Bioinformatical Analysis of Alternative Splicing

By

JUAN PABLO ROMERO RIOJAS

TECNUN
UNIVERSITY OF NAVARRA

A dissertation submitted to the University of Navarra in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in TECNUN, Faculty of Engineering.

SEPTEMBER 2017

Donostia - San Sebastian
七転び八起き
nanakorobi yaoki
Abstract

Splicing is a natural process happening in every living cell. As a fundamental process, all genes undergo splicing before tuning into a functional molecule such as proteins. However, the splicing process is still under active research in order to completely understand how it is regulated. As of now, the splicing machinery is characterized in what we call the spliceosome, but a complete understanding of it is still under active research. The discovery of alternative splicing set a breakthrough in molecular biology. This process allows a simple gene to turn into more than one protein by creating different transcripts or isoforms. The transcriptome is the whole set of known isoforms that have been characterized through research and experimental procedures. Still, many isoforms might remain unknown. Through research, alternative splicing has been shown to be responsible for the development of different pathologies, including cancer. On the other hand, alternative splicing provides a novel mean to characterize potential biomarkers for drug resistance or survival. Due to this, alternative splicing is constantly being studied in order to completely understand its functioning mechanism.

With the development of sequencing techniques, it has been possible to develop methods to quantify the different isoforms present in specific samples or conditions. From the method, developed by Frederick Sanger, also known as Sanger sequencing, to the development of new protocols, like third generation sequencers, vast amounts of data have been generated. Eventhough, sequencing has taken a lead role in the study of alternative splicing, other platforms like junction arrays have been developed to study such phenomenon. Affymetrix recently developed the Clariom microarray, which seems to be the most up to date array to identify splicing events.

Most of the available algorithms to identify and quantify alternative splicing events, provide more than one figure of merit per event and does not take into account a coherence in such events. For example, in a cassette exon, not only the skipped exon should show a change in expression, but the flanking junctions should display a similar behaviour but in opposite directions. In order to take this into account, a novel method to identify, classify and state statistical significance of splicing events has been developed. EventPointer allows users to identify alternative splicing events and provide the statistical significance of such events. The algorithm can be applied to data from both microarrays or RNA-Seq. Also, EventPointer generates files that can be loaded into genome browsers to ease the interpretation of the results and the desing of primers for standard PCR validations.

The performance of EventPointer has been tested in two independent experiments using
both platforms. The overall results show a promising validation rate in both technologies. EventPointer, also estimates the percent spliced index for every detected event and not only skipping exons, as most of the available software. The results, obtained through end-point PCR demonstrate that the estimated $\Psi$ values, provided by EventPointer, are highly correlated with the experimental results. EventPointer shows an improved method to identify and quantify alternative splicing events.

A comparison between microarrays and RNA-Seq, in their ability to identify alternative splicing events was performed using the same experimental data from three different cell lines treated with a drug that severely affects the splicing machinery. The results show that RNA-Seq is the most flexible and trustable platform for the identification of splicing events, but microarrays are a viable option to analyse alternative splicing due to reasons of cost and convenience. Microarrays can be an alternative when compared to shallow sequencing.
Antes que nada, quiero agradecer a mi director de tesis, el Dr. Ángel Rubio. Gracias Ángel, por todos los consejos durante estos años. Gracias por ayudarme a crecer tanto en lo profesional como en lo personal.

Me gustaría agradecer a mi familia. A mis padres y a mis hermanos por todo el apoyo que me han mostrado durante estos años. Un simple mensaje o una llamada fue suficiente para empezar el día con más energía y, sobretodo, con ánimos para seguir haciendo esto que tanto disfruto.

También quiero mencionar a todos los compañeros de despacho que he tenido durante estos años: Dr. Francis Planes, Dr. Ander Aramburu, Dr. Ander Muniategui, Dr. Luis Tobalina, Iñigo Apaolaza y Fernando Carazo. Llegar al despacho y encontrar gente así era más que suficiente para tener un buen día y afrontar cualquier reto que se presentara.

Quiero hacer una especial mención a todos los amigos que he hecho durante estos años. Desde que empezamos viviendo juntos en el Colegio Mayor Ayete, hasta ahora que estamos todos repartidos por el mundo. Charpen, Alfredo, Yeyo, Pollo, Echániz, Fran, Iñigo, Jaime, Johnny, Cheché, Richard y Tano. ¡Gracias a todos por tanto!

I would like to say some words in English just to thank all the people I had the chance to meet during my short stay in Leiden, The Netherlands. First of all, thanks Dr. Eric Van der Veer for making me feel at home during those 3 months. I really enjoyed spending time over there. Hope that in the future, not so far, I can visit you guys again.

Gracias también a todos los alumnos, amigos y compañeros durante estos años.

Deo Gratias!
# Table of Contents

<table>
<thead>
<tr>
<th>List of Tables</th>
<th>xi</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
</tbody>
</table>

1 Biology Behind Alternative Splicing  
1.1 Cell biology  
1.2 From DNA to proteins  
1.2.1 Protein post-translational modifications  
1.3 Splicing mechanism  
1.4 Alternative splicing  
1.4.1 Canonical classification of alternative splicing  
1.4.2 Regulation of alternative splicing  
1.4.3 Importance of alternative splicing  

2 Bioinformatics On Alternative Splicing  
2.1 Microarrays  
2.1.1 Oligonucleotide microarrays  
2.1.2 Splicing arrays  
2.1.3 Microarray data processing  
2.2 Next generation sequencing  
2.2.1 Sanger sequencing  
2.2.2 Second generation sequencers  
2.2.3 Third generation sequencers  
2.2.4 NGS data processing  
2.2.5 Visualization of NGS data  
2.3 Algorithms for the detection of alternative splicing  

vii
# TABLE OF CONTENTS

## 3 EventPointer
- 3.1 Overview of the algorithm ........................................... 33
- 3.2 Splicing graph creation ............................................. 34
  - 3.2.1 Junction arrays ............................................. 36
  - 3.2.2 RNA-Seq .................................................. 37
- 3.3 Event identification and classification ............................. 37
  - 3.3.1 Identification of alternative splicing events ............... 37
  - 3.3.2 Classification of alternative splicing events .............. 39
- 3.4 Statistical analysis ................................................ 40
  - 3.4.1 Differential expression ...................................... 40
  - 3.4.2 Differential splicing ......................................... 42
  - 3.4.3 Estimation of the percent spliced index ..................... 45
- 3.5 Visualization ...................................................... 47
- 3.6 Installation ........................................................ 47

## 4 Application of EventPointer to SRSF1
- 4.1 Serine/Arginine-rich splicing factor 1 ............................. 49
- 4.2 Experimental design .............................................. 50
- 4.3 Differential expression at gene level ............................. 51
- 4.4 Determining differential splicing events ......................... 55
- 4.5 Protein domains .................................................. 56
- 4.6 Comparison with other software .................................. 59
  - 4.6.1 TAC 3.0 .................................................... 59
  - 4.6.2 AltAnalyze ............................................... 60
  - 4.6.3 Comparison of top ranked events ............................ 61
- 4.7 Discussion ....................................................... 64

## 5 Comparison of Microarrays and RNA-Seq
- 5.1 Introduction ...................................................... 67
- 5.2 Experimental design .............................................. 68
- 5.3 Matching events .................................................. 70
- 5.4 Results .......................................................... 71
  - 5.4.1 PCR validation ............................................. 74
- 5.5 Discussion ....................................................... 78

## 6 Conclusion and Future Lines

viii
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Conclusions</td>
<td>81</td>
</tr>
<tr>
<td>6.2 Future lines</td>
<td>83</td>
</tr>
<tr>
<td>6.2.1 Multi-path events</td>
<td>83</td>
</tr>
<tr>
<td>6.2.2 Automatic design of PCR primers</td>
<td>83</td>
</tr>
<tr>
<td>6.2.3 Junction arrays probe quality</td>
<td>84</td>
</tr>
<tr>
<td>6.2.4 Biological interpretation of alternative splicing</td>
<td>84</td>
</tr>
<tr>
<td>A Primers SRSF1 experiment</td>
<td>85</td>
</tr>
<tr>
<td>B Primers CX4945 experiment</td>
<td>89</td>
</tr>
<tr>
<td>C Additional figures microarrays vs RNA-Seq</td>
<td>93</td>
</tr>
<tr>
<td>D EventPointer in BMC Genomics</td>
<td>99</td>
</tr>
<tr>
<td>E EventPointer poster</td>
<td>119</td>
</tr>
<tr>
<td>Bibliography</td>
<td>123</td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Genes whose splicing variants are associated to pathological conditions.</td>
</tr>
<tr>
<td>2.1</td>
<td>Affymetrix splicing arrays</td>
</tr>
<tr>
<td>3.1</td>
<td>Value of the coefficients for the splicing linear model and their interpretation.</td>
</tr>
<tr>
<td>4.1</td>
<td>Top 5 differentially expressed genes in SRSF1 knock down experiment</td>
</tr>
<tr>
<td>4.2</td>
<td>Top 5 differentially expressed long intergenic noncoding RNAs in SRSF1 knock down experiment</td>
</tr>
<tr>
<td>4.3</td>
<td>Enriched gene ontology terms for the SRSF1 experiment.</td>
</tr>
<tr>
<td>4.4</td>
<td>Tested splicing events for SRSF1 dataset</td>
</tr>
<tr>
<td>4.5</td>
<td>Enrichment of domains in the list of splicing events regulated by SRSF1</td>
</tr>
<tr>
<td>4.6</td>
<td>Comparison of top ranked events</td>
</tr>
<tr>
<td>5.1</td>
<td>Detected alternative splicing events for each platform</td>
</tr>
<tr>
<td>5.2</td>
<td>Detected alternative splicing events using both technologies</td>
</tr>
<tr>
<td>5.3</td>
<td>PCR validation for RNASeq and microarrays</td>
</tr>
<tr>
<td>5.4</td>
<td>Computational requirements</td>
</tr>
<tr>
<td>A.1</td>
<td>PCR primers for SRSF1 experiment</td>
</tr>
<tr>
<td>B.1</td>
<td>PCR primers for CX4945 experiment</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Eukaryotic cell diagram</td>
<td>2</td>
</tr>
<tr>
<td>1.2 DNA packaging</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Post-transcriptional pre-mRNA processing</td>
<td>5</td>
</tr>
<tr>
<td>1.4 Amino acid table</td>
<td>6</td>
</tr>
<tr>
<td>1.5 Protein post-translational modifications</td>
<td>8</td>
</tr>
<tr>
<td>1.6 Spliceosome assembly and mechanism</td>
<td>11</td>
</tr>
<tr>
<td>1.7 Graphical representation of alternative splicing</td>
<td>12</td>
</tr>
<tr>
<td>1.8 Canonical classification of alternative splicing</td>
<td>13</td>
</tr>
<tr>
<td>2.1 Affymetrix oligonucleotide microarrays</td>
<td>19</td>
</tr>
<tr>
<td>2.2 Sanger sequencing</td>
<td>24</td>
</tr>
<tr>
<td>2.3 Oxford Nanopore MinION™</td>
<td>26</td>
</tr>
<tr>
<td>2.4 Sashimi Plot</td>
<td>29</td>
</tr>
<tr>
<td>2.5 BAM files in IGV</td>
<td>32</td>
</tr>
<tr>
<td>3.1 Overview of EventPointer</td>
<td>35</td>
</tr>
<tr>
<td>3.2 Splicing graph</td>
<td>36</td>
</tr>
<tr>
<td>3.3 Classification of alternative splicing</td>
<td>40</td>
</tr>
<tr>
<td>3.4 Visualization of splicing events</td>
<td>48</td>
</tr>
<tr>
<td>4.1 Gene expression heatmap</td>
<td>52</td>
</tr>
<tr>
<td>4.2 lincRNAs expression heatmap</td>
<td>53</td>
</tr>
<tr>
<td>4.3 Number of splicing events by type</td>
<td>55</td>
</tr>
<tr>
<td>4.4 SRSF1 PCR validation</td>
<td>58</td>
</tr>
<tr>
<td>4.5 Venn diagram of common events</td>
<td>62</td>
</tr>
<tr>
<td>4.6 NAA50 and DNM2 PCR Validation</td>
<td>64</td>
</tr>
<tr>
<td>5.1 Sankey diagram of the relationship between matched events</td>
<td>74</td>
</tr>
</tbody>
</table>
5.2 Estimated PSI for three technologies ............................................... 76
5.3 Increment of $\Psi$ for both microarrays and RNA-Seq ....................... 77
5.4 Number and type of detected events for each platform ...................... 78

C.1 Scatterplot of the splicing Z-value ................................................. 93
C.2 Estimated $\Psi$ for $R^+ M^+$ .......................................................... 94
C.3 Estimated $\Psi$ for $R^+ M^\emptyset$ ...................................................... 95
C.4 Estimated $\Psi$ for $R^+ M^-$ .......................................................... 96
C.5 Estimated $\Psi$ for $R^\emptyset M^+$ ...................................................... 97
C.6 Estimated $\Psi$ for $R^- M^-$ .......................................................... 98
The necessary instructions required to carry out all the processes for the correct functioning and development of organisms are encoded in the genome. This information is contained in the DNA and is tightly packed inside the nucleus of every cell. Every specific function requires proteins to be carried out properly. However, the process of transforming DNA into functional molecules is extremely complex and, probably, still not completely deciphered by researchers. One of the steps for protein synthesis is splicing, which relies on a complex molecular machinery known as the spliceosome to remove non-coding regions. With the discovery of alternative splicing, it was possible to state that one single gene could produce different proteins and as a result, the central biology dogma - one gene encodes one protein - was proved wrong.

1.1 Cell biology

Living organisms are classified into hierarchical groups, known as taxons according to shared characteristics among them. These groups are called taxonomic ranks and are used to distinguish different levels of similarities between all organisms. The broadest rank corresponds to Domain, which classifies life forms based on the properties of the basic structure of life: the cell. The Domain taxon is divided into three subsets: Archaea, Bacteria and Eukarya. Both Archaea and Bacteria group unicellular organisms made up by prokaryotic cells, while complex multicellular organisms composed of eukaryotic cells form Eukarya.
Prokaryotic cells lack a nucleus and all its genomic information is dispersed in the cytoplasm. On the other side, eukaryotic cells are complex structures that contain membrane-bound organelles and a nucleus to store the genetic material (see Figure 1.1). Organelles (e.g. golgi apparatus, mitochondria, ribosomes) are specialized compartments that develop specific functions for the cell, from protein synthesis to selective permeability, all organelles contribute to the correct functioning of the cell.

The nucleus is an spherical-like organelle with specific components. It is surrounded by the nuclear envelope that serves as a barrier from the cytoplasm; its membrane holds tiny apertures, known as nuclear pores, that allow certain molecules in and out of the nucleus and it is also connected to a network of microtubules, called the endoplasmic reticulum, where proteins are modified and sent to the proper location to perform their function.

All the necessary instructions required by the cell are located inside the nucleus in the form of deoxyribonucleic acid (DNA). This information is composed of four chemical bases (Adenine, Thymine, Cytosine and Guanine) attached to a sugar and a phosphate group to assemble what is known as nucleotides. These chemical bases can be divided into purines, which contain two carbon nitrogen rings and include adenine and guanine, while thymine and cytosine only hold one carbon nitrogen ring and are named pyrimidines.
The sequence of nucleotides determines the set of instructions to be carried out by the cell and these specific sequences are commonly referred to as genes. The bases of DNA align with each other by complementary base pairing, i.e., A-T and C-G to form two strands that are organized into a double helix structure. The strands are oriented in opposite directions to allow base pairing. In order to identify each of them, the nomenclature positive and negative strand has been adapted. Both ends of the DNA sequence are referred to as the 5’ and 3’ end, respectively, where the 5’ is the leftmost end if read from left to right. The positive strand goes from 5’ to 3’ while the negative one in the opposite direction.

Through research, it has been estimated that the human genome contains about 3 billion base pairs. All this information is stored inside the nucleus of every cell, so an extremely efficient packaging method is required to store the DNA. This is done by histones, a special type of proteins that compact DNA into such a microscopic space. Due to phosphate groups, chromosomal DNA has a negative charge, whereas histones are positively charged. This allows DNA to coil around the histones to form a DNA-protein complex known as chromatin. As this process is repeated throughout the whole genome, a basic structural unit is created called nucleosomes. Each nucleosome consists of eight histones and approximately 146 base pairs. Different histones are used during the packaging of DNA (e.g., H1, H2A, H2B, H3 and H4). When the histone H1 is added to a nucleosome, a different structure is formed known as chromatosome. All these molecules together create a fiber-like structure which is organized into what we know as chromosomes. All the process is depicted in Figure 1.2.

In normal conditions, every cell holds twenty three pairs of chromosomes, that are divided into twenty two pairs of autosomes (1-22) and one pair of sex chromosomes (X and Y).

1.2 From DNA to proteins

When a gene is said to be expressed, it means that the information contained in it is used to obtain a functional product. In order to transform DNA into protein, a specific process is performed to transfer the information, encoded in the DNA, into the building blocks of proteins or amino acids and then obtain a functional molecule.

The first part of this process is transcription, which is divided into three steps: initiation, elongation and termination. In a simplified explanation, during transcription a sequence of nucleotides in the double stranded DNA is used as a template to generate a single stranded
Figure 1.2: The different levels of the structures that are created to pack DNA inside the nucleus of a cell. Figure taken from [11]

ribonucleic acid (RNA\(^1\)) molecule.

The initiation of transcription begins when RNA polymerase II (RNA pol II), an enzyme, binds to the gene that is being transcribed in a specific region called promoter. This sequence usually contains a TATA box, which is a region that is enriched in A and T nucleotides [26]. The promoter is usually located ten base pairs upstream from the site where transcription starts, although it can be located in nearby regions [26]. Upon binding, RNA pol II binds to another molecule known as the sigma subunit which forms a holoenzyme [26] that is capable of unwinding the strands of DNA. Once the RNA polymerase II is bound to the DNA and the strands are separated, the elongation step of transcription starts.

During elongation, the DNA strand serves as a template to add the corresponding com-

\(^{1}\)In RNA, the Thymine nucleotide from DNA is changed into Uracil (U), which pairs with Adenine
1.2. FROM DNA TO PROTEINS

A complementary nucleotide (changing T for U) to form a single stranded RNA molecule. This process is carried out in a specific direction: starting in the 5’ end and finishing in the 3’ one. This step is performed until a terminator sequence is detected, which triggers specific signals to initiate termination [26].

Once termination is triggered, cleavage factors bind to specific sequences in the RNA strand to obtain one functional strand and a disposed one that is later degraded. The resulting product is a precursor messenger RNA molecule (pre-mRNA) that still requires to undergo post-transcriptional processing before becoming a mature messenger RNA (mRNA).

In the post-transcriptional processing, both ends of the pre-mRNA are modified by adding specific sequences. The 5’ end is capped with an altered nucleotide that allows other molecules to recognize the strand before translation. The opposite end of the strand (3’), is modified by the addition of a repetitive string of adenine, known as poly-A tail [27]. The addition of this string is performed by poly(A) polymerase, which functions in a similar manner as RNA pol II.

At this point, the result from transcription and its further modifications is a single stranded pre-mRNA. However, in this chain there are two types of nucleotide sequences: the portion of nucleotides that will actually be transformed into proteins, also known as exons, and non-coding regions that are called introns. It is necessary to remove introns from the pre-mRNA molecule before any further transformation and this is achieved by a process known as splicing, which is still considered as one of the post-transcriptional steps. This process along with its mechanism is described in detail in section 1.3.

![Diagram of post transcriptional processing to the pre-mRNA molecule.](image)

**Figure 1.3:** Diagram of post transcriptional processing to the pre-mRNA molecule. Figure taken from [87]

Once the splicing procedure is completed, the resulting product is a mature mRNA
CHAPTER 1. BIOLOGY BEHIND ALTERNATIVE SPLICING

This molecule moves from the nucleus into the cytoplasm where the protein synthesis responsible organelle is located. These specialized organelles are ribosomes and are composed of two subunits: 50S (large subunit) and 30S (small subunit) \[27\]. Each of them contain specific proteins and RNA molecules, such as ribosomal RNA (rRNA) and transfer RNA (tRNA) to perform the translation procedure.

The translation process can be described, alike transcription, in three main steps: initiation, elongation and termination. The difference is that instead of building a chain of nucleotides, a chain of amino acids is built. Amino acids are the building blocks of proteins and, as of now, twenty have been characterized. Each amino acid is defined by a triplet of nucleotides, called codon; the relationship of codons and amino acids is displayed in the chart depicted in Figure 1.4.

