 7 are characterized by PSI distributions with very different variances for normal and tumor samples (and not only medians).

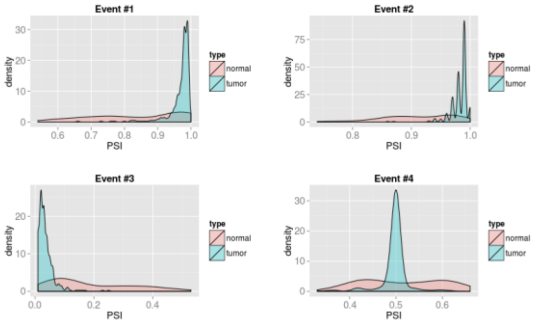


Figure 6 Distributions of PSI values for the 4 events with the most significant Levene's test p-values. The number presented on the title of each plot is the rank of the event.

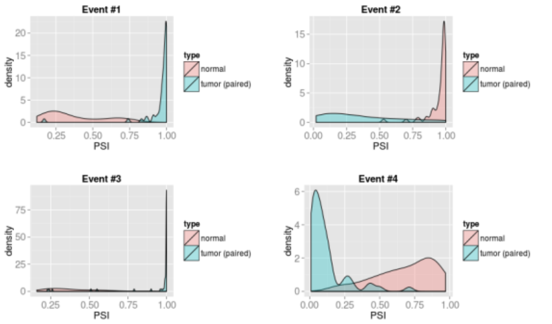


Figure 7 Distributions of PSI values for the 4 events with the most significant Wilcoxon Signed Rank test p-values. The number presented on the title of each plot is the rank of the event.

Figures 8 and 9 show the cRPKM distributions of the transcripts having the most significant p-values from the same pair of hypothesis tests. In other words, those figures depict the transcripts with the greatest differences of cRPKM variances and medians, respectively, between normal and tumor samples. All the transcripts of Figure 8 present bimodal cRPKM distributions, suggesting potentially interesting subgrouping of patients.

The heat maps of Figure 10 show the clustering of normal and tumor samples considering the AS events/transcripts having the most significant differences in PSI/cRPKM values. There is a clear distinction between the types of samples, for both types of data.

The evaluation of statistical moments of different orders, namely the mean and variance, seems to be a good approach for finding novel biomarkers of ccRCC. Like the rank product method, hypothesis testing may be used in order to get more insights about AS in cancer.

All the ranked lists referring to transcripts revealed significant results (nominal p-value, FDR q-value, and FWER p-value < 0.05) for the selected gene set databases in GSEA.

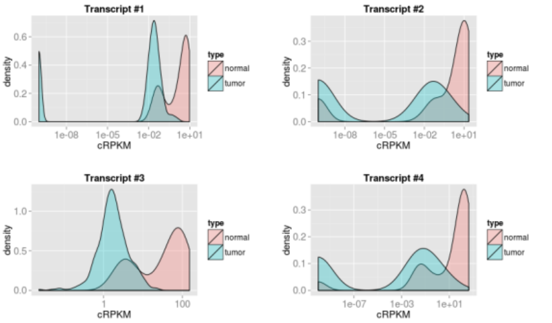


Figure 8 Distributions of cRPKM values for the 4 transcripts with the most significant Levene's test p-values. The horizontal axis of each plot is in logarithmic scale. The number presented on the title of each plot is the rank of the transcript.

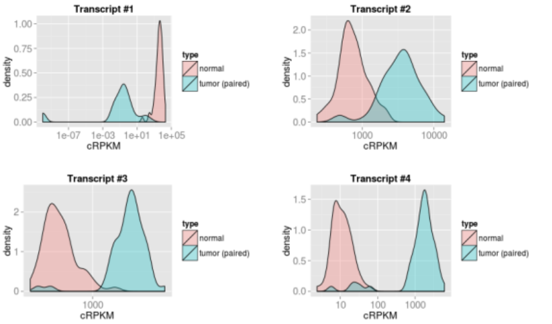


Figure 9 Distributions of cRPKM values for the 4 transcripts with the most significant Wilcoxon Signed Rank test p-values. The horizontal axis of each plot is in logarithmic scale. The number presented on the title of each plot is the rank of the transcript.