![Figure 1.4: Relationship between amino acids and nucleotides. Each codon represents a particular amino acid. The start codon AUG is related to methionine amino acid and the codons UAA, UAG and UGA serve as stop codons. © 2014 Nature Education All rights reserved](image)

Both the tRNA and rRNA play specific roles during translation. The transfer RNA serves as an adaptor: one end reads the sequence in the mRNA and binds to a specific codon by complementary base pairing, while the other end, carries the corresponding amino acid to
be added into the growing chain. Ribosomal RNA is a catalyzer during the addition of new amino acids [23].

During translation initiation, the mRNA binds to the ribosomes and is read to detect the start codon, identified by the AUG sequence. When located, the methionine amino acid is added and serves as the starting point of the newly built protein [27]. Once initiation is completed, the amino acid chain continues to grow by reading the corresponding codon and adding the proper amino acid by base pairing with tRNA molecules, this process is often referred to as elongation and continues until one of the stop codons (UAA, UAG or UGA) is detected, initiating the terminantion step. No tRNAs recongize the stop codons, but other molecules, also known as release factors, bind to help in releasing the mRNA molecule [27].

The functionality of the new molecule is not only defined by its aminoacid sequence, but on its structure, conformational disposition and chemical modifications.

Protein structure can be divided into four levels: 1) *Primary structure*, defined by the sequence of amino acids in the protein, 2) *Secondary structure*, three dimensional disposition of local segments (α-helices and β-sheets), 3) *Tertiary structure*, corresponds to the three dimensional conformation of the protein and 4) *Quaternary structure*, conformational change due to interaction with other proteins.

1.2.1 Protein post-translational modifications

Protein post-translational modifications (PTMs) are chemical modifications in proteins that lead to functional changes. These modifications help to regulate activity, localization and interaction with other molecules.

Many of the functions carried out by proteins are context specific and can be modified due to enviromental stimuli. All these responses are controlled by PTMs and can be activated or inactivated, depending on different factors. PTMs can occur in different regions of the protein such as amino acid side chains or peptide linkages which are mediated by enzymes. Some of the enxymes that rule PTMs include kinases, phosphatases, transferases and ligases.

Protein PTMs can occur at any step in the cycle of proteins. Some occur right after translation, to ensure the proper localization of the resulting protein, while others can be caused after localization and folding to achieve a proper biological activation.

One of the key aspects of PTMs include a reversible property. For example, kinases are used to phosphorylate proteins in different chains as a common method for catalytic activation or inactivation. On the other side, phosphathases hydrilyze the phosphate group
to remove it from the protein and reverse its biological activity.

**Figure 1.5:** Proteome complexity is not only defined by the different gene isoforms, but by other factors such as post-translational modifications. Taken from [107]

Because the number of PTMs is high and complex, only the most common examples are described.

Phosphorylation is a reversible mechanism of PTMs. This process is usually performed on serine, threonine and tyrosine residues. It is known that phosphorylations take an important role in the regulation of cellular processes such as cell cycle, growth and apoptosis.

Glycosylations correspond to the incorporation of sugar groups. These groups range from simple monosaccharides to complex branched polysaccharide changes. Among its functions, most of them are related to protein folding, conformation and stability.

Ubiquitin is a 8-kD polypeptide consisting of 76 amino acids. The modification can occur from an initial monoubiquitination event to the formation of a ubiquitin polymer. Ubiquitinated proteins are recognized and can be catalyzed to recycle ubiquitin in further modifications.

S-nitrosilation is a PTM used by cells to stabilize proteins and regulate gene expression. Nitric oxide (NO) is a chemical messenger that reacts with cysteine residues to form S-nitrothiols (SNOs). This modification is a reversible reaction.
1.3 Splicing mechanism

Splicing is one of the post-transcriptional steps that are carried out to the pre mRNA. During it, introns are removed from the nucleotide sequence and the resulting product is a mature mRNA molecule which consists of the 5’ and 3’ untranslated regions (5’/3’ UTRs) and the coding regions or exons, joined together in a single RNA strand.

The splicing process is orchestrated by the spliceosome, a complex machinery made up of different subunits known as small nuclear ribonucleoproteins (snRNPs\(^2\)). There are two unique spliceosomes in the majority of eukaryotes: The U2 dependent spliceosome, which removes U2 type introns, and the U12 dependent which is far less abundant than the former [131]. All the following descriptions are related to the U2 dependent spliceosome.

The spliceosome is composed of five snRNPs: U1, U2, U4, U5 and U6. These subunits are organized and assembled dynamically to create different complexes, both active and inactive, to perform the two catalytic steps in splicing. Each snRNP is composed of small nuclear RNA (snRNA) and specific protein co-factors. In order to form the spliceosome complex, in a first step, snRNPs components must be generated via specific processes.

snRNAs are non coding and non poly-adenylated sequences that carry out their functions in the nucleoplasm [81]. Due to their similarity to protein-coding genes, its transcription must be regulated by specific molecules that serve as signals to trigger the corresponding reactions. Some of these signalling molecules include RNA pol II promoters and nucleotide sequences, analogous to the TATA box [131]. Specifically, spliceosomal snRNAs transcription is initiated by the binding of a factor, known as the snRNA-activating protein complex (SNAPc), and is performed by RNA pol II [81]. As the snRNA is being transcribed, a multi-subunit factor called the Integrator complex [81] recognizes a sequence of nucleotides near the 3’ end to cleave the transcript and release it from RNA pol II.

In order to become active molecules, newly transcribed snRNAs must be transported into the cytoplasm. As RNA pol II is involved in both, transcription of snRNAs and mRNAs, similar properties are shared by both molecules such as the 5’ cap structure, so they should be differentiated between one another. Some of the factors that help to discriminate between both products are: length (mRNA molecules tend to be longer) and interaction with heterogeneous nuclear ribonucleoproteins (hnRNP) [81]. In the cytoplasm, snRNAs create a series of interactions with Sm-type proteins and cofactors to assemble the different snRNP subunits. Once generated, the snRNPs must return to the nucleus to perform their function.

\(^2\)Named snRNP to differentiate them from other RNPs such as the ribosomal subunits
As the molecules are moved back to the nucleus, they are moved through a specialized structure known as Cajal Bodies [81] where different maturation steps are also performed. All the snRNP subunits and other protein cofactors are stored in nucleoplasmic subcompartments called nuclear speckles [65] and when transcription is initiated the molecules are recruited together to carry out spliceosome assembly and splicing.

Specific nucleotide sequences contribute to define where an intron is located. These sequences are: 1) The 5’ splice site (ss) in the form of a GU dinucleotide near the 5’ end of the intron. 2) The 3’ss, a AG dinucleotide close to the 3’ limit of the intron and the adenosine branch point, located 18 to 40 nucleotides upstream of the 3’ss [131]. In between the branch point and the 3’ss a polypyrimidine tract, a repetitive sequence of pyrimidines is also located.

The spliceosome assembly is done at the same time as splicing is performed. First, the U1 subunit recognizes the 5’ss by base pairing of the U1 snRNA to the mRNA creating the E complex (Figure 1.3). The interaction of the U1 subunit and the pre-mRNA molecule is weak and it is stabilized by other factors [81]. Identically, the U2 snRNA recognizes the polypyrimidine tract and the adenosine branch point to enable the interaction between the U2 snRNP with the branch point and create complex A (often referred to as the pre-spliceosome).

As introns tend to be longer than exons, the splicing sites are identified and located by pairs due to the interactions between U1 and U2 subunits, often called the exon definition complex. Note that the exon definition complex consists of the interaction of U1 and U2 snRNPs while the pre-spliceosome is the interaction between U2 snRNA and U1 snRNP.

Furthermore, U1 and U2 snRNPs undergo certain rearrangements that bring close together the 5’ss, the 3’ss and the adenosine branch point. Then, the rest of the spliceosomal subunits (U4, U5 and U6) are brought together as a tri-snRNP complex [131]. At this point, all five subunits are assembled together into complex B, which is a non activated version of the spliceosome. In order to obtain a functional molecule, certain rearrangements must be carried out. The arrangement releases subunits U4 and U1 to form the U5-U2/U6 snRNA structure or activated complex B, that catalyses the first splicing reaction. The intron is cleaved at the 5’ss and the G nucleotide binds, via base pairing, to the A in the adenosine branch point to form a loop-like structure known as lariat. Here, complex C consists of a free exon and a lariat-exon structure.

During the second catalytic reaction, the spliceosomal complex C cleaves the 3’ end of the intron and binds both free exons by a transesterification reaction [131]. The U2,U5 and U6 snRNPs are released for further use in other splicing rounds while the lariat is degraded.
1.3. SPlicing MECHANISm

Figure 1.6: Pre-mRNA splicing process. A) Representation of the mechanism of pre-mRNA splicing. B) Splicing signals in the 3’ and 5’ splice sites and branch point in metazoans and budding yeast. Y = pyrimidine and R = purine. C) Schematic representation of the spliceosome assembly and splicing catalytic steps. Figure taken from [131]. © 2011 Cold Spring Harbor Laboratory Press All rights reserved.
CHAPTER 1. BIOLOGY BEHIND ALTERNATIVE SPLICING

1.4 Alternative splicing

The *Central Dogma* of molecular biology, proposed by Francis Crick in 1958 [32], states the relationship and workflow from DNA to proteins. In this dogma, the relationship between genes and proteins was defined as a one-to-one relation (one gene coded one protein). However, the complete decoding of the human genome showed a significant difference between the number of genes ($\sim 25,000$) and the generated proteins ($> 90,000$) [29]. This finding demonstrated that the central dogma of molecular biology was not completely correct.

Alternative splicing, is the mechanism by which a single pre-mRNA molecule can lead to different functional proteins. In this process, exons can be either included or excluded and thus creating specific gene isoforms or transcripts. The transcriptome is the complete set of gene isoforms that can be expressed for every single gene in the human genome. The phenomenon of alternative splicing was first described by Susan M. Berget in [15], where it was shown that the Adenovirus produced different transcripts during its infectious cycle.

![Figure 1.7: Graphical representation of alternative splicing](image)

**Figure 1.7:** Graphical representation of alternative splicing

1.4.1 Canonical classification of alternative splicing

In Figure 1.8 the canonical classification of alternative splicing events is displayed. These groups can be further divided according to their complexity and presence in eukaryotes. Among the first four types (cassette exon, alternative 3', alternative 5' and intron retention),

---

3Only considering protein coding genes
cassette exons are the most prominent ones, accounting for nearly 40% of splicing events in eukaryotes [7, 117]. In cassette exons, a particular exon is skipped in the mRNA molecule, while alternative 5’ and 3’ describe the recognition of different splice sites in the ends of an exon. Intron retention shows isoforms in which the intron is not spliced out, but retained as part of the mature mRNA.

Figure 1.8: Canonical classification of alternative splicing events. Constitutive exons are depicted in blue and alternatively spliced in orange.

In the group of less common alternative splicing events, mutually exclusive exons show pairs of exons in which only one is maintained in the mRNA, but not both. Both alternative first and last exons show events in which the splicing mechanisms recognizes alternative 5’ capping and poly A tails, respectively. Other types of alternative splicing events can be described, but are often referred to as complex events such as double cassette exons.

1.4.2 Regulation of alternative splicing

Splicing events in eukaryote cells can be divided in two main groups: constitutive splicing events, which are spliced in the same way for every pre-mRNA and alternative splicing events that lead to different gene isoforms. The mechanisms that control and regulate alternative splicing are still subject to active research. This is confirmed by the vast amount of publications [24, 40, 76, 86] related with deciphering the mechanisms that rule splicing regulation.
The most common approach, to describe the regulation of alternative splicing, considers two main elements: *cis-acting* and *trans-acting* elements. The first group is composed of RNA sequences in the pre-mRNA that allow the recognition of specific exonic/intronic regions by the spliceosome. Within this group we can find the canonical splicing signals: the 5’ splice site, the 3’ splice site and the branch point. Other *cis-acting* elements can be identified along the pre-mRNA that serve as target sites for trans-acting elements. This second group encloses different molecules such as proteins and hnRNPs [19] which take an important role in splicing regulation.

### 1.4.2.1 Cis-acting regulatory elements

The effect of canonical splicing signals can be altered due to position, in relation with the regulated exon, the surrounding context and proper recognition by the spliceosome [40]. For instance, weak 3’ splice sites are better recognized when a functional 5’ splice site is located within 100 to 150 nucleotides downstream [40]. For these splicing signals, the most important "regulatory" role is the proper assembly of the spliceosome in those regions. Certain decoy splice sites, located through the pre-mRNA sequence, can form pseudo-exons that hamper the correct spliceosomal assembly. [81].

Other sequence based motifs, that serve as targets for trans-acting elements, can be included as a part of *cis-acting* elements and are often referred as splicing regulatory elements (SREs). For this specific class, the typical classification consists of exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs) and intronic splicing silencers (ISSs). Even though the activity of SREs is context dependent, the general action mechanism is described by recruiting molecules that have positive or negative effects in the steps of splicing reaction [129]. One common example is the polypyrimidine tract-binding (PTB) protein, which can cause exon skipping by recognition of an ESS and avoiding the formation of the exon definition complex [54]. Also, the accessibility of SREs along the RNA structure can difficult the interaction with specific splicing regulating molecules [83].

### 1.4.2.2 Trans-acting regulatory elements

The typical trans-acting elements are usually referred as splicing factors, which mainly group serine/arginine rich proteins (SR) and hnRNPs. SR proteins are mainly characterized as positive splicing regulators, by promoting exon inclusion [40], while hnRNPs seem to have the opposite effect as they avoid the formation of the splicing machinery. However, this
reasoning does not always correct as specific experiments have shown an equal ratio of exon inclusion/exclusion when using either regulator [73, 90, 136]. One possible explanation is due to competitive binding not only between these two molecules, but with other factors that can be located in the cell or in the regulatory complex.

Studies have shown that specific genes encode proteins that can be included as trans-acting elements. Some examples of these splicing regulators are: NOVA [122], RBFOX [138] or QKI [49]. The complete mechanism of action for these regulatory elements is not completely understood. However, with different experimental techniques such as: cross-linking immunoprecipitation (CLIP) or photo activated ribonucleoside-enhanced CLIP (PAR-CLIP) [48] new binding motifs and splicing factors are being actively characterized.

### 1.4.3 Importance of alternative splicing

It has been shown that alternative splicing plays a key role in human biology. The identification of splicing variants has potential uses for diagnosis, prognosis and therapeutic target evaluation in different pathologies, including cancer [61, 140]. Aberrant splicing can be caused by both: the inclusion/exclusion of exons and the misregulation of the splicing mechanism. Both cases lead to the production of aberrant protein isoforms, which may contribute to the pathological condition.

In the following paragraphs, some examples in which alternative splicing can be used to describe pathological conditions or provide the opportunity for biomarker discovery are detailed.

Spinal muscular atrophy (SMA) is a genetic disease that affects the part of the nervous system that controls voluntary muscle movement. In particular, SMA is caused by the loss of the survival of motor neurons 1 gene (SMN1) [75]. The severity of SMA is also linked to the production of SMN protein, which is encoded by SMN2 gene. Both genes (SMN1 and SMN2) are almost identical, but a specific point mutation in exon 7 seems to create an ESS that increases the skipping of the exon [59]. The SMN protein is only active when exon 7 is included so the greater the skipping, the increased severity of the disease among SMA patients [44].

Retinitis pigmentosa (RP) refers to a group of inherited diseases that cause retinal degradation. Patients with RP experience a progressive decline in their vision, because photoreceptor cells in the retina degenerate. In this case, the relation with alternative splicing comes with mutations in specific splicing factors: PRPF3 [22], PRPF8 [82] and PRPF31 [126]. These point mutations affect trans-acting regulatory elements that misregulate the splicing
CHAPTER 1. BIOLOGY BEHIND ALTERNATIVE SPICING

Table 1.1: Genes whose splicing variants are associated to pathological conditions.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Disease</th>
<th>Role</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMN2</td>
<td>SMA</td>
<td>Increased severity due to exon skipping</td>
<td>[44]</td>
</tr>
<tr>
<td>PRPF</td>
<td>RP</td>
<td>Mutation in splicing factor leads to disease</td>
<td>[22, 82, 126]</td>
</tr>
<tr>
<td>SYK</td>
<td>Ovarian cancer</td>
<td>Exon skipping enhances cancer proliferation</td>
<td>[140]</td>
</tr>
<tr>
<td>MST1R</td>
<td>Cancer</td>
<td>Exon skipping enhances cancer proliferation</td>
<td>[135]</td>
</tr>
<tr>
<td>HMGA1A</td>
<td>Alzheimer</td>
<td>Isoform overexpression found in patients</td>
<td>[77]</td>
</tr>
<tr>
<td>CD44</td>
<td>Colorectal cancer</td>
<td>Biomarker for survival</td>
<td>[137]</td>
</tr>
<tr>
<td>MKNK2</td>
<td>Pancreatic cancer</td>
<td>Biomarker for drug resistance</td>
<td>[3]</td>
</tr>
<tr>
<td>ENHA</td>
<td>Lung cancer</td>
<td>Biomarker for survival</td>
<td>[18]</td>
</tr>
<tr>
<td>BRAF</td>
<td>Melanoma</td>
<td>Biomarker for drug resistance</td>
<td>[92]</td>
</tr>
</tbody>
</table>

mechanism and thus, creating aberrant protein isoforms.

Spleen tyrosine kinase (SYK) gene is a known tumor suppressor in breast cancer [30]. An extensive research by Prinos et al [93] showed that SYK expresses two distinct splice isoforms that differ in the presence of exon 9. When the short isoform is produced, apoptosis is induced in ovarian cancer cells, but the long isoform leads to cancer cell proliferation [140].

Besides the potential effect of alternative in the development of pathologies, aberrant splicing variants can also be used as biomarkers and therapeutic targets in cancer [118]. One of the known examples is CD44v6 isoform of CD44 gene. A low expression of this isoform has been related with improved survival of colorectal cancer patients [119].

Table 1.1 shows different genes associated with pathologies in which splicing variants seem to have an implication with either the development of the disease or potential biomarkers for survival and drug resistance.
Through research, it has been shown that the human transcriptome is composed of hundreds of thousands of gene isoforms. In order to obtain relevant measurements, to decipher the concentration and structure of those transcripts, different experimental techniques have been developed such as junction arrays and next generation sequencing. Each of these technologies provide different experimental measurements to quantify gene or exon expression levels. Specific data processing methods must be applied in order to compare the results between or within samples. Different bioinformatic pipelines have been developed to identify alternative splicing events. Some provide a complete method that includes detection and statistical analysis while others require additional toolkits to perform a complete analysis.

2.1 Microarrays

DNA microarrays allow researchers and scientists to measure the expression levels of thousands of genes in one single experiment. The logic behind microarray technology is based on DNA complementary base pairing property. When a single stranded DNA molecule is brought in contact with its complementary sequence, under a process referred to as hybridization, a double stranded molecule is created that can be measured using imaging techniques.

In simple terms, a microarray is a solid substrate (nylon membrane, glass or plastic) on
which single stranded DNA molecules, known as probes, are immobilized in an ordered grid-like structure, hence the name array. Probes are designed to hybridize with cDNA\(^4\) from the biological sample of interest.

The general workflow to perform a microarray experiment consists of: 1) Extraction of mRNA from the cells that are present in the sample under study, 2) conversion of mRNA into cDNA by reverse transcription, 3) fluorescent labelling of the cDNA molecules, 4) immobilization via hybridization and, finally, 5) measurement of the emitted light by phospho-imaging or fluorescence scanning. [106]. The amount of light, emitted by the fluorophores, is expected to be proportional to the concentration of the gene under study. These values give the capability to quantify genes or exons that are present in the sample.

Different microarray systems have been developed, with oligonucleotide microarrays being the most widespread platform among researchers.

### 2.1.1 Oligonucleotide microarrays

Probes used in oligonucleotide microarrays are fabricated directly in the solid substrate using photolithography i.e. the same technique to build integrated circuits. These probes have a length of 25 nucleotides and are packed at high densities (300,000 probes on a 1.28 x 1.28 cm surface [72]). Newer arrays include more than 1 million probes. Probe design should be carefully planned to maximize specificity and avoid cross-hybridization. This allows to discriminate correctly between background and specific signals. Figure 2.1 shows the basic properties of oligonucleotide microarrays.

The mRNA molecule of interest (often related to a gene) is represented by a probe set that includes 11 to 20 probes. By using multiple independent detectors, it is possible to improve the signal-to-noise ratio, reduce the effect caused by cross-hybridization and allows to reduce the number of false positives and miscalls [72].

The main manufacturer of oligonucleotide microarrays is Affymetrix, recently acquired by Thermo Fisher Scientific. Most probe sets are composed, depending on the array, of 4-10 probes called Probe Selection Regions (PSR). Usually, one PSR represents an exon or part of it. The number of PSRs per gene depends on the length of the gene, the number of exons and the number of known isoforms or transcripts. In average, there are around 8-10 PSRs per gene (32-40 probes).

Affymetrix relies on antigenomic probes to quantify background noise. These parti-

\(^4\)Complementary DNA
cular probes, are designed not to hybridize to any particular region of the genome in the corresponding species (human, rat, mouse, etc) and also range across the spectrum of GC content\(^5\), which is known to affect hybridization strength [132].

**Figure 2.1**: Fundamental properties of oligonucleotide microarrays. 1) Complementary base pairing between target sample and probes. 2) Probes are tightly packed and grouped in small dimensions. 3) Fluorescent dyes are added to the target sample before hybridization. 4) Imaging techniques allow the measurement of the emitted fluorescent light.

### 2.1.2 Splicing arrays

Splicing arrays are developed to measure alternative splicing events. In order to obtain such values, not only exon probes are required, but probes that map to known exon junctions are also included. For example, in a cassette exon, the probes that target the skipping junction can be used to determine if the exon is present or not.

---

\(^5\)Percentage of bases that are either guanine or cytosine
Exon probes are generally selected to meet specific criteria. As these probes can be located anywhere in the exon, its specificity can be optimized to ensure an optimal performance during hybridization. In contrast, junction probes have little room for optimization as the sequence of the junction is specific.

The first array that made use of junction probes was based on Agilent technology and included approximately 125,000 junction probes, but lacked exon probes [55]. These were included in a later version of the array [21].

Later on, in 2006, ExonHit introduced the Splicearray, also using the Affymetrix technology (currently they provide a version using Agilent technology). However, the use of Exonhit arrays is not as widespread as the use of standard ones from Agilent or Affymetrix.

In 2008, Affymetrix presented the Human Junction Array (HJAY). It was their first experimental array with exon and junction probes [67]. This microarray included approximately 6 million probes, comprising ~315,000 exons and ~260,000 junctions. Each exon and junction was interrogated by 8 different probes. The probes for this array were selected using RefSeq, ExonWalk and Ensembl annotations.

Two years after, Oryzon Genomics, in collaboration with our group, introduced an array based on Agilent technology, covering 7,958 genes with a total number of 115,318 exon probes and 105,141 junction probes [91]. This new array made use of a massive number of control probes (as much as 20% of the array) to ensure the proper normalization of the measurements.

In 2011, Affymetrix, together with Stanford University, designed the custom Glue Grant Human Transcriptome Array (GG-H array) [134]. And in 2013, Affymetrix launched the GeneChip Human Transcriptome Array 2.0 (HTA 2.0), a more up to date catalogue of the HJAY and GG-H arrays. The HTA 2.0 array interrogates a total of 670,402 exons or exon clusters with more than 6 million probes (approximately 10 probes per exon) and more than 339,000 exon-exon junctions with more than 1 million probes (around 4 probes per junction).

Finally in 2016, Affymetrix deployed the Clariom™ array in two different versions: Clariom™ S (shallow) and Clariom™ D (deep). The Clariom™ array can be considered as the successor of the HTA 2.0. In the Clariom™ S version, more than 20,000 well-annotated genes and 300,000 transcripts are measured by ~211,300 different probes [5], including both exons and junctions. The Clariom™ D provides a highly detailed view of the transcriptome providing the opportunity to analyze alternative splicing events not only in protein coding genes, but also long non-coding ones. This version interrogates more than 130,000
genes and 540,000 transcripts with a catalogue of more than 6.5 million probes. In order to construct such a detailed transcriptome, Affymetrix used more than 15 different sources of annotation. Some of the included databases compromise well annotated genes while others include publication-specific gene sets [4]. Both versions of the Clariom™ array are available for human, rat and mouse species. Table 2.1 shows the main characteristics for the splicing arrays produced by Affymetrix.

<table>
<thead>
<tr>
<th>Table 2.1: Main characteristics of Affymetrix exon-junction microarrays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
</tr>
<tr>
<td>---------------------------------------</td>
</tr>
<tr>
<td>Clariom™ D</td>
</tr>
<tr>
<td>Clariom™ S</td>
</tr>
<tr>
<td>HJAY</td>
</tr>
</tbody>
</table>

2.1.3 Microarray data processing

In order to analyze microarray data, the information must be pre-processed to correct for background signal and other artifacts that can appear during experimental and preparation procedures. Usually, pre-processing is done in three different steps: 1) Background correction, 2) Normalization and 3) Summarization.

Background correction is used to remove noise caused due to print tips, PCR reactions or spatial hybridization effects. The most common method is RMA (Robust Multi-array Average), that uses maximum likelihood to deconvolve the signal of the arrays [112]. This method is based on the assumption that the signal of the probes follows an exponential distribution while noise is normally distributed among the probes. Using maximum likelihood estimation, the parameters of the mixture and the distribution are estimated in order to remove background signal or noise.

The normalization process ensures the minimization of technical variation across different arrays. This allows to compare different arrays or even to use information from experiments that were carried out at different time points in different laboratories. The normalization methods can be divided into three groups: 1) scaling methods, 2) transformation methods and 3) error based models. In scaling methods, intensities are assumed to differ by a constant. Since intensity values are log2 transformed, the scaling factors become additive. For transformation methods, the most common one is Quantile normalization, which equalizes the distribution of expression intensities across arrays [53]. For error based models, the transformation methods correct the variance dependency on intensity by assuming that low values are associated with higher variance (on a log-scale).
Finally, during summarization, the probe intensities are combined into a single expression value per probeset. This provides a single value for each of the targets in the array, i.e. genes, exons, junctions or any other sensible grouping.

The accurate correction of the signal for background noise, hybridization artifacts or other effects is a fundamental step to avoid erroneous quantifications of the target molecules that can turn into false positives in further processing steps. In previous paragraphs, some of the methods used for signal correction are described. However, none of these methods rely on the information given by antigenomic probes. Some methods that use this information have been described\cite{42}, however the complexity of such models increase due to the number of parameters that must be estimated. As a result, no method is considered as the gold-standard for signal correction using antigenomic probes. The accurate correction of the signal can lead to a much more detailed and specific quantification of gene expression levels using microarrays.

2.2 Next generation sequencing

With sequencing techniques, it has been possible to decipher the genetic code of different organisms. From the bacteriophage \(\Phi\)X174, sequenced by Fred Sanger and his team back in 1977 \cite{102}, to the first version of the human genome in the human genome project. Initial approaches were completed using Sanger sequencing techniques. The development of massive parallel sequencers allowed to quantify and sequence larger organisms in a reduced amount of time.

Sanger sequencing is considered as a "first generation" technique, while new high-throughput methods are often referred to as next generation sequencing (NGS).

2.2.1 Sanger sequencing

Sanger sequencing is based on the natural properties of DNA elongation. First, DNA is denatured and amplified to obtain multiple copies of a single stranded molecule with a radioactively labelled primer in the 3’ end. In the next step, four different flasks are prepared by adding DNA polymerase, deoxynucleotides (dNTPs) and dideoxynucleotides (ddNTPs). In each flask, one of the four ddNTPs (ddATP, ddTTP, ddGTP and ddCTP) is added.

dNTPs have an hydroxyl group in the 3’ position while ddNTPs lack the OH group. When DNA elongation is started, dNTPs are added to continue the growth of the nucleotide chain,
however if a ddNTP is incorporated, no new nucleotide can be added and the elongation process is terminated.

This process results in different sized DNA fragments that are terminated by a specific nucleotide, depending on the ddNTP that was incorporated. Using gel electrophoresis the fragments are separated by size. This is performed in four different channels (one per nucleotide), allowing to decipher the corresponding nucleotide in every position of the single stranded DNA molecule.

Figure 2.2 summarizes the Sanger sequencing method.

2.2.2 Second generation sequencers

Next generation sequencing techniques rely on massive parallel sequencing methods to decode the information contained in different molecules such as DNA or RNA. Among the common technologies, the most prominent ones are: RNA-Seq and DNA-Seq. In RNA-Seq, cDNA is used to sequence RNA fragments.

Even though there are many platforms for sequencing technologies, the most widely spread are: Illumina and Ion Torrent. Depending on the platform of choice, specific steps and protocols are used.

RNAseq is used for quantification and sequencing of transcriptomes. In general terms, RNAseq is performed in three steps: 1) library preparation, 2) amplification and 3) sequencing. These steps will be briefly described in the following paragraphs without providing specific details for each platform. For a detailed description refer to [84].

The initial step is library preparation. This process is used to obtain small random fragments of the target molecule (RNA or DNA). This fragments are usually referred to as reads. In order to fragment the RNA (or DNA), different procedures can be used including: sonication, hydrolysis or nebulation. Once the fragments are obtained, specific sequencing adaptors are attached to one or both ends of the fragment. The overall result is a mixture of fragments from the target template. The length of the fragments varies depending on the selected platform for sequencing. In the case of Illumina, reads have a maximum length of 250 bp (HiSeq series) and Ion Torrent uses 400 bp reads.

Amplification consists in creating millions of copies of each of the reads after library preparation. Both Illumina and Ion Torrent use clonal amplification procedures. In general terms, the fragments are immobilized to a solid surface in order to allow multiple sequencing reactions.
CHAPTER 2. BIOINFORMATICS ON ALTERNATIVE SPlicing

Figure 2.2: Sanger sequencing process. 1) DNA elongation is prepared by adding the molecule with unknown sequence, DNA polymerase, dNTPs and the labeled primer. 2) Elongation is performed in four independent flasks and is terminated by the inclusion of ddNTPs. 3) Gel electrophoresis is used to separate fragments due to their size. 4) The obtained sequence is used to deduce the corresponding sequence from the template.

Finally, the sequencing step is performed. In the case of Illumina, the logic behind sequencing reactions is to associate a fluorescent light to the addition of each nucleotide during sequencing. Ion Torrent combines software with integrated circuits and metal-oxide semiconductors (CMOS). The ions that are released by DNA polymerase during DNA synthesis are used to determine the addition of a nucleotide.

Among the different platforms, Illumina is the most extended around researchers. Due
to this, it is important to state specific properties about its library preparation protocols and read types. For RNAseq, two types of reads can be obtained: single-end and paired-end. The first type are composed of a single RNA fragment while the paired-end reads group a pair of reads known as mates, that correspond to both sides of the same RNA fragment. When paired-end protocol is used, the distance between the mates is defined and is called insert size.

### 2.2.3 Third generation sequencers

Second generation methods still rely on standard procedures, such as PCR, to amplify target fragments. This amplifications can lead to sequencing errors or similar problems. In order to avoid such issues, third generation sequencers remove the amplification step and thus, sequence the complete RNA (DNA) fragments obtained during library preparation.

One of the examples of third generation sequencers is Oxford Nanopore. Where the mechanism of action is based on the use of nanopores, either synthetic or natural proteins, which are located along a membrane (graphene). An electric current is applied on both sides of the membrane creating a potential difference that attracts DNA, which is negatively charged. As the molecules move through the nanopores, an electric resistance is created that results in variations in the electric potential between both sides of the membrane. Each specific nucleotide creates a particular variation that leads to the possibility to state which nucleotide is going through the pore. By adding multiple pores in the membrane it is possible to simultaneously sequence different fragments.

In May 2015, Oxford Nanopore deployed the MinION™. This mechanism is a portable sequencing usb device. During intial testing, before the product was commercially available, a rate of 90 nucleotides per second per nanopore with an error rate of 30% was guaranteed [89]. In October 2016, an updated version of the MinION™ was presented with astonishing properties such as a sequencing rate of 450 bases per second per nanopore for 10Gb data per MinION™ Flow Cell [88]. Figure 2.3 depicts the MinION™ workflow.

The device can be connected to a desktop computer and the prepared sample is directly given as an input to the sequencer. This type of technology allows real time sequencing. Oxford Nanopore also provides software solutions for the treatment and storage of sequencing data either in local or cloud drives.

Pacific Biosciences (PacBio) uses the natural properties of DNA polymerase for sequencing. PacBio incorporates phospholinked nucleotides to visualize DNA polymerase activity. Each nucleotide has a fluorescent label incorporated in the terminal phosphate group. When
CHAPTER 2. BIOINFORMATICS ON ALTERNATIVE SPLICING

**Figure 2.3:** The Oxford Nanopore MinION™ is a usb sequencer. By measuring the variation of electrical potential between both sides of a membrane, each nucleotide can be identified.

The incorporation of nucleotides is performed, the label is released and can be detected in a nano-photonic visualization chamber. This chamber is a cylindrical and is approximately 70nm wide that is illuminated through a glass support. As the nucleotides are incorporated, they diffuse in and out of the chamber and when polymerase finds the required nucleotide, the fluorescent label is excited emitting light that is captured by a detector. This process is repeated creating a burst of light that is translated in the corresponding DNA sequence. PacBio allows simultaneous multiplexing of thousands or millions of independent chambers that replicate DNA in real-time.

Other developed approaches include Single Cell Sequencing techniques. For these methods, the DNA from a single cell is sequenced instead of taking the molecules from a
whole sample. The main goal, for this type of technology, is the identification of new cell types in different tissues or the detailed characterization of known cell types. This allows the identification of specific markers that can provide novel means for cell classification. In 2013, Nature publishing group selected this technique as method of the year [1].

Even though single sequencing approaches seem to provide accurate and specific results, different issues arise when dealing with this technique. One of the few problems is the sparsity of the sequenced information: Many of the genes are not present in every cell type thus, many genes appear with few or no reads at all so the statistical assumptions made when dealing with bulk sequencing cannot be used. Also, there is no method to ensure that the gathered DNA comes from one single cell. If two or more cells are sequenced, the information can lead to erroneous results. In order to obtain significant results, hundreds or thousands of cells must be sequenced.

### 2.2.4 NGS data processing

After sequencing, the analysis of the data starts by mapping all the fragments to a reference genome. The common output from a sequencer is a plain text file, given as a *fastq* file which contains the sequence of every read along with quality in a phred-32 or phred-64 scale. Each sample turns into a fastq file, if single-end protocol is used, and two files for paired-end experiments. Specific information from the library preparation protocol, such as the distribution of the length of the fragments or the insert size, might be required for the processing.

Different protocols exist to map the reads against a reference transcriptome. Most of them use a similar approach that consists of mapping the reads against exons and obtaining the possible exon-exon junctions. Some reads might appear as unmapped in a first step. In a second iteration, unmapped reads are used against the pool of possible junctions. The files that can be retrieved after mapping are: Sequence Alignment/Map (SAM), Binary Alignment/Map (BAM) and Browser Extensible Data (BED). The first one contains sequences aligned to a reference annotation, while the second one is a compressed format for SAM files. BED files contain the junctions found during the mapping procedure.

Different software alternatives are available to perform the mapping of the reads. Some are often referred to as spliced aligners, which look for possible splice junctions, while unspliced aligners only map reads against exons. Some examples of aligners are: STAR[37], TopHat[120], SOAP-splice[52], BWA[70], Bowtie [66] and GSNAP[133]. Each of them use specific ways to speed up the analysis or reduce the amount of unmapped reads. For a
CHAPTER 2. BIOINFORMATICS ON ALTERNATIVE SPLICING

thorough evaluation of the available software refer to [71].

Once the mapping is completed and the resulting files are retrieved, different procedures can be used based on the desired analysis: From simple differential expression to transcriptome reconstruction.

The number of reads mapped to a specific target is often referred to as counts. However, certain normalization of the data must be performed in order to compare expression levels between or within samples. One standard normalization procedure is based on scaling to library size.

As longer genes or transcripts are more likely to have sequenced reads, a proper normalization method should be applied to the data before any other type of analysis. Dividing the counts by the product of length and library size results in a scaling normalization method. The result is known as RPKM (Reads Per Kilobase per Million mapped reads). Equation 2.1 shows the normalization unit where \( C \) is the number of reads mapped to a gene, \( N \) is the library size and \( L \) is the length of the gene. The term RPKM is also called FPKM when paired-end reads are used, where the \( R \), that stands for reads, is substituted for an \( F \) denoting fragments.

\[
RPKM = 10^9 \cdot \frac{C}{N \cdot L}
\]  

(2.1)

2.2.5 Visualization of NGS data

BAM files can be loaded into genomic visualizers such as Integrative Genomics Viewer (IGV) or UCSC genome browser. This allows to perform a visual inspection of the results after mapping. In IGV, each BAM file is displayed in three independent tracks: 1) coverage, 2) junctions and 3) reads. Figure 2.5 displays two RNA-Seq samples from different tissues (heart and liver) as displayed in IGV.

A common tool to display RNA-Seq densities along exons and junctions is Sashimi Plot, which is part of Mixture of Isoforms (MISO) software [60]. Sashimi plot can be used to display the number of junctions and RPKMs that correspond to specific regions of interest (exons or junctions). As the corresponding isoforms are also displayed, it is possible to relate the displayed values with specific transcripts. Figure 2.4 displays an example of visualization using Sashimi Plot.
2.3 Algorithms for the detection of alternative splicing

Alternative splicing can be studied from two complementary points of view: with focus on transcripts or splicing events respectively. In the former, the subject of analysis is the transcript (or isoform), whereas in the latter, the subject(s) are the splicing events themselves.

The pipeline of the transcript-focused approach uses RNAseq data with and without known annotations in order to reconstruct the transcriptome and estimate the concentration values of the transcripts. Finally, the significance of change in absolute or relative concentrations is assessed using suitable statistical methods [10, 43, 68]. Transcript reconstruction is challenging [116] and any error in reconstruction of transcript structure may be propagated to the estimation of corresponding concentrations and thereby to the output of statistical analysis. Moreover, the challenge of estimating isoform concentrations for genes with many transcripts yields wide confidence intervals.

In order to circumvent the problem of transcript reconstruction, the transcriptome may be taken as algorithmic input in RNAseq when performing a direct estimation of isoform concentrations [17]. This approach, however, only measures annotated transcripts and is hence unable to detect and quantify novel isoforms. Despite this obstacle, methods to estimate isoform concentrations using microarrays have been proposed [12, 13, 121, 128]. A recent comparison of the isoform deconvolution using both RNA-seq and microarrays has
been published using non-junction Affymetrix HuEx arrays [34].

One potential criticism of transcript-focused methods is that they may miss local alternative splicing variability because of the inherent challenge of isoform deconvolution [51]. Even the better methods display transcriptome reconstruction levels below 50% when using simulated reads, i.e. less than 50% of transcripts are recovered and less than 50% of predictions are correct [116]. On this basis, therefore, an event-based method appears a more suitable approach via which to compare alternative splicing detection technologies, with the additional benefit of straightforward validation using PCR.

Event-based methods focus directly on the analysis of differential splicing events, rather than first attempting to estimate transcript concentration levels. These events can be classified into five canonical categories [16]: cassette exon, alternative 3’, alternative 5’, mutually exclusive exons and intron retention. In some cases, alternative start and termination sites are included also when defining splicing events. This approach has gained traction and several algorithms have been developed recently for detection of splicing events using RNA-seq data, including rMats, SplAdder, spliceGrapher or SGSeq [47, 56, 99, 109]. SpliceGrapher and SGSeq detect events prior to application of separate software in order to state corresponding statistical significance, whereas rMats and SplAdder perform both detection and statistical analysis. Alongside NGS-based approaches, alternative splicing event detection methods are available for exon arrays [141], and exon-junction arrays [38, 100, 108, 110]. The latter methods display validation rates well above 50%.

The Transcriptome Analysis Console (TAC) 3.0 software offered by Affymetrix, AltAnalyze [38] and iGEMS [115] are the available options to analyze HTA 2.0 and HJAY arrays. FIRMA [94], using CDFs generated by Brainarray [33], can be applied to extract and summarize exon expression but the junction probes would be missing from this analysis pipeline.