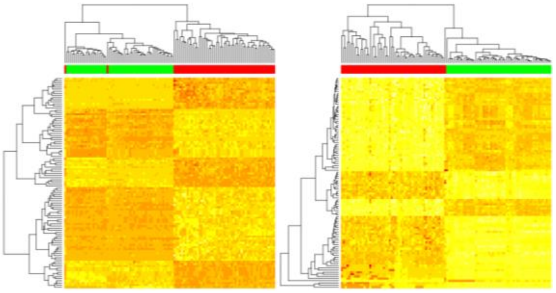


Figure 10 Heat maps of PSI (left) and cRPKM (right) values for the 100 events and transcripts (rows), respectively, with the most significant Wilcoxon Signed Rank test p-values. The normal and paired tumor samples (columns) are indicated in green and red, respectively.

Figure 11 shows the enrichment plot of one of the most enriched gene sets that were found. It refers to the biological pathway of the TCA cycle (series of chemical reactions that takes place in the mitochondrion), which is often altered in ccRCC [27]. Another result that stood out from the GSEA was the enrichment of an oncogenic signature and a hallmark associated with the oncogene *KRAS*. This gene encodes a protein that, if mutated, is known to be implicated in various types of cancer [28]. Various gene sets related to gene ontologies regarding metabolic processes were also enriched, which is consistent with the reported modifications at the metabolism level in ccRCC [27].

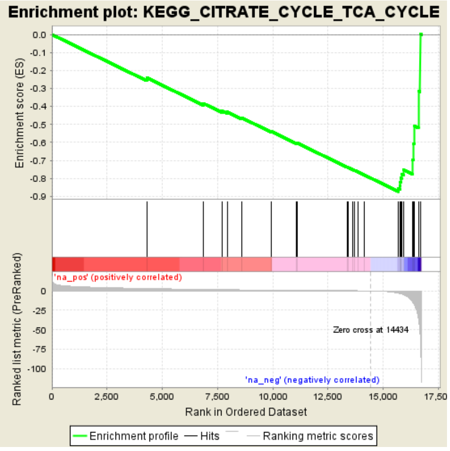


Figure 11 Enrichment plot of a gene set enriched in the list of genes associated with transcripts having a greater cRPKM variance for normal samples than for tumor ones. The ES is approximately -0.87. There are 25 genes of the set contained in the ranked list, 11 of which contribute most to the ES.

The enrichment plot of one of the two gene sets significantly enriched in phenotypes referring to events is shown in Figure 12. This set refers to a biological pathway associated with antigen presentation, which is a vital process of the immune system that is compromised in cancer [29].

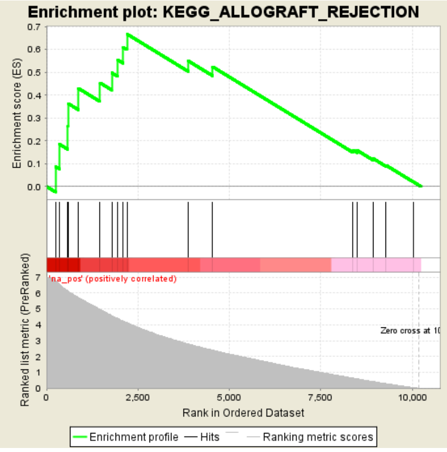


Figure 12 Enrichment plot of a gene set enriched in the list of genes associated with events having different PSI medians between normal and tumor samples. The ES is approximately 0.67. There are 17 genes of the set contained in the ranked list, 10 of which contribute most to the ES.

Overall, the GSEA confirmed that the AS events/transcripts having PSI/cRPKM distributions with different characteristics for normal and tumor samples are associated with genes altered in cancer. Therefore, ranking genes by a score based on the statistical significance of Levene's test, Wilcoxon Signed Rank test, or rank product seems relevant to find molecular signatures for ccRCC.

Representative PSI distributions of the events (equally classified for independent and paired samples) related to each distribution-based classification group are depicted in Figures 13-21. Since the density plots are generally similar to the expected results, this type of classification seems adequate to distinguish PSI distributions. The same statement can be done for cRPKM distributions, whose results are similar to the ones obtained for PSI.

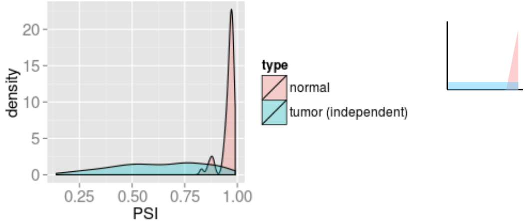


Figure 13 Distributions of PSI values for a HIN event. A schematic form of the expected distributions is also presented.