The main drawback of the TAC and FIRMA approaches, is that neither of them combines the information provided by the junction probes with the corresponding exon probes in the event under study. For example, in the detection of a cassette event, it is not sufficient to detect the altered expression values of exon probes. In addition to that, the flanking junctions must behave coherently and the skipped junction must have a negative correlation. A similar argument can be done for other alternative splicing events. On the other hand, AltAnalyze using the ASPIRE algorithm [123] combines the information of two probesets to get a figure of merit for each event. For example, in a cassette exon, AltAnalyze would provide three figures of merit: one corresponding to the probeset of the exon and the junction that skips it and two more combining the flanking junctions with the junction that skips the
exon. Although, this approach is intrinsically better, it would be still desirable to have a single figure of merit per event. There is an algorithm (MADS+) developed to exploit combined information from exons and junctions [110]. However, its development has been discontinued and it cannot be applied to the HTA 2.0 platform. Furthermore, most of the algorithms developed to detect alternative splicing events (including MADS+) are limited to the analysis of case–control studies. Its extension to more complex experimental designs, such as case–control studies with paired samples or time course studies, is non-trivial.
Figure 2.5: Representation of BAM files for the SLC25A3 gene in two different tissues (heart and liver) as depicted in IGV.
EventPointer is an R package to identify alternative splicing events that involve either simple (case-control experiment) or complex experimental designs such as time course experiments and studies including paired-samples. The algorithm can be used to analyze data from either junction arrays (Affymetrix arrays) or sequencing data (RNA-Seq). The software returns a table with the detected alternative splicing events along with general information such as: gene name, type of event (cassette, alternative 3′,...,etc), genomic position, statistical significance and increment of the percent spliced in (ΔΨ) for all the events. The algorithm can generate a series of files to visualize the detected alternative splicing events in IGV. This eases the interpretation of results and the design of primers for standard PCR validation.

3.1 Overview of the algorithm

The developed algorithm provides users a simplified manner to identify, classify and visualize alternative splicing events using information from microarrays or RNA-Seq. For both platforms, the required steps are almost identical as only the initialization is different for each technology. Also, the input is different due to the nature of the technologies and pre-processing procedures. For junction arrays, signal values must be provided as CEL files and for sequencing, BAM files are given as input.

In a simplified manner, EventPointer can be described in four independent steps: 1)
splicing graph creation, 2) event identification and classification, 3) statistical analysis and 4) visualization. This is graphically represented in Figure 3.1.

In step one, the algorithm builds the corresponding splicing graphs to represent the different isoforms for a given gene. Once the graphs are created, EventPointer analyzes each splicing graph in order to identify and classify the alternative splicing events. With all the events detected, a statistical analysis is performed to obtain the significance of each event. Briefly, EventPointer considers that there is differential splicing if the isoforms in the associated event change their expression in opposite directions. Finally, to ease the interpretation of the splicing events, EventPointer generates GTF (gene transfer format) files that can be loaded into Integrative Genomics Viewer (IGV) [98]. The visualization allows users to design primers to validate the detected events using standard polymerase chain reaction (PCR).

In the case of junction arrays, the steps of splicing graph creation and event identification are performed together to give as an output a chip definition file (CDF). This file groups the probes of the microarray into probesets to provide a single value to each of the elements that define an alternative splicing event. This process allows the user to use the same CDF file for independent experiments, as long as the same microarray is used. For sequencing technologies, the splicing graphs must be generated for every single experiment in order to take into account reads obtained from the sequencer and not only a reference transcriptome.

The developed algorithm uses an event-based approach. This means, that the subject of analysis are the splicing events as opposed to transcript-focused algorithms, were the analysis is based on the transcripts. With this approach, all the need to reconstruct the transcriptome and estimate the concentration values of the transcripts is removed and thus, the result is less prone to false positives.

3.2 Splicing graph creation

The splicing graph [50] is a directed acyclic graph used to represent the structure of a gene. In our model, nodes are the genomic coordinates that define the start and end positions of non overlapping regions in the genome that belong to the same set of transcripts (hereinafter referred to as subexons). There are two types of nodes: $a$ and $b$ nodes, which set the start and end coordinates, respectively, of subexons. All $a$ nodes are exclusively connected to $b$ nodes and vice-versa, giving the graph a bipartite property. This extension allows either probes or reads to be mapped to the edges of the splicing graph. Finally, two additional
3.2. SPlicing GRAPH CREATION

**Figure 3.1:** Main steps for EventPointer pipeline. **A)** Splicing graphs are constructed for each technology. For microarrays, a reference transcriptome must be provided and for sequencing it can be constructed from reads and/or a reference annotation. **B)** Every splicing graph is analyzed to identify and classify alternative splicing events. **C)** The statistical significance for each event is computed from signal values (CEL files) or counts (BAM files) **D)** Visualization files are created to view the events in IGV genomic browser to ease interpretation and primer design for standard PCR validation.
nodes, start and end nodes, are added to the graph and nodes in 5’ or 3’ locus of any isoform are connected to them. Figure 3.2 displays the **DDX11L1** gene, as depicted in Affymetrix HTA 2.0 GTF file and the corresponding splicing graph.

**Figure 3.2:** A) Representation of **DDX11L1** transcripts in Affymetrix HTA 2.0 array GTF file, exons are depicted as blue boxes and dashed red lines denote subexons. B) Regions representing subexons 2 through 5. C) Constructed Splicing graph using EventPointer pipeline

As RNA-Seq allows the discovery of novel events (non annotated in a reference genome), the creation of the splicing graphs is different as for microarrays. In the following sections the specific properties for each technology are described.

### 3.2.1 Junction arrays

For microarrays, the splicing graph is built based on a given transcriptome provided as an input in the form of a GTF file. The user can either provide a custom file or the algorithm will download reference transcriptomes from Ensembl [62] or UCSC known genes [58]
3.3. EVENT IDENTIFICATION AND CLASSIFICATION

The probes of Affymetrix junction arrays are designed using the union of several transcriptomes, that include a range of annotations, e.g. RefSeq, Vega, Ensembl, UCSC known genes and other sources for non-coding isoforms.

Once the splicing graph is created, the algorithm takes the genomic coordinates of the probes and maps them against the different edges of the graph. This creates a weighted graph, in which the weight of the edges correspond to the number of probes mapped to a specific edge.

3.2.2 RNA-Seq

EventPointer for RNA-seq relies on SGSeq [47] to build the corresponding splicing graphs from BAM files. The complexity of the splicing graph can be controlled in SGSeq by setting different thresholds in the expression of splicing junctions in order to be included in the graph (by default set to 2 FPKM). As this parameter is modified, the complexity of the graph changes. A lower expression value increases the complexity of the graph.

SGSeq builds one splicing graph per gene and an external reference transcriptome can be provided to assigning names to the elements of the graph. The resulting graphs contain the corresponding reads that map to the edges of the splicing graph in the form of counts and FPKMs. Any element in the graph, that is not annotated in the transcriptome, is also included as part of the splicing graph. These elements lead to the discovery of novel splicing events.

3.3 Event identification and classification

3.3.1 Identification of alternative splicing events

A splicing event is described as a triplet \( (P_R, P_1, P_2) \) of subgraphs. The fluxes, traversing the splicing graph, are the sum of the concentrations of all the transcripts that share the specific path under study. These subgraphs are composed of sets of edges and nodes which share certain characteristics: 1) the flow traversing any of the elements of each triplet is identical, and 2) the flow traversing any element in \( P_R \) is equal to the sum of the flows traversing \( P_1 \) and \( P_2 \). To be consistent, \( P_1 \) is assigned to be the set of edges with the largest genomic length and \( P_2 \) to the shortest one (i.e. for a cassette exon, \( P_1 \) is the path that includes the exon).

Let \( G(V, E) \) be a directed acyclic graph with \( V = v_1, v_2, \ldots, v_n \) representing \( n \) vertices or
nodes and $E = e_1, e_2, \ldots, e_m$ describing $m$ edges. The incidence matrix $I$ of $G$ is an $n \times m$ matrix $I = (a_{ij})$, where

$$a_{ij} = \begin{cases} -1, & \text{if edge } e_j \text{ leaves node } n_i \\ 1, & \text{if edge } e_j \text{ enters node } n_i \end{cases}$$  \hspace{1cm} (3.1)$$

With the incidence matrix, it is possible to find a flux that is compatible with the graph. For any valid flux in the graph, the sum of the fluxes entering a node is equal to the fluxes leaving the same node (except for the start and end nodes). If the input and output fluxes to the start and end nodes are arbitrarily set to 1, any set of fluxes $v_i$ guarantees that equation 3.2 holds.

$$Iv = b$$  \hspace{1cm} (3.2)$$

Where $I$ denotes the incidence matrix, $v$ is a column vector with the fluxes traversing each of the edges of the graph and $b = [-1, 0, 0, \ldots, 0, 0, 1]^T$ is a valid flux in the graph.

A particular solution for the system depicted in equation 3.2 is defined. The particular solution is set to be of the form:

$$v_p = I^+ b$$  \hspace{1cm} (3.3)$$

Where $I^+$ denotes the Moore-Penrose pseudoinverse of the incidence matrix. The reason to use the Moore-Penrose pseudoinverse is that it provides a least squares solution to the system of linear equations and ensures that every edge in the graph has a flux. This particular solution can be represented as a minimization problem depicted in equation 3.4.

$$\min \sum_{i=1}^{m} v_i^2$$  \hspace{1cm} (3.4)$$

s.t. \hspace{0.5cm} Av = b$$

Once the particular solution is defined, the basis of the left null space of the incidence matrix is obtained by solving the system depicted in equation 3.5. This provides the trivial solution of the homogeneous system of equations described by the incidence matrix.

$$Iv_h = 0$$  \hspace{1cm} (3.5)$$
3.3. EVENT IDENTIFICATION AND CLASSIFICATION

By adding up both the particular and homogeneous solutions, the compatible fluxes traversing each of the edges in the splicing graph are defined.

The triplets are determined as follows. If the flow is identical for several edges, all of them are part of the same set. After the edges are grouped into sets, it is possible to find triplets of sets in which the sum of the flow in two of the triplets $P_1$ and $P_2$ equals the third one $P_R$. The same set can participate in different events as playing the $P_R$ role, but only at most on one as $P_1$ or $P_2$.

3.3.2 Classification of alternative splicing events

Splicing events can be classified into 7 canonical categories: cassette exons, alternative 3’ donor site, alternative 5’ donor site, intron retention, alternative last exon, alternative first exon and mutually exclusive exons. Any event not classified into them is considered to be a complex event. It is possible to label an event by searching for specific patterns in the corresponding graph. Any graph can be represented in matricial form using the adjacency matrix $A$.

The adjacency matrix $A$ for a graph $G$ with $n$ nodes is an $n \times n$ matrix whose $(i, j)$ entry $a_{ij}$ is 1 if the $i^{th}$ node and $j^{th}$ nodes are connected, and 0 if they are not.

$$ a_{ij} = \begin{cases} 1, & \text{if node } n_i \text{ is connected to node } n_j \\ 0, & \text{Otherwise} \end{cases} $$ (3.6)

With matrix $A$ it is possible to identify specific patterns, depicted in figure 3.3.

Cassette exons, alternative 3’, alternative 5’ and intron retentions share the same pattern. In order to differentiate them, the lengths of junctions 1.b - 2.a and 2.b - 3.a are used. For cassette exons, both lengths must be greater than 1; in the case of retained introns both must be equal to 1; and for alternative 3’ (5’) donor sites, junction 1.b - 2.a (2.b - 3.a) must be equal to one. The rest of the events can be identified by looking for the specific graph patterns.
3.4 Statistical analysis

3.4.1 Differential expression

A convenient formalism to describe an experiment is using design and contrast matrices. In order to define the concepts, a simple differential expression model is used. Let us consider a set of samples that are measured in 2 different conditions Normal (N) and Tumoral (T). The corresponding design matrix (D) is shown in equation 3.7.

\[
D = \begin{bmatrix}
    1 & 0 \\
    1 & 1
\end{bmatrix}
\] (3.7)

With matrix \(D\) a linear model, of the form \(y = X\beta + \epsilon\) is defined. Where \(y\) is the vector of data points, \(X\) is the design matrix, \(\beta\) are the coefficients of the model and \(\epsilon\) is an i.i.d\(^6\) error term that follows a \(N(0, \sigma^2)\) distribution. This system can be represented in matricial form as:

\[
\begin{bmatrix}
    y_1 \\
    y_2
\end{bmatrix} = \begin{bmatrix}
    1 & 0 \\
    1 & 1
\end{bmatrix} \begin{bmatrix}
    \beta_0 \\
    \beta_1
\end{bmatrix} + \epsilon
\] (3.8)

\(^6\) independent and identically distributed
And the corresponding equations, defined by the system shown in equation 3.8, can be written as follows (omitting the error term):

\[
\begin{align*}
Y_N &= \hat{\beta}_0 \\
Y_T &= \hat{\beta}_0 + \hat{\beta}_1
\end{align*}
\] (3.9)

Where \( Y_N \) and \( Y_T \) correspond to the expression of a given gene, transcript or exon in both conditions under study.

The contrast matrix \( C \) used to test for differential expression is:

\[
C = [0, 1] \quad (3.10)
\]

Linear models test for the relationship between the observed values (in this case the expression for each condition) and the coefficients in the model. The null hypothesis \((H_0)\) states that \( \beta = 0 \), in case that the null hypothesis does not hold, the alternative hypothesis \((H_1)\) is accepted. By giving an interpretation the the coefficients, it is possible to state a meaning of having a value of \( \beta \neq 0 \).

Returning to our example and equation 3.9, we get that:

\[
\hat{\beta}_1 = Y_T - \hat{\beta}_0 = Y_T - Y_N
\] (3.11)

According to equation 3.11, the coefficient \( \beta_1 \) tests for the difference in expression owing to the tumoral condition when compared to the normal condition. This can be summarized as follows:

\[
\beta_1 = \begin{cases} 
> 0, & \text{Overexpression} \\
= 0, & \text{No change in expression} \\
< 0, & \text{Underexpression}
\end{cases}
\] (3.12)

In relation with alternative splicing, this model would not be valid to test for differential splicing as it only tests for an increase or decrease in the expression levels owing to the conditions under study, but does not show a differential usage of specific isoforms in both conditions.
3.4.2 Differential splicing

In order to test for differential splicing, an extended design matrix is constructed as follows.

Let \( Q \) be the following \( 3 \times 3 \) matrix:

\[
Q = \begin{pmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 1 & 1
\end{pmatrix}
\]  

(3.13)

Using the previously described design matrix \( D \), from equation 3.7. Then, the proposed design matrix \( D^* \) to detect differential splicing events is \( D \otimes Q \) where \( \otimes \) is the Kronecker product, which is equivalent to replacing each of the entries of matrix \( D \) by matrix \( Q \) multiplied by the corresponding entry in \( D \)

\[
D^* = D \otimes Q = \begin{pmatrix}
d_{1,1}Q & \cdots & d_{1,n}Q \\
\vdots & \ddots & \vdots \\
d_{m,1}Q & \cdots & d_{m,n}Q
\end{pmatrix}
\]  

(3.14)

Using the previous example, let us consider a differential splicing event that consists of an alternative last exon. Let us assume that this event is measured in two different conditions, normal (\( N \)) and tumoral (\( T \)). According to EventPointer, there are 6 different measurements: \( Y_{RN}, Y_{1N}, Y_{2N}, Y_{RT}, Y_{1T}, Y_{2T} \). Where \( R, 1 \) and \( 2 \) denote the reference and alternative paths. The proposed extended design matrix \( D^* \) includes an intercept term for all the samples and an increment for the tumoral sample.

Using the auxiliary matrix \( Q \) and the Kronecker product, the splicing design matrix \( D^* \) is obtained.

\[
D^* = D \otimes Q = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}
\]  

(3.15)

With the extended design matrix \( D^* \), the proposed linear model is:
3.4. STATISTICAL ANALYSIS

\[
\begin{pmatrix}
  Y_{RN} \\
  Y_{1N} \\
  Y_{2N} \\
  Y_{RT} \\
  Y_{1T} \\
  Y_{2T}
\end{pmatrix} =
\begin{pmatrix}
  1 & 0 & 0 & 0 & 0 \\
  1 & 1 & 0 & 0 & 0 \\
  1 & 1 & 1 & 0 & 0 \\
  1 & 0 & 0 & 1 & 0 \\
  1 & 1 & 0 & 1 & 1 \\
  1 & 1 & 1 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
  \beta_0 \\
  \beta_1 \\
  \beta_2 \\
  \beta_3 \\
  \beta_4 \\
  \beta_5
\end{pmatrix} + \epsilon
\]

(3.16)

Where \( \epsilon \) is an i.i.d error term following a \( N(0, \sigma^2) \) distribution.

Using some algebra, it is possible to get the value of the coefficients \( \beta_i \) in the proposed model. The corresponding values are displayed in Table 3.1.

**Table 3.1:** Value of the coefficients for the splicing linear model and their interpretation.

<table>
<thead>
<tr>
<th>Value of ( \beta )</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 = \log(a_R) + \log(t_{1N} + t_{2N}) )</td>
<td>No special interest</td>
</tr>
<tr>
<td>( \beta_1 = \log\left( \frac{a_1}{a_r} \right) + \log\left( \frac{t_{1N}}{t_{1N} + t_{2N}} \right) )</td>
<td>No special interest</td>
</tr>
<tr>
<td>( \beta_2 = \log\left( \frac{a_2}{a_1} \right) + \log\left( \frac{t_{2N}}{t_{1N}} \right) )</td>
<td>No special interest</td>
</tr>
<tr>
<td>( \beta_3 = \log\left( \frac{t_{1T}}{t_{1N} + t_{2T}} \right) )</td>
<td>Logarithm of the overall fold change of the event. Differential expression present if ( \neq 0 )</td>
</tr>
<tr>
<td>( \beta_4 = \log\left( \frac{t_{1T}}{t_{1N} + t_{2T}} \right) - \log\left( \frac{t_{1N}}{t_{1N} + t_{2N}} \right) )</td>
<td>Difference of the logarithms of the fold change using relative concentrations of isoform 1 in both conditions. Alternative splicing present if ( \neq 0 )</td>
</tr>
<tr>
<td>( \beta_5 = \log\left( \frac{t_{2T}}{t_{1T}} \right) - \log\left( \frac{t_{2N}}{t_{1N}} \right) )</td>
<td>Difference of the logarithms of the fold change of both isoforms. Alternative splicing present if ( \neq 0 )</td>
</tr>
<tr>
<td>( \beta_4 + \beta_5 = \log\left( \frac{t_{2T}}{t_{1T} + t_{2T}} \right) - \log\left( \frac{t_{2N}}{t_{1N} + t_{2N}} \right) )</td>
<td>Difference of the logarithms of the fold change using relative concentrations of isoform 2 in both conditions. Alternative splicing present if ( \neq 0 )</td>
</tr>
<tr>
<td>( \beta_3 + \beta_4 = \log\left( \frac{t_{1T}}{t_{1N}} \right) )</td>
<td>Logarithm of the fold change of isoform 1</td>
</tr>
<tr>
<td>( \beta_3 + \beta_4 + \beta_5 = \log\left( \frac{t_{2T}}{t_{2N}} \right) )</td>
<td>Logarithm of the fold change of isoform 2</td>
</tr>
</tbody>
</table>

Within Table 3.1 there are some coefficients that involve affinities (or equivalent lengths) and, therefore, have little interest from a biological point of view.

There are a different alternatives to detect alternative splicing using the proposed coefficients. Either of \( \beta_4, \beta_5 \) or \( \beta_4 + \beta_5 \) are theoretically able to detect alternative splicing events. Some of them are more sensitive than others depending on the relative concentrations of the isoforms. For example, if isoforms in path 2 are much more highly expressed than isoforms in path 1 in both conditions, \( \beta_4 \) will be more sensitive than \( \beta_4 + \beta_5 \) since in the latter case,
the numerator and denominator of the logarithms of both therms are similar, and hence their logs are close to zero. A contrast on $\beta_5$ would seem to be more sensitive than a contrast on $\beta_4$ or $\beta_4 + \beta_5$; however, in practice, this contrast proved to be hardly specific and led to many false positives especially in weakly expressed isoforms. If one of the paths is not expressed in any condition, its signal will be similar in either condition and a change in the expression of the other isoform will drive to a false positive detection. This contrast can only be used if the signals are filtered to guarantee that they are above a threshold.

In the PCR validation, the contrast that provided the best performance was the combination of the fold changes of both isoforms plus the requirement that the fold changes have opposite directions, i.e. if isoform 1 significantly increases its expression, isoform 2 must significantly decrease its expression and vice versa. Therefore, this test requires that the detected alternative splicing events show a significant change of the expression both paths and this change must be in opposite direction.

Given the contrast matrix for the experiment (shown in equation 3.10), the splicing contrast matrix $C^*$ to detect splicing events is given by:

$$
C^* = \begin{bmatrix}
C \otimes [110] \\
C \otimes [111]
\end{bmatrix}
$$

(3.17)

Where each row represents the contrasts $\beta_3 + \beta_4$ and $\beta_3 + \beta_4 + \beta_5$. Each contrast is split into two different contrasts (to test the differential expression of both isoforms) and afterwards, they are summarized using the Irwin-Hall distribution.

In order to compute this contrast, the p-values (one-tailed) for both contrasts are summed up. If the null hypothesis ($H_0$) holds, the expected null distribution is triangular from 0 to 2 with the peak at 1, and the summation of the p-values must be close to 0 or close to 2 for genes with differential alternative splicing. Using this triangular distribution, it is possible to assign an overall p-value to their sum. This combination is preferred, rather than the classical Fisher method, since in the latter a single good p-value yields a good summary p-value for the event. Using this approach, both p-values must be close to zero or one in order to generate a significant overall p-value.

The contrast previously described is very sensitive. If one of the paths of an event is not expressed and the other one is, EventPointer would assign a significant p-value to the event. This would result in a large number of false positives due to the lack of expression.

In order to avoid this problem, EventPointer allows the user to filter the events to ensure that all the paths are expressed above a fixed threshold. For every path, the algorithm gets the
maximum value of expression from all the samples. The maximum values for the references are used to set the threshold. The user provides a quantile that will set the threshold.

Once the threshold is selected, an event will be considered as expressed if the maximum value of expression for all the paths is above the threshold previously set.

### 3.4.3 Estimation of the percent spliced index

The Percent Spliced Index (PSI or $\Psi$) is a measure of the inclusion level of an exon or exons involved in a particular splicing event. This proposed model allows to measure the inclusion of all the isoforms mapped to $P_1$ relative to the expression of both paths.

Assuming that the signal of a probeset in microarrays and the number of reads within a region of the transcriptome in RNA-Seq depend on the product of an affinity value of the probeset (or the equivalent length in RNA-Seq) and the concentration of the interrogated isoforms in the paths, equation 3.18 holds.

\[
S_i = a_i t_i + \epsilon \quad (3.18)
\]

where $S_i$ is the measured expression value of path $i$, $a_i$ is the affinity of the probes or equivalent length of the path $i$, $t_i$ is the concentration of the isoforms mapped to path $i$ and $\epsilon$ the error term. The affinity values (or equivalent lengths) and concentration values are unknown and must be estimated from the data.

Particularizing the above equation to each of the paths and taking into account that the concentration of the isoforms in the reference path must be the sum of those of paths 1 and 2, the following equations are obtained:

\[
S_1 = a_1 t_1 + \epsilon_1 \quad (3.19)
\]
\[
S_2 = a_2 t_2 + \epsilon_2 \quad (3.20)
\]
\[
S_R = a_R t_R + \epsilon_R = a_R(t_1 + t_2) + \epsilon_R \quad (3.21)
\]

In turn, the signal value of the reference path can be expressed as the sum of the signal values of paths 1 and 2 as follows,

\[
S_R = a_R a_1^{-1} S_1 + a_R a_2^{-1} S_2 + \zeta = uS_1 + vS_2 + \zeta \quad (3.22)
\]
where \( u \) and \( v \) represent the fraction of the affinities of the mapped probeset (or equivalent lengths) in the reference path and paths 1 or 2 respectively. The values of \( u \) and \( v \) can be estimated from signal data.

Dividing equation 3.19 with equation 3.21 we get,

\[
\frac{S_1}{S_R} = \frac{a_1 t_1}{a_R (t_1 + t_2)}
\]  

(3.23)

Combining equations 3.22 and 3.23, the desired equation of \( \Psi \) used in EventPointer is obtained:

\[
\Psi = \frac{t_1}{t_1 + t_2} = \frac{u S_1}{S_R} = \frac{u S_1}{u S_1 + v S_2}
\]  

(3.24)

Note that \( \Psi \) can be directly obtained from signal values once \( u \) and \( v \) are known. This equation does not require the estimation of the affinities (difficult to predict accurately) to compute \( \Psi \). On the contrary, it simply requires to estimate \( u \) and \( v \) from signal values using equation 3.22. In the case of RNA-Seq, the equivalent lengths are known a priori and hence \( u \) and \( v \). However, using this approach has an advantage: the estimates of these lengths can accommodate the potential lack of uniformity of the reads.

Note that \( u \) and \( v \) must be positive, similar between them and close to one. The first affirmation is trivial since affinity values (or equivalent lengths) are always positive. In microarrays, probesets are composed by several probes and their overall affinity are expected to be similar to each other, since these affinities are a median of the average of the affinities of the probes that build up them. Therefore \( a_1 \approx a_2 \approx a_R \), and \( u \approx v \approx 1 \). A similar reasoning can be applied to RNA-seq, if using coverage instead of read counts, since the coverage of the reference path is expected to be close to the sum of the coverages of paths 1 and 2.

These two fractions can be estimated from equation 3.22 by using non-negative least squares as depicted in equation 3.25.

The penalty factor \( \lambda \) is added to force the equation to fulfill the previous considerations: \( u \) and \( v \) must be similar and close to 1. In our results, we found that the estimates were not too sensitive to the specific value of \( \lambda \) if there is differential alternative splicing. If the relative usage of both paths is similar and therefore, \( \Psi \) is constant, the results are more sensitive to the value of \( \Psi \).
\[ \begin{align*}
\min & \|Ax - b\|_2 \\
\text{s.t} & \quad x \geq 0, x \in \mathbb{R}^2, A \in \mathbb{R}^{m \times n}
\end{align*} \]  
(3.25)

where,

\[
A = \begin{pmatrix}
S_1 & S_2 \\
\lambda & -\lambda \\
\lambda & 0 \\
0 & \lambda
\end{pmatrix}, \quad x = \begin{pmatrix}
u \\
v
\end{pmatrix}, \quad b = \begin{pmatrix}
S_R \\
0 \\
\lambda \\
\lambda
\end{pmatrix}
\]  
(3.26)

### 3.5 Visualization

EventPointer enables visualization of alternative splicing events using the Integrative Genomics Viewer (IGV) [98]. The algorithm generates a GTF file that can be loaded to display the events (the reference and both paths) as well as the location of the probes for each of the paths in the case of microarrays. For RNASeq, novel events can also be displayed, along with a reference transcriptome. An example is shown in figure 3.4.

### 3.6 Installation

EventPointer is available for download at Bioconductor [46] R repository. The algorithm can be installed using the following R-based commands.

```r
source("http:\/\/www.bioconductor.org\biocLite.R")
biocLite("EventPointer")
```
event identified by EventPointer using RNA-Seq.

displayed in different colors. In the lower part, probes are shown and colored accordingly. Each of the alternative paths is

Figure 3.4: (A) IGV visualization of a cassette exon identified by EventPointer using microarrays. Each of the alternative paths is

(B) Representation of a novel splicing event identified by EventPointer using RNA-Seq.
APPLICATION OF EventPointer TO SRSF1

The performance of EventPointer was tested in an experiment where the splicing factor SRSF1 was knocked down using siRNA on the A549 lung adenocarcinoma cell line. The samples were hybridized to the Affymetrix Human Transcriptome (HTA 2.0) microarrays to identify differential splicing events. Top ranked events, predicted by EventPointer, were validated using standard PCR and the results show that the proposed methodology is very reliable as only one false positive was found within the top 200 tested events. A statistical analysis of the enrichment of the protein domains that may be affected by the alternative splicing was also performed. The algorithm was compared with other available software (TAC and AltAnalyze) for the detection of splicing variants.

4.1 Serine/Arginine-rich splicing factor 1

The Serine/Arginine Splicig Factor 1 (SRSF1), formerly known as SF2/ASF, is a member of the serine-arginine (SR) rich protein family of splicing regulators. This protein family has a key role in the regulation of alternative splicing among eukaryotes, as most of the SR proteins are phylogenetically conserved [74]. Currently, 12 human SR proteins are characerized and, most of them, share a structural conformation composed of one or two RNA-recognition motifs (RRMs) and a terminal domain formed by multiple arginine-serine dipeptide repeats [78]. These splicing factors are mostly nuclear, located in the nuclear speckles [35], but six of them (SRSF1, SRSF3, SRSF4, SRSF6, SRSF7 and SRSF10) are known to be dynamically located between the cytoplasm and nucleus [20, 31, 104].
It has been previously characterized [35] that SR proteins are responsible for constitutive pre-mRNA splicing and function as regulators of alternative splicing. Besides their function as splicing regulators, the splicing factor SRSF1 [35] has a pleiotropic effect: regulates nonsense-mediated mRNA decay, has a role on RNA metabolism (translation), RNA protein binding, has a potential oncogenic role in cancer, regulates the mitosis among other processes [9, 36, 80, 101]. Also, SRSF1 has been identified as an oncogene, proving the importance of alternative splicing in tumorigenesis [57].

SRSF1 undergoes different post-translational modifications (PTMs) that take a role in both the localization and functioning of the protein. Most of these PTMs are related with phosphorylation of the serine residues [28]. The arginine residues are also methylated in the proximal region of the nearby RNA recognition motif (RRM) [113]. The transition between methylated and/or phosphorylated states influence in the overall protein.

Most of the functions carried out by SRSF1, are caused due to its RNA-binding potential and interaction with other proteins. This capability of interaction is closely related to its structure. Different cross-linking immunoprecipitation and high-throughput sequencing (CLIP-seq) experiments have shown the preference of SRSF1 to bind to exonic regions in the transcriptome [90, 101].

### 4.2 Experimental design

The performance of the developed algorithm was tested in an experiment where the splicing factor SRSF1 was knocked down using siRNA on the A549 lung adenocarcinoma cell line. This cell line was obtained from the American Type Culture Collection (ATCC). The experiment included three conditions: cells treated only with the vehicle of the transfection (Lipofectamine 2000, Invitrogen), cells treated with scramble siRNA (i.e. a sequence that will not lead to the specific degradation of any cellular mRNA) and cells transfected with a siRNA that targets SRSF1. Each condition had three biological replicates that, in turn, were hybridized three times to Affymetrix Human Transcriptome arrays (HTA 2.0) for a total of 9 hybridizations per condition.

Downregulation of SRSF1, expression analyses and microarray hybridization were done as previously described [36]. Samples from two independent experiments were used for validation of splicing events by endpoint PCR. Briefly, RNA was retro-transcribed using PrimeScript RT reagent Kit (Takara). PCR was performed using PCR Master Mix (Promega) using the following program: 94°C 2 min; 30 cycles at 94°C 30 sec, 57°C 30 sec, 72°C 30 sec;
and 72°C 10 min. The PCR products were loaded in 2% agarose gels and separated for 40-60 min at 100-120V. Pictures were taken in a UV transilluminator and the densitometry of the bands was analyzed using Fiji software [105].

The corresponding design and contrast matrices used for the experiment are shown in equation 4.1. For this particular experiment, the contrast matrix compares the knock down samples using siRNA with the samples of the cells transfected with scramble siRNA.

\[
D = \begin{bmatrix}
C_1 & SCR & SRSF1 \\
C_2 & 1 & 0 & 0 \\
& 1 & 0 & 0 \\
& \vdots & \vdots & \vdots \\
C_9 & 1 & 0 & 0 \\
SCR_1 & 1 & 1 & 0 \\
SCR_2 & 1 & 1 & 0 \\
& \vdots & \vdots & \vdots \\
SCR_9 & 1 & 1 & 0 \\
SRSF1_1 & 1 & 1 & 1 \\
SRSF1_2 & 1 & 1 & 1 \\
& \vdots & \vdots & \vdots \\
SRSF1_9 & 1 & 1 & 1 \\
\end{bmatrix}
\]

\[
C = \begin{bmatrix}
SRSF1 vs SCR \\
Control \\
SCR \\
SRSF1 \\
\end{bmatrix}
\]

(4.1)

In the design matrix depicted in equation 4.1, no distinction was done between technical and biological replicates.

### 4.3 Differential expression at gene level

For the sake of completeness, a differential expression analysis of the experiment was performed. In order to run the analysis, a CDF file was downloaded from Brainarray [33] using gene annotation from Ensembl v.74. Limma [97] was used to perform the analysis and the corresponding design and contrast matrices are the ones depicted in equation 4.1 and previously described in section 4.2.

Table 4.1 shows the top 5 differentially expressed genes. Within this genes, SRSF1 is the 5th ranked and the negative value for the t-statistic confirms the effect of the siRNA for the knock down samples.
Table 4.1: Top 5 differentially expressed genes in SRSF1 knock down experiment

<table>
<thead>
<tr>
<th>Gene</th>
<th>log2 Fold Change</th>
<th>t-statistic</th>
<th>P.value</th>
<th>Adjusted P.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCAPD2</td>
<td>-2.37</td>
<td>-70.5786</td>
<td>1.86e-32</td>
<td>6.58e-28</td>
</tr>
<tr>
<td>E2F7</td>
<td>-1.87</td>
<td>-62.5878</td>
<td>4.86e-31</td>
<td>8.59e-27</td>
</tr>
<tr>
<td>HIATL1</td>
<td>-2.14</td>
<td>-61.3062</td>
<td>8.52e-31</td>
<td>1.00e-26</td>
</tr>
<tr>
<td>CDC20</td>
<td>-1.51</td>
<td>-58.8083</td>
<td>2.63e-30</td>
<td>2.32e-26</td>
</tr>
<tr>
<td>SRSF1</td>
<td>-1.53</td>
<td>-55.8495</td>
<td>1.06e-29</td>
<td>7.54e-26</td>
</tr>
</tbody>
</table>

An expression heatmap is shown in figure 4.1 using the top 15\textsuperscript{th} differentially expressed genes. It can be seen, that samples with siSRSF1 show a decrease, in expression levels, compared with the Lipofectamine and scramble siRNA. Also, the hierarchical clustering of the samples show a similar behavior for each of the different replicates within each experimental condition.

Figure 4.1: Gene expression heatmap for the top 15 differentially expressed genes. Clustering by columns (samples) shows a coherent behavior for the three experimental conditions.

The Affymetrix HTA 2.0 array includes many probes that measure different non-coding
4.3. DIFFERENTIAL EXPRESSION AT GENE LEVEL

Table 4.2: Top 5 differentially expressed long intergenic noncoding RNAs in SRSF1 knock down experiment

<table>
<thead>
<tr>
<th>Gene</th>
<th>log2 Fold Change</th>
<th>t-statistic</th>
<th>P value</th>
<th>Adjusted P.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP11-64D22.5</td>
<td>-2.43</td>
<td>24.0766</td>
<td>2.35e^-21</td>
<td>1.34e^-17</td>
</tr>
<tr>
<td>LINC01504</td>
<td>0.84</td>
<td>22.0767</td>
<td>2.86e^-20</td>
<td>8.19e^-17</td>
</tr>
<tr>
<td>RP11-791G15.2</td>
<td>1.10</td>
<td>21.6443</td>
<td>5.04e^-20</td>
<td>9.62e^-17</td>
</tr>
<tr>
<td>RP11-221N13.3</td>
<td>1.13</td>
<td>20.6177</td>
<td>2.01e^-19</td>
<td>2.88e^-15</td>
</tr>
<tr>
<td>RP11-215G15.5</td>
<td>-1.56</td>
<td>-19.2231</td>
<td>1.46e^-18</td>
<td>1.67e^-15</td>
</tr>
</tbody>
</table>

genes (i.e micro-RNAs, lncRNAs, pseudogenes or lincRNAs). This additional information can be used to check for the expression levels, within these gene subtypes, and analyze differentially expressed non-coding genes. Similar to previous examples, table 4.2 displays the top 5 differentially expressed lincRNAs and figure 4.2 displays the corresponding expression heatmap.

Finally, an enrichment analysis for gene ontology was performed using topGO R package.

Figure 4.2: Gene expression heatmap for the top 15 differentially expressed lincRNAs.
One of the advantages of the topGO package is that avoids providing redundant information. For example, if a function is strongly enriched, also its parent in the ontology will probably be enriched. By using proper pruning, only the most significant leaves of the tree are provided. In some cases, a function expected to be affected does not appear simply because one of its descendant (and therefore more specific) functions are included in the output. The "classical" analysis can be performed by using a different algorithm that does not prune the significant functions.

The analysis was divided in two independent procedures: one using overexpressed genes \((t > 0)\) and other using subexpressed ones \((t < 0)\). The selected threshold for choosing differentially expressed genes was set at \(t > 3\) and \(t < -3\). Table 4.3 depicts the obtained results.

**Table 4.3:** Enriched gene ontology terms for the SRSF1 experiment.

<table>
<thead>
<tr>
<th>GO ID</th>
<th>Term</th>
<th>Annotated</th>
<th>Significant</th>
<th>Expected</th>
<th>P.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006614</td>
<td>SRP-dependent cotranslational protein</td>
<td>105</td>
<td>67</td>
<td>28.34</td>
<td>2.0^{-19}</td>
</tr>
<tr>
<td>GO:0006414</td>
<td>Translational elongation</td>
<td>198</td>
<td>77</td>
<td>44.95</td>
<td>2.8^{-17}</td>
</tr>
<tr>
<td>GO:0016259</td>
<td>Selenocysteine metabolic process</td>
<td>87</td>
<td>57</td>
<td>19.75</td>
<td>4.5^{-17}</td>
</tr>
<tr>
<td>GO:0000184</td>
<td>Nuclear-transcribed mRNA catabolic process</td>
<td>116</td>
<td>66</td>
<td>26.34</td>
<td>3.2^{-15}</td>
</tr>
<tr>
<td>GO:0006415</td>
<td>Translation termination</td>
<td>171</td>
<td>65</td>
<td>38.82</td>
<td>5.8^{-15}</td>
</tr>
<tr>
<td>GO:0006413</td>
<td>Translation initiation</td>
<td>256</td>
<td>96</td>
<td>58.12</td>
<td>2.3^{-13}</td>
</tr>
<tr>
<td>GO:0019083</td>
<td>Viral transcription</td>
<td>187</td>
<td>79</td>
<td>42.45</td>
<td>2.9^{-10}</td>
</tr>
<tr>
<td>GO:0018279</td>
<td>Protein N-linked glycosylation</td>
<td>245</td>
<td>98</td>
<td>55.62</td>
<td>7.6^{-10}</td>
</tr>
<tr>
<td>GO:0035666</td>
<td>TRIF-dependent toll-like receptor</td>
<td>81</td>
<td>38</td>
<td>18.39</td>
<td>1.3^{-06}</td>
</tr>
<tr>
<td>GO:0045454</td>
<td>Cell redox homeostasis</td>
<td>87</td>
<td>40</td>
<td>19.75</td>
<td>1.3^{-06}</td>
</tr>
</tbody>
</table>

**Underexpressed Genes**

<table>
<thead>
<tr>
<th>GO ID</th>
<th>Term</th>
<th>Annotated</th>
<th>Significant</th>
<th>Expected</th>
<th>P.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0051301</td>
<td>Cell division</td>
<td>628</td>
<td>282</td>
<td>144.50</td>
<td>1.6^{-27}</td>
</tr>
<tr>
<td>GO:0007067</td>
<td>Mitotic nuclear division</td>
<td>414</td>
<td>223</td>
<td>95.26</td>
<td>1.9^{-19}</td>
</tr>
<tr>
<td>GO:0000724</td>
<td>Double-strand break repair</td>
<td>145</td>
<td>94</td>
<td>33.36</td>
<td>2.5^{-17}</td>
</tr>
<tr>
<td>GO:0031047</td>
<td>Gene silencing by RNA</td>
<td>141</td>
<td>72</td>
<td>32.44</td>
<td>4.5^{-16}</td>
</tr>
<tr>
<td>GO:0000086</td>
<td>G2/M transition of cell cycle</td>
<td>188</td>
<td>95</td>
<td>43.26</td>
<td>9.4^{-16}</td>
</tr>
<tr>
<td>GO:0000082</td>
<td>G1/S transition of cell cycle</td>
<td>242</td>
<td>123</td>
<td>55.68</td>
<td>4.9^{-15}</td>
</tr>
<tr>
<td>GO:0000398</td>
<td>mRNA splicing</td>
<td>273</td>
<td>136</td>
<td>62.82</td>
<td>7.5^{-15}</td>
</tr>
<tr>
<td>GO:0006271</td>
<td>DNA strand elongation</td>
<td>38</td>
<td>34</td>
<td>8.74</td>
<td>2.1^{-14}</td>
</tr>
<tr>
<td>GO:0034080</td>
<td>CENP-A containing nucleosome assembly</td>
<td>38</td>
<td>31</td>
<td>8.74</td>
<td>3.3^{-14}</td>
</tr>
<tr>
<td>GO:0016925</td>
<td>Protein sumoylation</td>
<td>123</td>
<td>66</td>
<td>28.30</td>
<td>1.9^{-13}</td>
</tr>
</tbody>
</table>
4.4 Determining differential splicing events

The identification of alternative splicing events was done using the Ensembl transcriptome (Ensembl v.74) [39]. An alternative splicing event is considered to be differentially spliced if the concentration of the isoforms mapped to either paths of the event are differentially expressed in opposite directions.