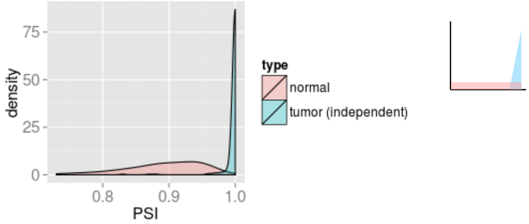


Figure 14 Distributions of PSI values for a HIT event. A schematic form of the expected distributions is also presented.

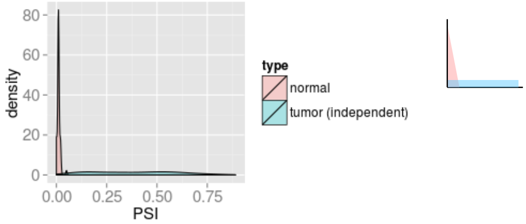


Figure 15 Distributions of PSI values for a LON event. A schematic form of the expected distributions is also presented.

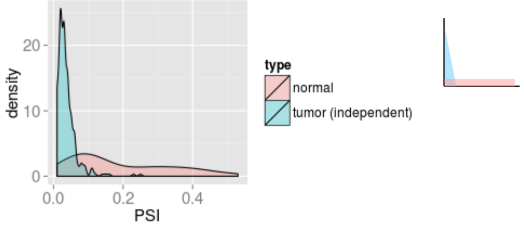


Figure 16 Distributions of PSI values for a LOT event. A schematic form of the expected distributions is also presented.

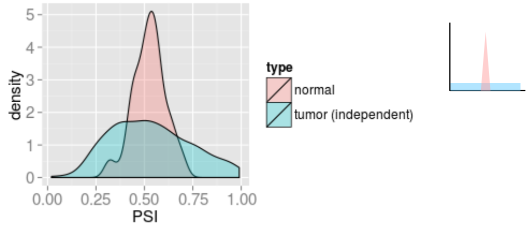


Figure 17 Distributions of PSI values for a SIN event. A schematic form of the expected distributions is also presented.

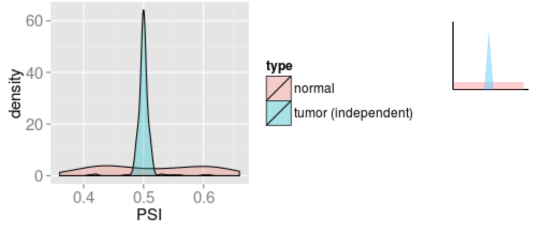


Figure 18 Distributions of PSI values for a SIT event. A schematic form of the expected distributions is also presented.

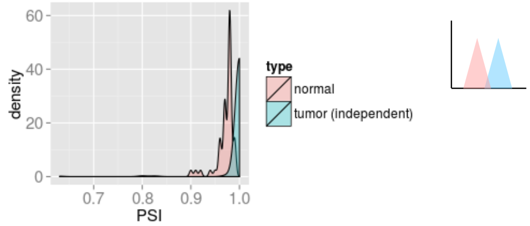


Figure 19 Distributions of PSI values for a SWIN event. A schematic form of the expected distributions is also presented.

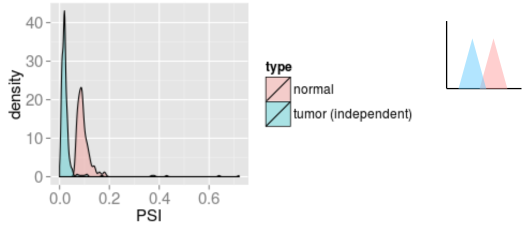


Figure 20 Distributions of PSI values for a SWIT event. A schematic form of the expected distributions is also presented.

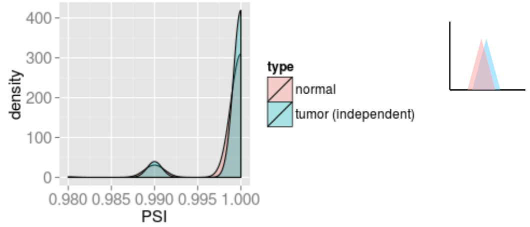


Figure 21 Distributions of PSI values for a NOC event. A schematic form of the expected distributions is also presented.