For every experiment, the output of hybridization (i.e. CEL files returned by the scanner) must be summarized using any standard pipeline (in our case, RMA). Following probeset summarization (according to the previously prepared CDF file), EventPointer uses the provided design and contrast matrices to compute the statistical significance of alternative splicing events. By construction, the alternative splicing events can always be validated with at most three primers.

A total of 70,886 events were identified and, out of those, 3,718 showed a p-value < $10^{-3}$ (about 5% of the events). Figure 4.3 shows a barplot with the number of detected events for the different types of events according to the classification provided by EventPointer.

![Barplot with the number of detected alternative splicing events by type of event](image)

Figure 4.3: Barplot with the number of detected alternative splicing events by type of event

The majority of detected event types correspond to complex events (i.e. events that cannot be included in any of the standard categories). As the transcriptome annotation improves, it also becomes more complex and as a result, it is more challenging to provide a standard definition to classify correctly every detected alternative splicing event.

Standard PCR was used to validate the five top-ranked splicing events within each of the eight different types of alternative splicing events (i.e. cassette, mutually exclusive, complex, etc). In total, 40 different events were tested. Each event was validated on two different samples. The events were ranked according to their p-values. In turn, the p-value represents the "differential opposite expression" of each of the isoforms interrogated by an event. The
statistical significance of the events (and their ranking) was very different across the different alternative splicing types. All the five top ranked cassette and complex events were within the overall top 20 ranked events. In contrast, only one out of the five top ranked alternative 3' was found within the top 150 ranked events and the top ranked mutually exclusive exons was in position 500. The complete list of tested events is shown in table 4.4.

The 17 top ranked events appeared within the top 5 of any of the categories and were validated. In all of them the validation was positive although, in a few cases, the PCR-band analysis (using ImageJ \[2\]) did not pass statistical significance. Figure 4.4 shows the PCR results for all the events.

The validation rate is quite high. The reasons of this are on one hand EventPointer gets single statistics that combines information within all the probesets interrogating the event and, on the other hand, uses a proper reference region for each event: most methods select a number of probesets (all of them, only the constitutive exons, core or quasi-constitutive exons, etc.) that are used for all the events in the gene. However, this reference is different for each event.

### 4.5 Protein domains

Each of the paths\(^7\) in an event is annotated with the protein domains (if any) included in the Ensembl database. In some cases, the domain is disrupted in one of the paths compared with the other. A statistical analysis of the enrichment of the domains that may affected by the alternative splicing events was performed.

Using the Ensembl database is possible to relate each of the paths \(P_1\) and \(P_2\) in the events with the presence of protein domains within them. For each domain, it is possible to know in which paths \(P_1\) and \(P_2\) is included. In order to state the statistical significance of the events, a Wilcoxon paired test was performed for each event (\(P_1\) and \(P_2\)). The results of this analysis are summarized in table 4.5

The enrichment analysis illustrates one of the potentials of this analysis. The laminins are proteins of the extracellular matrix. The modification of their domains is known to induce a pro-invasive phenotype \[103\].

Several of the motifs are related to RNA binding: RNA recognition motif (a.k.a. RRM, RBD, or RNP domain). These results are coherent with the known functioning of SR proteins, as their structural conformation and RRMs permit a wide interaction with other molecules.

\(^7\)The word path is used here to refer to any of the two configurations of each event
Table 4.4: The first column shows the ranking of the event in EventPointer according to the p-value. The last column states if the validation was positive: a (√) mark is shown if the validation is significant in PCR, a (★) mark is shown if differences in the PCR can be observed but the statistical significance is above 0.05. A (×) is shown if the event was not validated either because there was no differential splicing, no alternative splicing or no expression at all.

<table>
<thead>
<tr>
<th>Ranking</th>
<th>HGNC Symbol</th>
<th>Type</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MYCBP2</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>KIF23</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td>3</td>
<td>AC024560.3</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>FBXO22</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td>5</td>
<td>SRSF3</td>
<td>Complex Event</td>
<td>★</td>
</tr>
<tr>
<td>6</td>
<td>SUPT16H</td>
<td>Alternative Last Exon</td>
<td>✓</td>
</tr>
<tr>
<td>7</td>
<td>HMBOX1</td>
<td>Alternative First Exon</td>
<td>✓</td>
</tr>
<tr>
<td>8</td>
<td>ACAD11</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td>9</td>
<td>NCO1</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td>10</td>
<td>AUP1</td>
<td>Retained Intron</td>
<td>✓</td>
</tr>
<tr>
<td>11</td>
<td>IFT27</td>
<td>Alternative Last Exon</td>
<td>✓</td>
</tr>
<tr>
<td>12</td>
<td>GALNT10</td>
<td>Alternative Last Exon</td>
<td>★</td>
</tr>
<tr>
<td>13</td>
<td>PARD3</td>
<td>Alternative Last Exon</td>
<td>★</td>
</tr>
<tr>
<td>14</td>
<td>PRMT2</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td>15</td>
<td>HORMAD1</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td>16</td>
<td>ANAPC7</td>
<td>Alternative Last Exon</td>
<td>✓</td>
</tr>
<tr>
<td>17</td>
<td>OGT</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td>23</td>
<td>MSL3</td>
<td>Alternative First Exon</td>
<td>✓</td>
</tr>
<tr>
<td>26</td>
<td>NT5C</td>
<td>Alternative 5’</td>
<td>✓</td>
</tr>
<tr>
<td>29</td>
<td>ALG2</td>
<td>Alternative First Exon</td>
<td>✓</td>
</tr>
<tr>
<td>33</td>
<td>MSL3</td>
<td>Alternative First Exon</td>
<td>✓</td>
</tr>
<tr>
<td>39</td>
<td>BAAP2L1</td>
<td>Alternative First Exon</td>
<td>✓</td>
</tr>
<tr>
<td>40</td>
<td>HIST1H2AC</td>
<td>Alternative 5’</td>
<td>★</td>
</tr>
<tr>
<td>44</td>
<td>DDX52</td>
<td>Retained Intron</td>
<td>★</td>
</tr>
<tr>
<td>71</td>
<td>TMEM214</td>
<td>Retained Intron</td>
<td>★</td>
</tr>
<tr>
<td>74</td>
<td>GABPB1</td>
<td>Alternative 5’</td>
<td>✓</td>
</tr>
<tr>
<td>77</td>
<td>EIF3B</td>
<td>Retained Intron</td>
<td>★</td>
</tr>
<tr>
<td>101</td>
<td>LAMP1</td>
<td>Retained Intron</td>
<td>★</td>
</tr>
<tr>
<td>113</td>
<td>LMO7</td>
<td>Alternative 3’</td>
<td>✓</td>
</tr>
<tr>
<td>149</td>
<td>SCAMP3</td>
<td>Alternative 5’</td>
<td>×</td>
</tr>
<tr>
<td>226</td>
<td>UHRF2</td>
<td>Alternative 5’</td>
<td>×</td>
</tr>
<tr>
<td>355</td>
<td>COPS3</td>
<td>Alternative 3’</td>
<td>×</td>
</tr>
<tr>
<td>464</td>
<td>SLSC9A8</td>
<td>Alternative 3’</td>
<td>★</td>
</tr>
<tr>
<td>473</td>
<td>FLNA</td>
<td>Alternative 3’</td>
<td>×</td>
</tr>
<tr>
<td>480</td>
<td>C21orf58</td>
<td>Alternative 3’</td>
<td>✓</td>
</tr>
<tr>
<td>500</td>
<td>CALU</td>
<td>Mutually Exclusive</td>
<td>×</td>
</tr>
<tr>
<td>688</td>
<td>CCT6P1</td>
<td>Mutually Exclusive</td>
<td>★</td>
</tr>
<tr>
<td>844</td>
<td>ST20</td>
<td>Mutually Exclusive</td>
<td>✓</td>
</tr>
<tr>
<td>1388</td>
<td>ACO2</td>
<td>Mutually Exclusive</td>
<td>×</td>
</tr>
<tr>
<td>1956</td>
<td>KIAA0100</td>
<td>Mutually Exclusive</td>
<td>×</td>
</tr>
</tbody>
</table>
CHAPTER 4. APPLICATION OF EVENTPOINTER TO SRSF1

Figure 4.4: PCR bands and relative concentrations of the isoforms based on the PCR image. The red bar corresponds to the shorter isoform and the blue bar to the longer isoform. For the event in KIAA0100, it was not possible to get PCR results, and thus the image is not included.
4.6 COMPARISON WITH OTHER SOFTWARE

Table 4.5: Enrichment of domains in the list of splicing events regulated by SRSF1. The domains are sorted by its statistical significance. The upper part of the table shows overexpressed domains. The second part of the table (divided by a thick line) shows underexpressed domains. The statistical significance was larger for overexpressed than for underexpressed domains.

<table>
<thead>
<tr>
<th>Domains (PFAM)</th>
<th>P-value</th>
<th>Z-value</th>
<th>Description</th>
<th>Characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF00052 PF06908</td>
<td>4.15e-15, 4.81e-16</td>
<td>7.56,5.46</td>
<td>Laminin B, Laminin Domain I and II</td>
<td>Laminin family</td>
</tr>
<tr>
<td>PF08373 PF00076</td>
<td>4.61e-14, 1.11e-14</td>
<td>6.23,5.31</td>
<td>RAP domain, RNA recognition motif</td>
<td>RNA binding</td>
</tr>
<tr>
<td>PF07040</td>
<td>1.19e-09</td>
<td>6.08</td>
<td>WD Domain, G-beta repeat</td>
<td>WD domain</td>
</tr>
<tr>
<td>PF02780 PF02779</td>
<td>2.30e-09, 2.05e-08</td>
<td>5.97,4.75</td>
<td>Transketolase, C-terminal domain</td>
<td>Transketolase</td>
</tr>
<tr>
<td>PF04386 PF050804</td>
<td>2.38e-09, 1.66e-09</td>
<td>5.67,5.94</td>
<td>Interferon-related protein conserver region</td>
<td>Interferon related</td>
</tr>
<tr>
<td>PF00038 PF04732</td>
<td>5.97e-08, 9.48e-08</td>
<td>5.42,5.34</td>
<td>Intermediate filament protein and head region</td>
<td>Intermediate filament</td>
</tr>
<tr>
<td>PF00038</td>
<td>9.48e-08</td>
<td>5.34</td>
<td>GRAM domain</td>
<td>GRAM domain</td>
</tr>
<tr>
<td>PF08373</td>
<td>1.35e-08</td>
<td>5.27</td>
<td>Plexin cytoplasmic RasGAP domain</td>
<td>Plexin</td>
</tr>
<tr>
<td>PF02996</td>
<td>2.18e-07</td>
<td>3.18</td>
<td>60Kd inner membrane protein</td>
<td>Cytochrome oxidase</td>
</tr>
<tr>
<td>PF00152</td>
<td>6.05e-07</td>
<td>4.99</td>
<td>tRNA synthetases class II</td>
<td>tRNA synthetases</td>
</tr>
<tr>
<td>PF00884</td>
<td>3.38e-06</td>
<td>4.65</td>
<td>Sulfatase</td>
<td>Sulfatase</td>
</tr>
<tr>
<td>PF01485</td>
<td>3.63e-06</td>
<td>4.63</td>
<td>IBR domain</td>
<td>Protein Quality Control</td>
</tr>
<tr>
<td>PF08725</td>
<td>2.95e-04</td>
<td>-3.62</td>
<td>Integrin beta cytoplasmic domain</td>
<td>Cell-cell receptors</td>
</tr>
</tbody>
</table>

As of current knowledge, the relation between SRSF1 and the WD40 domain was unknown. The underlying common function of all WD40-repeat proteins is the coordination of multi-protein complex assemblies, where the repeating units serve as a rigid scaffold for protein interactions [114]. In addition to this, the RNA domains are also targets of PRPF8, another splicing factor. IBR, as well as WD40, domains are related to ubiquitin ligase complexes [79].

A few of the domains that were underexpressed after the knock-down of SRSF1 are included. It is important to point out that the statistical significance is much smaller. An intriguing result is that the tRNA synthetases domains are overexpressed and underexpressed depending on their class.

4.6 Comparison with other software

Affymetrix Transcriptome Analysis Console (TAC 3.0) and AltAnalyze are examples of available software to detect alternative splicing events using HTA 2.0. The main features from each of them are briefly explained below and the comparisons with EventPointer is discussed in the latter paragraphs.

4.6.1 TAC 3.0

The software from Affymetrix is publicly available for download and provides the user with different tools to go beyond simple differential expression analysis. Some of the options are...
gene pathway networks, miRNA and target gene interactions and alternative splicing events identification. It works only in Windows operating system.

Given the corresponding CEL files, the software automatically runs the analysis based on the options provided by the user.

TAC uses Splicing Index (SI) \([45]\) as a measure to detect alternative splicing events. Briefly, the SI of a probeset compares two ratios:

\[
SI = \frac{\text{Probeset } n \text{ signal condition 1}}{\text{Overall signal of the gene in condition 1}} \left( \frac{\text{Probeset } n \text{ signal condition 2}}{\text{Overall signal of the gene in condition 2}} \right)
\]  

(4.2)

If the SI is close to one, then the behavior of the probeset is coherent with the behavior of the gene. If it much larger or much smaller than 1, then the probeset signal may indicate the presence of alternative splicing.

TAC applies several filters based on expression levels (adjustable by the user) prior to calculation of the splicing index for any given PSR (probe selection region, i.e. subexon) or junction. It also classifies (some) of the events according to the standard categories. And, for the ones that are classified, TAC includes a splicing score. This value is based on "how well the data fits into pre-defined splicing patterns" (from the TAC manual). Besides, both "PSRs and their related junctions all contribute to an events score" that is bounded between 0 and 1 (also from the TAC manual). No additional information is provided on how this score is computed or how the events are classified and can be considered and experimental method. Only around half of the events are given a splicing score. In our case, 23/40 (57\%) of the validated events included the score.

All the events found by TAC were sorted according to the absolute value of the logarithm of the SI (negative values indicate the lack of the exon in the case samples).

### 4.6.2 AltAnalyze

This is an open-source software developed in the Nathan Salomonis lab at Cincinnati Children's Hospital Medical Center and the University of Cincinnati. This project began in the laboratory of Bruce Conklin at the Gladstone Institute. It can be downloaded from their webpage (http://www.altanalyze.org/) and it can be run in different operating systems such as Windows, Mac OS and Ubuntu (Linux distribution). As stated by the developers: "requires no advanced knowledge of bioinformatics programs or scripting".
The software enables analysis of data produced by both conventional and splicing sensitive microarrays (e.g. exon and junction arrays) as well as RNA-Seq data and the pipeline enables a complete analysis that includes identification of alternative splicing events and differential expression as well as different functional annotations of the genomic regions identified as alternatively spliced.

For the detection of alternative splicing events, AltAnalyze uses two different algorithms to measure splicing events: Splicing Index (as in TAC) and analysis of splicing by isoform reprocity (ASPIRE). AltAnalyze provides the user the option to set different thresholds to filter genes and alternative splicing events depending on the expression levels. The ASPIRE algorithm is used when two probesets (A and B) measure the exclusion and inclusion of an exon respectively. It provides a score similar to a fold change, bounded between -1 and 1, where negative values indicate that the expression in the probeset (A or B) in experimental group is higher than the control group. A single splicing event can have (and usually does) several inclusion indexes per event. Each of them correspond to the pairwise comparisons between the probesets that include and exclude the event respectively.

In order to identify an event as statistically significant, AltAnalyze relies on three different values: ratio of inclusion, ratio of exclusion and $\delta I$. The first ratios measure the proportion of the inclusion or exclusion of an exon with the mean expression of the whole gene. Both ratios must be in opposite directions to continue the analysis. The $\delta I$ value measures the difference between both the inclusion and exclusion ratios. As a default value, any event must have a $\delta I$ above 0.2. A detailed explanation of the algorithm can be found in the AltAnalyze user manual.

EventPointer takes into account all the exons and junctions involved in the alternative splicing event to give its statistical significance. The results for EventPointer are normalized using the probeset in the reference path and using the whole gene for AltAnalyze and TAC. On the other side, TAC (using SI) and AltAnalyze (using ASPIRE) provide up to 3 statistical values for each event (skipping junction vs flanking junction 1, skipping junction vs exon and skipping junctions vs flanking junction 2). Both AltAnalyze and TAC are focused on case-control studies. EventPointer can be applied to any experimental design that can be described with a design and a contrast matrix.

### 4.6.3 Comparison of top ranked events

Using the "GRanges" R package [69], the events expressed and detected by the three algorithms (EventPointer, AltAnalyze and TAC) were matched. As explained before, an event
detected by EventPointer can be matched to more than one element of either TAC or Analyze, as a result only the unique matched events were used in order to create the Venn diagram depicted in figure 4.5. This diagram shows the expressed events for all of them. These events do not necessarily show differential usage of the variants.

Figure 4.5: Venn diagram of common events identified by EventPointer, AltAnalyze and TAC. The diagram displays the total number of events detected by each algorithm (in parenthesis) and the corresponding values for intersections.

As already mentioned, TAC provides a different SI for each of the probesets that interrogate an event. These values were summarized taking the most significant SI for each event. Table 4.6 provides the ranking of the top ten events detected by EventPointer in both AltAnalyze and TAC.

Table 4.6: Ranking of the top ranked events according to EventPointer and their ranking positions in other algorithms. (EP: EventPointer, AA: AltAnalyze)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Event Type</th>
<th>Genomic Position</th>
<th>EP</th>
<th>AA</th>
<th>TAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYCBP2</td>
<td>Complex Event</td>
<td>13:77673020-77695630</td>
<td>1</td>
<td>12</td>
<td>383</td>
</tr>
<tr>
<td>KIF23</td>
<td>Cassette Exon</td>
<td>15:69713986-69714774</td>
<td>2</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>AC024560.3</td>
<td>Cassette Exon</td>
<td>3:197348575-197350253</td>
<td>3</td>
<td>9</td>
<td>113</td>
</tr>
<tr>
<td>FBXO22</td>
<td>Cassette Exon</td>
<td>15:76196323-76205608</td>
<td>4</td>
<td>55</td>
<td>1392</td>
</tr>
<tr>
<td>SRSF3</td>
<td>Complex Event</td>
<td>6:36566626-36568967</td>
<td>5</td>
<td>2095</td>
<td>280</td>
</tr>
<tr>
<td>SUPT16H</td>
<td>Alternative Last Exon</td>
<td>14:21837979-21852105</td>
<td>6</td>
<td>NA</td>
<td>1026</td>
</tr>
<tr>
<td>HMBOX1</td>
<td>Alternative First Exon</td>
<td>8:28902878-28904970</td>
<td>7</td>
<td>6</td>
<td>98</td>
</tr>
<tr>
<td>ACAD11</td>
<td>Complex Event</td>
<td>3:132297677-132298402</td>
<td>8</td>
<td>396</td>
<td>3371</td>
</tr>
<tr>
<td>NCO1</td>
<td>Cassette Exon</td>
<td>17:16052765-16055312</td>
<td>9</td>
<td>365</td>
<td>5757</td>
</tr>
<tr>
<td>AUP1</td>
<td>Retained Intron</td>
<td>2:74754863-74755133</td>
<td>10</td>
<td>NA</td>
<td>10501</td>
</tr>
</tbody>
</table>
The events can be clustered into three groups: events with low ranking in the three algorithms, events with high ranking in EventPointer and low ranking in AltAnalyze and TAC and, finally, events low ranked to EventPointer and high ranked in the other two algorithms.

Within the first group, \textit{KIF23} was ranked 2\textsuperscript{nd}, 11\textsuperscript{th} and 3\textsuperscript{rd} in EventPointer, AltAnalyze and TAC, respectively. The corresponding statistical values in each algorithm are: $1.00e^{-58}$ (EventPointer pvalue), $0.5523$ (ASPIRE $\delta I$) and $16.67$ (TAC SI). The three methods identify this event as low ranked.

\textit{HMBOX1} is also an event identified as high-ranked by the three algorithms. The corresponding rankings are 7\textsuperscript{th}, 6\textsuperscript{th} and 98\textsuperscript{th} in EventPointer, AltAnalyze and TAC, respectively. Even though the ranking in TAC is not as low as in the other two methods, the PCR validation confirms the alternative splicing event.

In the second group, \textit{NCOR1} was ranked 9\textsuperscript{th} in EventPointer while AltAnalyze and TAC rank the event in positions 365 and 5757, respectively. This event shows either the highest SI or $\delta I$ value when compared to the other events found for the same gene. The event was validated by PCR and shows a significant pvalue ($5.69e^{-38}$) in EventPointer.

\textit{ACAD11} is ranked as the 8\textsuperscript{th} event in EventPointer while AltAnalyze and TAC rank the event in positions 396 and 3371, respectively. It shows a similar behavior as \textit{NCOR1}.

\textit{NAA50} was ranked 10926\textsuperscript{th}, 6\textsuperscript{th} and 148\textsuperscript{th} in EventPointer, TAC and AltAnalyze, respectively. The reason is that this event is backed up by only one junction probeset. Since EventPointer finds no coherence with the other paths of the event, the ranking is low. This event was validated using PCR. The PCR results are shown in figure 4.6. For this event, the ranking of the event using TAC 3.0 and EventPointer are very different. The reason is the difference on the underlying statistical tests: EventPointer imposes that both isoforms (i.e. both bands in the PCR) must change in opposite directions. In this case, the weak isoform changes its expression (from not being expressed at all to being weakly expressed). However, the most expressed isoform does not change its expression at all and therefore, its statistical significance is low using the EventPointer test. The reason to implement this restriction is that this type of changes (one isoform strongly expressed and a weak change in the other) has a debatable biological implication. Nevertheless, the EventPointer test can also be changed to detect this type of events. If this is done, for this particular case the p.value is smaller than $1e^{-16}$ and therefore strongly significant.

\textit{DNM2} (2595\textsuperscript{th}, 26935\textsuperscript{th} and 1781\textsuperscript{th} in EventPointer, TAC and AltAnalyze, respectively) is a mutually exclusive exon event. Even though the ranking is not good in EventPointer, it was validated by PCR (figure 4.6). TAC shows that, although the change in expression between
both isoforms is not large, each of the paths has different signs in the corresponding SI.

![Figure 4.6](image)

**Figure 4.6:** PCR results for alternative splicing events in genes NAA50 and DNM2.

### 4.7 Discussion

EventPointer provides a complete pipeline to detect alternative splicing events using HTA 2.0 arrays. The main advantages of this method over the Splicing Index or ASPIRE are that:

1. it can be applied to any experimental design (by providing the corresponding design and contrast matrices) and not only to case-control studies,
2. it exploits the redundancy of all the junction probes involved in an alternative splicing event and
3. it labels all the events according to the different categories. All the suggested events can be validated using standard PCR by construction.

EventPointer is event-focused instead of transcript-focused [12, 13]. It estimates the statistical significance of splicing events without estimating the concentration of the underlying transcripts. For each event, EventPointer identifies the type of event (cassette, alternative start site, alternative donor, etc.) and provides its statistical significance according to the design and contrast matrices given by the user. EventPointer also generates a graphical output using IGV [98] to make the identification and validation of the events easier.

There are, however, some events that cannot be identified with EventPointer. For example, a couple of isoforms with different transcription start sites in which one of them is included in the other. These types of events are also very difficult to detect using RNAseq or PCR since there is no a specific sequence in the second isoform to design a primer.

EventPointer considers that there is a differential splicing event if the isoforms in the associated paths change their expression in opposite directions. Although this statistical test can miss some events, the selected events have a clear change on their expression. These
changes, usually, have more biological relevance than other subtler cases in which only one isoform, usually weakly expressed, changes its expression across the conditions.

Once the Affymetrix CEL files are analyzed (i.e. background corrected, normalized, and summarized using the standard procedures), the statistical analysis to detect the alternative splicing events is very fast by using the limma [97] R package. Using a standard Intel i5 processor, the analysis requires about 10 seconds. The hardware requirements are modest (a low-end desktop computer with 4GB of RAM is sufficient). The whole enrichment analysis takes only fractions of a second. This is an advantage compared to the requirements on storage, computational power and memory of RNA-seq analysis. In addition to that, the proposed methodology is very reliable: only one false positive was found within the top-200 tested events.

One of the key parts of the analysis of alternative splicing is to provide a biological interpretation of the splicing events, i.e. what is the difference between the isoforms expressed in a condition specific manner. EventPointer provides the protein family domains that are affected on each of the splicing events. It also performs an enrichment analysis (event-based) to identify which are the domains that are significantly over or underused in the condition under study. Even though in its present form, EventPointer only provides information on the PFAM domains the algorithm is being actively developed to provide annotation for other domain databases such as Pirsf, Superfamily, Smart, Prosite, or Interpro. In addition to protein domains, there are other interesting biological data that could be inferred. For example in [96], Ray et al. identify the binding motifs of several hundreds of RNA binding proteins and the potential binding sites in the human genome. Using this information is possible to predict which are the splicing factors that are driving the differential usage of isoforms. Another potential improvement would be to identify miRNA binding sites and check if the splicing pattern causes skipping these binding sites and therefore, the corresponding miRNAs may be no longer regulating the expressed isoforms. This functionality is already offered by AltAnalyze and it is expected to be included in EventPointer in the near future.
RNA-Seq is a reference technology for determining alternative splicing at genome-wide level. Exon arrays remain widely used for the analysis of gene expression, but show poor validation rate with regard to splicing events. Commercial arrays that include probes within exon junctions have been developed, in order to overcome this problem. In this chapter, the performance of RNA-seq (Illumina HiSeq) and junction arrays (Affymetrix Human Transcriptome array) for the analysis of transcript splicing events is compared. Three different cell lines were treated with CX-4945, a drug that severely affects splicing. Common results and discrepancies between the technologies were validated and/or resolved by over 200 PCR experiments.

5.1 Introduction

The development of exon microarrays enabled the transcriptomic study of differential splicing events, but PCR validation rates for identification of splice variants via microarray analysis tend to be lower than those observed for identification of differential gene expression using similar technologies [45, 64, 139]. Junction arrays [36, 100, 108, 110, 115] have been proposed to overcome this problem by using oligonucleotide probe-sets that interrogate junctions between exons in the transcriptome, as well as the exons themselves.

Since the advent of next-generation sequencing (NGS), RNA-Seq has become the technology of choice via which to detect and quantify alternative splicing (for a review see [130]).
Various published works compare the performance of RNA-Seq and expression microarrays for the analysis of gene expression [41, 127], but a thorough evaluation of both technologies in terms of their ability to detect differential alternative splicing events has yet to be presented.

The principal aim of this chapter is to compare RNA-Seq technology (using the Illumina HiSeq platform) and junction arrays commercialized by Affymetrix (Human Transcriptome array, or HTA 2.0) in their ability to detect differential alternative splicing events. To do so comprehensively, and to allow as close to a direct comparison as possible, EventPointer algorithm was applied to data from both platforms, generated from the same control experiment.

CX-4945 is a potent and selective orally bioavailable small molecule inhibitor of casein kinase CK2 [111], which has been proposed previously as a cancer therapy [25], and which has been shown to regulate splicing in mammalian cells [63]. RNA samples taken from three distinct triple negative breast cancer (TNBC) cell-lines, exposed to CX-4945 and also to a DMSO control, were profiled using both RNA-Seq and hybridization to exon-junction microarrays.

### 5.2 Experimental design

The experiment comprises three distinct triple-negative breast cancer (TNBC) cell-lines, each of them exposed to CX-4945 (a drug known to affect the transcriptional machinery) and DMSO respectively across five replicates.

Triple negative breast cancer cell lines MDA-MB-231 and MDA-MB-468 were obtained from ATCC (Manassas, VA) and SUM149 was purchased from Asterand plc (Detroit, MI). All cell lines were grown according to the suppliers’ recommendation. CK2 inhibitor CX-4945 (Selleckchem, Houston, TX) was dissolved in DMSO and stored frozen at -80°C until used.

To induce splicing events, cells were grown to 70% confluence and treated with 1µM CX-4945 or DMSO during 12 hours in a total of 5 replicates per condition. Total RNAs were isolated using the RNeasy Mini Kit (Qiagen, Germantown, MD) according to the manufacturer’s protocol. Integrity of RNA was quantified using the Agilent 2100 Bioanalyzer (Agilent Biosystems, Foster City, CA). Samples were labeled and hybridized in Human Transcriptome arrays (HTA 2.0) by the Genomics Core Facility of the Center for Applied Medical Research (CIMA) following manufacturer’s instructions.

RNA-Seq was performed in the Center for Cooperative Research in Biosciences (CICBio-
5.2. EXPERIMENTAL DESIGN

gene) using the Illumina HiSeq2000 sequencing technology, HiSeq Flow Cell v3 and TruSeq SBS Kit v3. 2µg of RNA of each sample was sent for this purpose. The run type was strand specific, multiplexed with paired-end reads of 100 nucleotides each. The amount of RNA for hybridization and validation purposes was 5 µg.

STAR 2.4.0h1 was used to align the reads against the human genome. The reference genome was Ensembl GRCh37.75. The output were sorted BAM files. All the other parameters were set to the default values. The average sequencing depth was 49 million reads (9.8 billion nucleotides sequenced per sample).

The microarray data preprocessing was performed using the aroma.affymetrix framework using the standard RMA algorithm [53].

The comparison and analysis of the profiling data was done using a linear model. The design matrix was built considering both the cell line and treatment with CX4945 as factors. The interaction between cell line type and treatment was not considered. The selected contrasts test for the difference between control samples (DMSO) and drug exposed ones (CX4945) controlling for the cell-type. The complete experimental design in the form of design and contrast matrices is included in equation 5.1.
In order to identify well-expressed events (more likely to be biologically significant and less prone to validation error), the comparison of alternative splicing detection was performed on a subset of the data with expression above a set threshold.

For arrays, the signal of the probesets interrogating each of the alternative paths involved in a splicing event, must be expressed more than a certain threshold in at least one sample. This threshold is the 25% quantile of the expression of the signal in the reference paths for all the events included in the array. For RNA-Seq, the edges of the splicing graph (junction reads) are included only if their expression is at least 2 FPKM in at least one sample (SGSeq defaults).

### 5.3 Matching events

Events detected by both technologies (referred to as "matched events" hereon) were defined by a stringent criterion in which nucleotide sequences of paths identified via one technology

\[
C = \begin{bmatrix}
    0 \\
    MD\text{A}231 \\
    MD\text{A}468 \\
    CX4945
\end{bmatrix}
\]

\[
D = \begin{bmatrix}
    Intercep & MD\text{A}231 & MD\text{A}468 & CX4945 \\
    SUM149 DMSO 1 & 1 & 0 & 0 & 0 \\
    SUM149 DMSO 2 & 1 & 0 & 0 & 0 \\
    SUM149 CX4945 1 & 1 & 0 & 0 & 1 \\
    SUM149 CX494 2 & 1 & 0 & 0 & 1 \\
    MDA231 DMSO 1 & 1 & 1 & 0 & 0 \\
    MDA231 DMSO 2 & 1 & 1 & 0 & 0 \\
    MDA231 CX494 1 & 1 & 1 & 0 & 1 \\
    MDA231 CX494 2 & 1 & 1 & 0 & 1 \\
    MDA468 DMSO 1 & 1 & 0 & 1 & 0 \\
    MDA468 DMSO 2 & 1 & 0 & 1 & 0 \\
    MDA468 CX494 4 & 1 & 0 & 1 & 1 \\
    MDA468 CX494 5 & 1 & 0 & 1 & 1
\end{bmatrix}
\]
must be a subset of sequences identified via the other.

Let’s assume that $A_R$ and $A_M$ are, possibly non-contiguous, regions of the genome that correspond to path $A$ using either technology ($A_R$ for RNA-Seq and $A_M$ for HTA). $B_R$ and $B_M$ have a similar description for path $B$ and $R_R$ and $R_M$ for the reference path in each technology. Two events are considered to match if any of the following two expressions is true:

\[
((A_R \subset A_M) \& (A_M \subset A_R)) \& ((B_R \subset B_M) \& (B_M \subset B_R)) \& ((R_R \cap R_M) \neq \emptyset) \\
((A_R \subset B_M) \& (B_M \subset A_R)) \& ((B_R \subset A_M) \& (A_M \subset B_R)) \& ((R_R \cap R_M) \neq \emptyset)
\] (5.2)

In these expressions, $(x \subset y)$ is true if the genomic region $x$ is a subset of the genomic region $y$ (the nucleotide sequence of $x$ is a substring of the nucleotide sequence in $y$). Besides, the operators ($\&$) and ($\|$) are the logical OR and AND operations. If $x \subset y$ or $y \subset x$, then one of the regions is contained in the other are considered to be "compatible". On the other hand, $(x \cap y) \neq \emptyset$ means that regions $x$ and $y$ overlap in the genome. Therefore, the first expression is true if both paths $A_R$ and $A_M$ are compatible, $B_R$ and $B_M$ are compatible and $R_R$ and $R_M$ overlap. The second expression is true if path $A_R$ and path $B_M$ are compatible and also path $A_M$ and $B_R$ are compatible and, again, and $R_R$ and $R_M$ overlap.

Within an event, the longer path in the transcriptome is assigned the name $A$ and the other the $B$. The second equation takes into account that, in some few cases, the name of the paths can be switched in both technologies.

### 5.4 Results

Table 5.1 displays the number of events detected via application of EventPointer to RNA-Seq and microarray data respectively as these thresholds are varied. As may be expected, setting more stringent expression thresholds yields fewer events detected with better False Discovery Rate (FDR) on both platforms.

Fixing $p$-value to 0.001 yields False Discovery Rates (FDRs) less than 1% for both technologies. The expected proportion of alternative splicing events appears high ($\pi_0$ approx. 54%) [14], i.e. more than 46% of the events have its splicing patterns altered, which reflects the anticipated strong effect of compound exposure on the splicing machinery. It is also apparent that, for a similar number of detected splicing events, the FDR corresponding to RNA-Seq analysis is smaller.
Table 5.1: Number of detected splicing events using both RNA-seq and array technologies for different expression thresholds

<table>
<thead>
<tr>
<th>Expression Threshold</th>
<th>RNA-Seq Detected Events</th>
<th>Microarrays Expression Threshold</th>
<th>Microarrays Detected Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Junction coverage &gt; 6 FPKM</td>
<td>9,277</td>
<td>Signal &gt; 50%</td>
<td>10,114</td>
</tr>
<tr>
<td>Junction coverage &gt; 2 FPKM</td>
<td>34,961</td>
<td>Signal &gt; 25%</td>
<td>31,506</td>
</tr>
<tr>
<td>Junction coverage &gt; 2.5 FPKM</td>
<td>92,986</td>
<td>No threshold</td>
<td>92,405</td>
</tr>
</tbody>
</table>

The method described in section 5.3 yielded 6,222 matched events. When reporting correspondence and divergence between alternative splicing events, below, the following naming convention is used: $R^+$ represents number of events deemed significantly altered in RNA-Seq analysis; $R^-$ represents number of events deemed not significantly altered in RNA-Seq analysis. $M^+$ and $M^-$ are the counterpart terms used to describe microarray results. Events not detected by each technology are labelled $R^∅$ and $M^∅$ respectively.

A subset of matched events are significant in both technologies ($R^+M^+$) and show coherent change in the corresponding $Ψ$. There are also significant events detected by only one of the technologies ($R^+M^∅$ and $R^∅M^+$). The summary of findings is presented in Table 5.2.

Table 5.2 shows that the FDR of the events detected only by RNA-Seq is similar to that for events detected by both platforms (0.0456% vs. 0.0496%). In other words, the reliability of events discovered only by RNA-seq is similar to that of events identified by both technologies. In the case of the arrays, the FDR of matched events is three times smaller than for those discovered solely by the arrays (0.199% vs 0.0623%), i.e. $R^∅M^+$ events are less reliable than $R^+M^+$ events for the same p-value threshold. In addition, Table 5.2 shows that the number of significant events that are RNA-Seq specific ($R^+M^∅$) is larger than the number of significant events detected only by arrays ($R^∅M^+$) (10,617 vs 3,297 events).

Figure 5.1 depicts a Sankey diagram of the relationship between matched events. It is apparent that many events that are significant for RNA-Seq are not detected by arrays, but also that events significantly detected via arrays are not detected by RNA-Seq. Most matched events are consistent across technologies: significant events for one technology are also significant for the other.
### Table 5.2: Number and statistical significance of detected alternative splicing events using both RNA-Seq and microarray technologies.

<table>
<thead>
<tr>
<th>Condition</th>
<th>RNA-Seq Detected Events</th>
<th>Microarrays Detected Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matched Events</td>
<td>6,222</td>
<td>6,222</td>
</tr>
<tr>
<td>Significant in both ($R^+M^+$)</td>
<td>1,324</td>
<td>1,324</td>
</tr>
<tr>
<td>FDR for matched events (RNA-Seq)</td>
<td>0.0496%</td>
<td>FDR for matched events (Microarrays) 0.0623%</td>
</tr>
</tbody>
</table>

#### Number of events in $R^+M^+$

<table>
<thead>
<tr>
<th>Expression Threshold</th>
<th>RNA-Seq Detected Events</th>
<th>Microarrays Detected Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Junction coverage &gt; 6 FPKM</td>
<td>2,973</td>
<td>Signal &gt; 50% 1,016</td>
</tr>
<tr>
<td>Junction coverage &gt; 2 FPKM</td>
<td>10,617</td>
<td>Signal &gt; 25% 3,297</td>
</tr>
<tr>
<td>Junction coverage &gt; $\frac{2}{3}$ FPKM</td>
<td>25,063</td>
<td>No threshold 7,581</td>
</tr>
</tbody>
</table>

#### FDR for $R^+M^+$

<table>
<thead>
<tr>
<th>Expression Threshold</th>
<th>RNA-Seq FDR</th>
<th>Microarrays FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Junction coverage &gt; 6 FPKM</td>
<td>0.0244%</td>
<td>Signal &gt; 50% 0.146%</td>
</tr>
<tr>
<td>Junction coverage &gt; 2 FPKM</td>
<td>0.0456%</td>
<td>Signal &gt; 25% 0.199%</td>
</tr>
<tr>
<td>Junction coverage &gt; $\frac{2}{3}$ FPKM</td>
<td>0.0690%</td>
<td>No threshold 0.485%</td>
</tr>
</tbody>
</table>
Figure 5.1: Correspondence between the events detected by arrays and RNA-Seq. An event is considered to be significant if the p-value is smaller than 0.001 and non-significant if it is larger than 0.2. Events with p-values between both are considered to be inconclusive cases.

5.4.1 PCR validation

PCR validation was performed on a subset of predicted alternative splicing events drawn from each of the subsets discussed in previous sections, i.e. events detected by one or both technologies. PCRs were performed on:

1. Five top-ranked events detected by both technologies regardless of the matching.
2. Five top-ranked events detected by one technology ($R^+ M^0$ and $R^0 M^+$)
3. Five top-ranked events significant in one technology ($R^+ M^-$ and $R^- M^+$)
4. Five top-ranked events detected by both technologies ($R^+ M^+$)

PCR for events in non-coherent classes ($R^+ M^-$, $R^- M^+$) required up to 40 PCR cycles and were harder to validate in general. The results of the PCR validations are shown in table 5.3.
Table 5.3: PCR validation for RNA-Seq and microarray technologies across events detected by one or both technologies. ✓ denotes positive validations and ✗ denotes negative results.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Gene</th>
<th>Type</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top 5 RNA-Seq</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DONSON</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>OSGIN2</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>ACTR10</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>KIF20B</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>USP9X</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Top 5 Microarrays</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DONSON</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>TOR1AIP2</td>
<td>Alternative Last Exon</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>PSMG1</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>COPS7A</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>POLR3C</td>
<td>Retained Intron</td>
<td>✓</td>
</tr>
<tr>
<td><strong>R⁺M⁻</strong></td>
<td>OSGIN2</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>ACTR10</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>KIF20B</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>USP9X</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>CSF2</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td><strong>R⁻M⁺</strong></td>
<td>KDM1B</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>C1orf43</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>MRFAP1</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>EXOSC10</td>
<td>Alternative Last Exon</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>UBQLN1</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td><strong>R⁺M⁺</strong></td>
<td>SAYSD1</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>RNF185</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>GPBP1</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>CASP6</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>ZNF692</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td><strong>R⁻M⁻</strong></td>
<td>UBIAD1</td>
<td>Complex Event</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>CTC1</td>
<td>Alternative 3’</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>POLR2J4</td>
<td>Complex Event</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>LRRC37BP1</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>WNK1</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td><strong>R⁺M⁺</strong></td>
<td>DONSON</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>MELK</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>ACADM</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>KIF14</td>
<td>Cassette Exon</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>TOP2B</td>
<td>Complex Event</td>
<td>✓</td>
</tr>
</tbody>
</table>

Figure 5.2 shows the Ψ estimates and the PCR bands for two of the top-ranked events in $R⁺M⁺$ (gene names DONSON and MELK), with clear concordance of the splice index, Ψ, across the three technologies despite use of end-point (i.e. non quantitative) PCR.
Figure 5.2: Estimated PSI (for RNA-Seq, microarrays and PCR image analysis), PCR bands, the reference HTA transcriptome and the alternative paths of the DONSON (panel A) and MELK (panel B) genes in R+M+. This concordance occurs for most of the genes validated by PCR. The last numbers shown are expected bands for the selected primers.