The proportions of the groups for both AS events and transcripts are very different from each other. NOC is always the most common group, so the majority of AS events/transcripts does not seem to be altered in ccRCC, as expected. From the classification of events, one may infer that there is a loss of AS regulation in ccRCC, since the PSI distributions have generally a higher variance for tumor samples than for normal ones.

Several functional annotation terms of DAVID presented significant p-values (Benjamini less than 0.05) for the various groups of transcripts and events. Because there was only one transcript classified as HIT, it was not possible to perceive the annotation associated with this particular group. From the 2,068 genes associated with the SWIT transcripts, for example, 252 (12.1%) are involved in the mitochondrion term. This result suggests that those genes are less expressed in ccRCC, which is consistent with the metabolic alterations previously reported for this disease that are related with cellular respiration [27]. Another example that may be emphasized refers to the 3,789 genes associated with the NOC events, 226 of which (6.0%) are involved in the ribonucleoprotein complex term.

The correlation between all samples, based on AS events and transcripts, may be perceived by looking at the heat maps of Figure 22. The clustering of samples is quite different when considering PSI or cRPKM values, as it was seen with the PCA results.

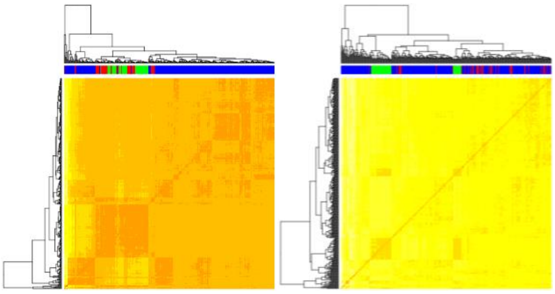


Figure 22 Heat maps of Spearman correlations between samples, based on PSIs (left) and cRPKMs (right). The independent tumor, paired tumor, and normal samples are identified in blue, red, and green, respectively.

Although there are events and transcripts that seem to be correlated with each other, there are also cases that suggest an independence between AS and GE. The identification of these situations is important for unveiling the molecular level at which the normal and tumor samples become distinct.

The survival analysis revealed 3 potential novel prognostic factors among the 1,000 AS events evaluated. Figures 23-25 show the Kaplan-Meier plots for those events and their associated transcripts. Even though the curves referring to the event of Figure 23 are significantly different, the separation of patients resulted in groups with very dissimilar dimensions (only 2 patients presenting a high PSI). The survival curves of this AS event are significantly different for other PSI boundaries, one of them leading to the identification of 15 patients with high PSI.

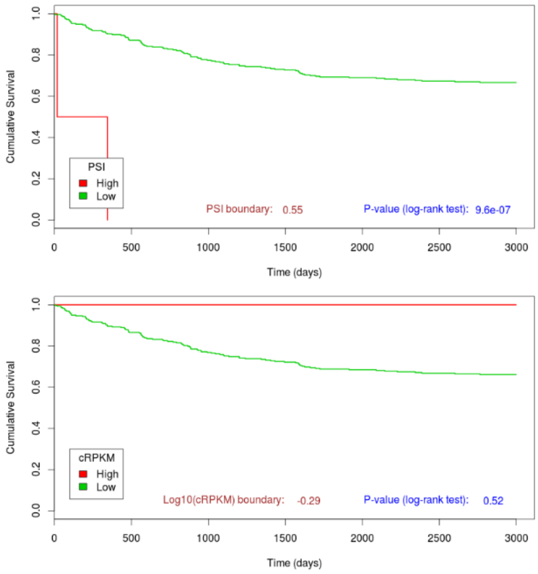


Figure 23 Kaplan-Meier plots of the AS event (top), and associated transcript (bottom), with the most significant difference between survival curves. The gene in question is *ECM2*, which codes for an extracellular matrix protein that promotes matrix assembly and cell adhesiveness [30].

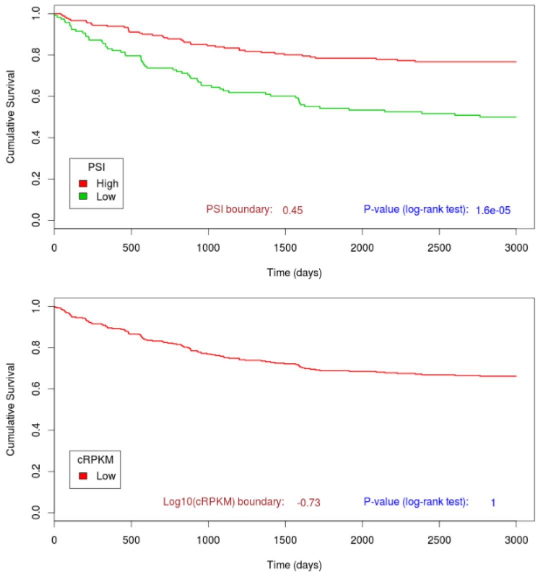


Figure 24 Kaplan-Meier plots of the AS event (top), and associated transcript (bottom), with the second most significant difference between survival curves. The gene in question is *CARKD*, which codes for a protein that catalyzes enzymatic reactions [31].

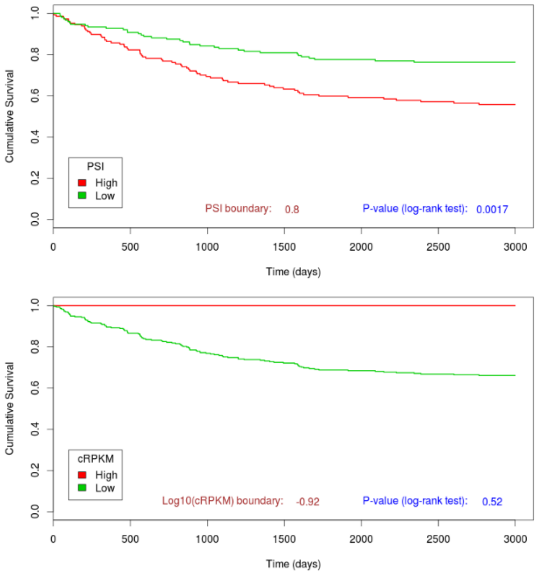


Figure 25 Kaplan-Meier plots of the AS event (top), and associated transcript (bottom), with the third most significant difference between survival curves. The gene in question is *ANKRD36B*, which codes for a protein related to protein binding [32].

The AS events whose plots are depicted in Figures 24 and 25 are characterized by important differences between survival curves that are not observable for their associated transcripts. The patients were almost or even completely undistinguishable by the cRPKM value of their tumor samples for the transcripts in question. This result is explained by the fact that the majority of the patients, or all of them, presented the same cRPKM value. The results obtained from the performed survival analysis evidence that AS may yield good prognostic factors independently from GE. Hence, their identification is clinically relevant and can be done using the proposed approach.

4. Conclusion

The present work confirmed that RNA-Seq data are an important source of information about AS alterations in cancer. The performed biostatistics analyses of transcriptomic data revealed themselves as effective in finding novel molecular signatures in ccRCC and yielding both potential novel biomarkers and prognostic factors.

Besides augmenting the current knowledge about molecular biology and cancer, AS analysis can be applied at different levels of clinical practice in oncology. The diagnosis of ccRCC may be improved by considering AS quantification, for events known to distinguish between normal and tumor samples. The patients' prognosis can also be assessed in a more accurate way by analyzing the AS events strongly associated with survival. Furthermore, a higher effectiveness of genetic therapies may be reached with the identification of therapeutic targets through AS analysis.

Acknowledgments

I thank Prof. Alexandra Carvalho and Principal Investigators Susana Vinga and Nuno Morais for their supervision. This work was partially supported by Fundação para a Ciência e a Tecnologia (FCT) through Projects CancerSys (EXPL/SEM-SIS/1954/2013) and IF (IF/00653/2012, IF/00595/2014).

References

1. Sebestyén, E. *et al*. (2015). Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Research*, *43*, 1345–1356. http://doi.org/10.1093/nar/gku1392.
2. Danan-Gotthold, M. *et al*. (2015). Identification of recurrent regulated alternative splicing events across human solid tumors. *Nucleic Acids Research*, *43*, 5130–5144. http://doi.org/10.1093/nar/gkv210.
3. Germann, S. *et al*. (2012). Splicing Programs and Cancer. *Journal of Nucleic Acids*, *2012*, 9 pages. http://doi.org/10.1155/2012/269570.
4. NIH – National Institutes of Health (2015). Site of The Cancer Genome Atlas Data Portal, https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp, accessed on 4th October 2015.
5. Labbé, R. M. *et al.* (2012). A Comparative Transcriptomic Analysis Reveals Conserved Features of Stem Cell Pluripotency in Planarians and Mammals. *Stem Cells*, *30*, 1734–1745. http://doi.org/10.1002/stem.1144.
6. Katz, Y. *et al.* (2010). Site of MISO (Mixture-of-Isoforms) software documentation, https://miso.readthedocs.org/en/fastmiso/#, accessed on 4th October 2015.
7. Katz, Y. *et al.* (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, *7*, 1009–1015. http://doi.org/10.1038/nmeth.1528. Retrieved from the site http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037023/ on 8th October 2015.
8. WTSI / EMBL-EBI – Wellcome Trust Sanger Institute / European Molecular Biology Laboratory-European Bioinformatics Institute (2015). Site of Ensembl, http://www.ensembl.org/index.html, accessed on 4th October 2015.
9. IBM (2015). Site of IBM, http://www-01.ibm.com/software/analytics/spss/products/statistics/, accessed on 5th October 2015.
10. Bewick, V. *et al.* (2004). Statistics review 12: Survival analysis. *Critical Care*, *8*, 389–394. http://doi.org/10.1186/cc2955.
11. Bland, J. M. (2004). The logrank test. *BMJ*, *328*, 1073–1073. http://doi.org/10.1136/bmj.328.7447.1073.
12. The R Foundation (2015). Site of R, https://www.r-project.org/, accessed on 5th October 2015.
13. RStudio (2015). Site of RStudio, https://www.rstudio.com/, accessed on 5th October 2015.
14. Subramanian, A. *et al.* (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*, 15545–15550. http://doi.org/10.1073/pnas.0506580102.
15. Broad Institute (2015). Site of Gene Set Enrichment Analysis, http://www.broadinstitute.org/gsea/index.jsp, accessed on 5th October 2015.
16. National Institute of Allergy and Infectious Diseases (NIAID), NIH (2015). Site of DAVID Bioinformatics Resources 6.7, https://david.ncifcrf.gov/home.jsp, accessed on 5th October 2015.
17. Abdi, H. & Williams, L. J. (2010). Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*, 433–470. http://doi.org/10.1002/wics.101.
18. Koziol, J. A. (2010). Comments on the rank product method for analyzing replicated experiments. *FEBS Letters*, *584*, 941–944. http://doi.org/10.1016/j.febslet.2010.01.031.
19. Koziol, J. A. (2010). The rank product method with two samples. *FEBS Letters*, *584*, 4481–4484. http://doi.org/10.1016/j.febslet.2010.10.012.
20. NIST/SEMATECH (2013). Site of Engineering Statistics Handbook, http://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm, accessed on 2nd October 2015.
21. Chang, W. (2013). Site of Cookbook for R, http://www.cookbook-r.com/Statistical_analysis/Homogeneity_of_variance/, accessed on 2nd October 2015.
22. Student (1908). The Probable Error of a Mean. *Biometrika*, *6*, 1–25. Retrieved from the site http://www.aliquote.org/cours/2012_biomed/biblio/Student1908.pdf on 2nd October 2015.
23. UC Regents (2015). Site of Institute for Digital Research and Education – UCLA, http://www.ats.ucla.edu/stat/mult_pkg/faq/general/mann-whitney.htm, accessed on 2nd October 2015.
24. Graph Pad Software, Inc. (2015). Site of GraphPad Software, http://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_nonparametric_tests_dont_compa.htm, accessed on 2nd October 2015.
25. Noble, W. S. (2009). How does multiple testing correction work? *Nature Biotechnology*, *27*, 1135–1137. http://doi.org/10.1038/nbt1209-1135.
26. Croux, C. & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications*, *19*, 497–515. http://doi.org/10.1007/s10260-010-0142-z.

27. The Cancer Genome Atlas Research Network (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, *499*, 43–49. http://doi.org/10.1038/nature12222.
28. National Center for Biotechnology Information, U.S. National Library of Medicine (2015). Site of NCBI, http://www.ncbi.nlm.nih.gov/gene/3845, accessed on 27th October 2015.
29. Hanahan, D. & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*, 646–674. http://doi.org/10.1016/j.cell.2011.02.013.
30. UniProt Consortium (2015). Site of UniProt, http://www.uniprot.org/uniprot/O94769, accessed on 29th October 2015.
31. UniProt Consortium (2015). Site of UniProt, http://www.uniprot.org/uniprot/Q8IW45, accessed on 29th October 2015.
32. EMBL-EBI (2015). Site of InterPro, http://www.ebi.ac.uk/interpro/protein/Q8N2N9;jsessionid=7F2FF2AEF6D57F73E102BD5D5332C7C6, accessed on 29th October 2015.