Figure 5.3 shows the increment of the $\Psi$ value estimated by EventPointer for events detected by both technologies. Correlation for alternative splicing events is over 0.90, and z-values of the statistical test are also similar. PCR figures also show high coherence between the estimated $\Psi$ using both technologies, especially for RNAseq, and the PCR results.

Regarding the type of alternative splicing events detected by both technologies, figure 5.4 shows the number and type of alternative splicing events detected by the EventPointer algorithm on data from both profiling technologies. The number of detected cassette exons using arrays is smaller than that using sequencing (p.value $<$ 1e-16, test for equality of proportions). In fact, after matching the events detected by both technologies, a large proportion of the cassette exons in RNA-Seq appear as complex in microarrays (see figure 5.4). The reason for this disparity is the complexity of the reference transcriptome used in the HTA array. For this analysis, we used the transcriptome provided by Affymetrix, which includes a range of annotation sources, e.g. RefSeq, Vega, Ensembl, MGC (v10), UCSC known genes and other sources for non-coding isoforms. The underlying transcriptome for HTA includes such a variety of isoforms that many detected events are labelled as complex.

In addition, the proportion of retained introns is smaller for RNA-Seq (p.value $<$ 1e-16,
5.4. RESULTS

Figure 5.3: Increment of $\Psi$ for both microarrays and RNA-Seq. The black (gray) dots represent events with high (low) standard deviation in the differential usage of the isoforms in both paths. Correlation between events with high and low variability are 0.90 and 0.61 respectively.

test for equality of proportions), perhaps owing to the coverage required to include a region as expressed by SGSeq (defaults to 0.5 FPKM) which may exclude weakly expressed introns.

The comparisons above suggest that that RNA-Seq, at the depth of sequencing deployed here, detects a larger number of splicing events at lower FDR.

Successive approximations using 30% and 10% of the initial RNA-Seq reads (approximately 30 million and 10 million reads respectively) yield the estimate that FDR obtained using junction arrays in the present comparison is equivalent to that which would be obtained by an RNA-Seq experiment with sequencing depth of 20 million reads.

The filter used here to detect RNA-Seq events was based on expression alone, however, rather than design or contrast matrices. Accordingly, we hypothesize that application of a filter based on an estimate of expression (for example by developing an accurate estimate of probe affinities and potential cross-hybridization) would improve the performance of the HTA 2.0 arrays significantly. In fact, if a similar approach is applied using only matched
CHAPTER 5. COMPARISON OF MICROARRAYS AND RNA-SEQ

Figure 5.4: A) Events detected using RNA-Seq and array technologies. B) Type of event after matching the events detected by both technologies

events, HTA 2.0 arrays are found to be equivalent to RNA-Seq with a depth of approximately 65 million reads.

5.5 Discussion

The main aim of this chapter was to quantitatively compare the performance of RNA-Seq and junction arrays technologies to detect splicing events. The fact of using the same algorithm for both technologies makes the comparison fairer and more accurate.

Both RNA-Seq and array technologies have great validation rates: 100% for the top-ranked events for both HTA 2.0 and RNA-Seq. There is also a strong coherence in their results. The estimated ΔΨ value for events detected by both technologies has a correlation close to 0.9. In other words, the events that are detected by both technologies show almost identical estimates of ΔΨ. This estimate is, in turn, confirmed by PCR validation. Incidentally, this coincidence on the estimate using the three technologies confirms the validity of the algorithm to compute ΔΨ.

Despite of the thorough gathering of transcriptome information to build the HTA array, there are still many unannotated events. More than 10,000 statistically significant events, that could be unveiled using only the RNA-Seq technology, were detected. Some of these
novel events were validated through standard PCR (100% validation rate). In them, the expression of both isoforms is acceptably high in some of the conditions and display alternative splicing. They are true events that may have a biological impact and no probes in the arrays interrogate them. Interestingly, the RNA-Seq specific events have FDR similar to events detected by both technologies. Therefore, it seems that the fact of being detected by HTA does not improve the reliability of the events detected by RNA-Seq.

Interestingly, there are also almost 3,300 events that were detected using microarrays and non-detected using RNA-Seq. Some of them (\textit{WNK1, LRRC37BP1}) were perfectly validated and are well-expressed. Other events in this group, for example \textit{CCNHT}, were weakly expressed or were not validated. The estimated FDR of these events is three times larger than the FDR for the matched events and the validation rate is also smaller (3 out of 5). It seems that this group is enriched with events with weaker expression that, in turn, are more difficult to validate. Probably, an accurate estimate of the affinities of the probes would help to discern these weakly expressed events.

Although RNA-Seq with this depth outperforms HTA, it is difficult to discern the threshold in number of reads to state which technology is better. The FDR was used to compare both technologies. Arguably, it is not advisable to compare both technologies with fabricated statistical properties and therefore, this value must be taken with caution. In terms of PCR validation, the results of both technologies are very good, and only when testing events in which both technologies are non-coherent, RNA-Seq outperforms HTA. Any problem of the technologies (cross hybridization of the probes, multi mapping reads, GC dependence, etc.) tends to enrich these groups of non-coherent results.

In addition to the statistical analyses and PCR validations, other issues must be taken into account. RNA-Seq is inherently more flexible than microarrays: its sensitivity can be improved simply by increasing the depth of the sequencing and is able to detect novel events not previously annotated in any database. These novel events can be especially valuable since can be specific to the condition under study and, therefore, be potential markers for diagnosis. Another advantage of RNA-Seq is its "quasi-absolute quantification" property: if the concentration of the genes/transcripts is required (for example to state a threshold based on the expression of an event), then RNA-Seq is the method of choice.

It seems that in some cases, the arrays are able to detect weakly expressed events perhaps because of a strong affinity of the interrogating probes. If a thorough study of the transcriptome is required, the arrays can be complementary to RNA-Seq. On the other hand, computer requirements and bioinformatician time are much shorter for microarrays than
CHAPTER 5. COMPARISON OF MICROARRAYS AND RNA-SEQ

Table 5.4: Resources required for both technologies

<table>
<thead>
<tr>
<th></th>
<th>Computing Time</th>
<th>Memory Requirements</th>
<th>Storage Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RNA-Seq</td>
<td>HTA</td>
<td>RNA-Seq</td>
</tr>
<tr>
<td>Mapping to the transcriptome</td>
<td>11.5 hours</td>
<td>-</td>
<td>32 Gb</td>
</tr>
<tr>
<td>(STAR)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Splicing graph generation</td>
<td>2 days (5 cores)</td>
<td>14 hours</td>
<td>8 Gb per core</td>
</tr>
<tr>
<td>(SGSeq)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Event Detection</td>
<td>7 min 16s (10 cores)</td>
<td>1.2 Gb per core</td>
<td>5 Gb</td>
</tr>
<tr>
<td>Statistical Analysis</td>
<td>1 min 43s</td>
<td>3 min 06s</td>
<td>2 Gb</td>
</tr>
</tbody>
</table>

for RNA-Seq analyses and the pipelines are much more settled for them. This has also economic implications when calculating the overall cost and time to get available results of each experiment. Table 5.4 is included to quantitatively state the computational requirements for each technology.

This table shows that both the computation time and storage requirements are much larger (one to two orders of magnitude larger) for RNA-Seq than for HTA 2.0. The analysis was performed on 16 cores (Intel Xeon E5-2670 @ 2.60 GHz) with 64 GB of RAM Linux server running 64 bit CentOS distribution.

In conclusion, the use of the same algorithm to compare RNA-Seq and array technologies ensures a fair comparison. In addition, the concordance of the $\Psi$ estimates with a third technology (PCR) confirms that the overall methodology is correct. Both RNA-Seq and junction-arrays have extremely good validation rates. The coherence between them is very good: the events that are predicted by both technologies give strikingly similar results when compared by PCR. RNA-Seq has an edge in terms of flexibility and validation rate for this depth of sequencing. It also provides better PCR validation rates with the (few) non-coherent results. On the other hand, HTA provides also very good validation rates and has smaller requirements in terms of computer storage, computation time and bioinformatics expertise. HTA 2.0 shows to be a reasonable alternative to shallow RNA-Seq.
In this chapter, a brief summary of the main findings obtained with EventPointer are described. In the second part of this chapter, the future lines for development of the algorithm are described. These future lines can provide insight in order to increase the application of the algorithm to broader datasets and more complex problems related with the detection of alternative splicing.

6.1 Conclusions

It has been shown that alternative splicing plays a key role in human biology. The identification of splicing variants has been linked with the development of different pathologies such as cancer, alzheimer or muscular atrophy. Also, the characterization of alternative splicing events can be used as novel means for diagnosis, prognosis and identification of therapeutic targets. Due to this, the complete understanding of the mechanisms that control alternative splicing has taken an important role in active research.

As novel isoforms are characterized, the human transcriptome becomes more complex. This creates a need to develop new technological platforms in order to obtain relevant measurements from the different splicing variants. With junction arrays and sequencing technologies it has been possible to develop such experimental procedures. However, both platforms are being actively improved to reduce cost, time and provide more data. Third generation sequencing devices start to take an important role in this problem.
CHAPTER 6. CONCLUSION AND FUTURE LINES

With the improvement of technological platforms, new algorithms are constantly developed to identify alternative splicing events. Some of the available algorithms work as a complete pipeline to identify and state the significance of each event, while others require the addition of different tools to perform a complete analysis.

In chapter 3 EventPointer has been described as a novel alternative to identify, classify and analyze alternative splicing events. The algorithm can be applied to data from both junction arrays and RNA-Seq. The statistical model used in EventPointer provides one single figure of merit per splicing event to state statistical significance, rather than providing multiple ones as other available algorithms. The proposed method looks for the coherence of the alternative splicing event: if the isoforms mapped to one of the alternative paths are changing its expression, the isoforms in the opposite path should change in the opposite direction. With this condition, authentic isoform switches can be identified that are more biologically relevant. The proposed model to estimate $\Psi$ is also prominent. The estimated $\Psi$ values obtained for both, microarrays and RNA-Seq, were confirmed using end-point PCR. This new method provides researchers with a new parameter that can be used to filter out splicing events without relying only on statistical significance. The visualization, provided by EventPointer, eases the interpretation of the results and allows to design primers for PCR validation and thus, the time required to perform a complete experiment is reduced.

The performance of the algorithm was tested in two independent datasets (SRSF1 and CX4945) for two different technological platforms such as Affymetrix junction-arrays and RNA-Seq. In both experiments and platforms, the top ranked events were positively validated using standard PCR. These results prove that the developed methodology is efficient and less prone to the detection of false positives. In the SRSF1 dataset, the validation rate was above 80% (37/40). In the case of CX4945, the 5 top ranked events for both technologies were properly validated.

Both available technologies (microarrays and RNA-Seq) to identify splicing events provide accurate results that are, in turn, correlated between each other in terms of statistical significance. Sequencing methods can provide better results just by increasing the library size, however still certain events might not be detected. The probes for new generation junction arrays are designed using different annotation sources, this allows for the detection of most of the known splicing events. Also, the required capabilities for processing data from microarrays is smaller in terms of knowledge and money. Eventhough RNA-Seq might still be the preferred technology for transcriptomic analysis, microarrays can be a good option for shallow RNA-Seq.
Recently, Affymetrix released a new version of the Transcriptome Analysis Console (TAC 4.0). In this version, EventPointer has been included as an option to identify splicing events. As stated in the TAC 4.0 datasheet: "Combining EventPointer with existing algorithms in TAC Software for alternative splicing analysis, offers two independent algorithms for better statistical measurement of splicing results". The inclusion of EventPointer in TAC allows researchers to perform the analysis without the need of coding knowledge in R.

6.2 Future lines

6.2.1 Multi-path events

The definition of alternative splicing event by EventPointer consists in a reference path and two alternative paths ($P_R$, $P_1$ and $P_2$). If there is a splicing event, with more than two alternative paths, the current method will not detect it as the sum of the fluxes will not be equal to the reference. This type of events can be identified in a splicing graph by looking for nodes with a single input and more than two outputs.

The mathematical model for both, statistical significance and $\Psi$ estimation, rely on two alternative paths only. In order to provide such figures of merit, for multi-path events, a different method should be used. One possible option is to provide a $\Psi$ value per path and use such results to perform the statistical analysis.

6.2.2 Automatic design of PCR primers

In the case of microarrays, the number of alternative splicing events remains constant as long as the same microarray is used. This number is independent of the samples or experimental design. Due to this, the required PCR primers to validate any of the detected event will always be the same. The definition of splicing event by EventPointer guarantees that any event can be validated with at most three primers.

A method can be developed to identify the required primers for all the events. By using similar approaches, related with graph theory, it is possible to identify the number and position of the primers. Using this approach, in combination with other available software (i.e Primer 3) the algorithm could provide as an output not only the statistical significance of the splicing events, but the required primers for PCR validation and thus, reducing the required time for a complete analysis.

For RNA-Seq, the events depend on many factors such as library size and the sequenced
samples. Eventhough, the events will change for every single experiment, it is possible to apply a similar methodology to obtain the required primers per RNA-Seq experiment.

6.2.3 Junction arrays probe quality

Affymetrix provides figures related with the quality of the probes included in junction arrays. This information could be used to improve the significance of every splicing event. When building the corresponding linear models, to test for differential splicing, a penalty factor could be included that is related with the quality of the probes that measure the event under study. This method would cause that events, that are measured by a set of probes with low quality, require much more degree of change before being considered as relevant. Certain events include paths that are measured by a few probes (2 or 3). The number of probes could also be used as a penalty when testing for differential splicing.

6.2.4 Biological interpretation of alternative splicing

EventPointer identifies alternative splicing events, but many unanswered questions appear once the events are detected. Some examples of these questions can be related with effect of specific splicing events, i.e. disrupting protein domains or creating novel binding motifs. Also, the affected splicing events can be used to predict which splicing factor is controlling such changes. These turns into two independent analyses can be performed. For example, with the results from the SRSF1 dataset, it is possible to identify specific events related with such splicing factor. In case new samples were analyzed, and those specific events appeared as significant, a relationship could be established between the condition and the splicing factor.

Due to the vast amount of available datasets in different online databases, i.e. Gene Expression Omnibus, Short Read Archive, Array Express, a splicing database could be created. This database could include the affected splicing events depending on different tissues, pathological conditions or other clinical phenotypes. Such databases exists for gene expression and can easily be extended to alternative splicing.
Primers used for the validation of the detected alternative splicing events using Event-Pointer with the SRSF1 dataset. The table includes: gene name, direction of the primers, primer sequence and expected product sizes of the splicing variants.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Direction</th>
<th>Sequence (5’→3’)</th>
<th>Product sizes (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYCBP2</td>
<td>Forward</td>
<td>CAGTGCCCAAGGATTTGATT</td>
<td>119 / 299</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>TGCTGTGTAGAGGACAGGCTC</td>
<td></td>
</tr>
<tr>
<td>KIF23</td>
<td>Forward</td>
<td>TGTGGCTAATCCCTTGCTGTA</td>
<td>98 / 235</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>GGCATCAACCTCACACTGTA</td>
<td></td>
</tr>
<tr>
<td>AC024560.3</td>
<td>Forward</td>
<td>CCTACCAGTCACACTGT</td>
<td>98 / 242</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>CCACCAATGCACGCTGTA</td>
<td></td>
</tr>
<tr>
<td>FBXO22</td>
<td>Forward</td>
<td>GTTGAGTAACCTGGCGGAG</td>
<td>101 / 240</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>ACTGTATGTGGTAGAGATGCGGAAC</td>
<td></td>
</tr>
<tr>
<td>SRSF3</td>
<td>Forward</td>
<td>GAGTCCTCCACCTGCGT</td>
<td>125 / 581</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>GGGACGGCTTGTGATTTC</td>
<td></td>
</tr>
<tr>
<td>SUPT16H</td>
<td>Forward</td>
<td>CTCGCGTGATTCTCTGAAAC</td>
<td>166 / 193</td>
</tr>
<tr>
<td></td>
<td>Reverse 1</td>
<td>TCAACCCACTGATACAACA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reverse 2</td>
<td>CTGTTAAGGGCCAGGCTTC</td>
<td></td>
</tr>
<tr>
<td>HMBOX1</td>
<td>Forward</td>
<td>CCCAGATGAAGCAAGAGG</td>
<td>152 / 200</td>
</tr>
<tr>
<td></td>
<td>Forward 2</td>
<td>ACTTGAACTCTCCCTTGGAAACA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reverse 1</td>
<td>TGGCTCTCCTCTTGATCTCC</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Forward 1</td>
<td>Forward 2</td>
<td>Reverse 1</td>
</tr>
<tr>
<td>-------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>ACAD11</td>
<td>AAAATGGGAGGCACACGTGGAG</td>
<td>ACAAGAATCATGCTGGCTGTGTT</td>
<td></td>
</tr>
<tr>
<td>NCOR1</td>
<td>GGTCTTGAGGAGGCAAAAGTTGA</td>
<td>GAGCTGATCATACAGCCGTCG</td>
<td></td>
</tr>
<tr>
<td>AUP1</td>
<td>GGAACCTGTCCTACACCAC</td>
<td>TGCATATCATAGTGCTGTGCTT</td>
<td></td>
</tr>
<tr>
<td>IFT27</td>
<td>AAGAAATTCGCGCGCTGATC</td>
<td>GACAGCTGAGACTAGCTGC</td>
<td></td>
</tr>
<tr>
<td>GALNT10</td>
<td>GATCCTCTCAGAATCGAGA</td>
<td>TCTGATGATAGTTGCCAC</td>
<td></td>
</tr>
<tr>
<td>PARD3</td>
<td>CCCATGAAACAGCAGTGGA</td>
<td>CAGGACGAGGTACGAGGAA</td>
<td></td>
</tr>
<tr>
<td>PRMT2</td>
<td>CTCAAGGGCTCTCGAGAAGGAC</td>
<td>CCACAAATCCTCTGGCTGTGTT</td>
<td></td>
</tr>
<tr>
<td>HORMAD1</td>
<td>CCTGTATACGTATTTGAGGGG</td>
<td>TGTCTGAGGACATTCTGGTT</td>
<td></td>
</tr>
<tr>
<td>ANAPC7</td>
<td>GCTGTCAAATGAGTACGGAGG</td>
<td>TCTCATTCTCTGCACTCCCC</td>
<td></td>
</tr>
<tr>
<td>OGT</td>
<td>CTGTGTTCCGAGTGACCT</td>
<td>ATGCTGCAGGTTACCTTTTG</td>
<td></td>
</tr>
<tr>
<td>MSL3 (23)</td>
<td>CCGGCAGGCAGCAGCACATC</td>
<td>GTTCTTCTCCCTCAAAGCCA</td>
<td></td>
</tr>
<tr>
<td>NT5C</td>
<td>TGCTGAGATTACACGACTGT</td>
<td>ACGTCTCTTGCTCCCTTGTGCA</td>
<td></td>
</tr>
<tr>
<td>ALG2</td>
<td>AGGAGTTGACGGTGGATG</td>
<td>GGATCTCTCTCCGGCCGCTC</td>
<td></td>
</tr>
<tr>
<td>MSL3 (33)</td>
<td>CCAGAAGCCATTGCGAGTCTCTCT</td>
<td>AACAGTCTTGGAGGCCAGTCTTT</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Forward 1</td>
<td>Reverse</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-----------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>BAIAP2L1</td>
<td>CGTCCTCCTCCAACTTTCTT</td>
<td>TCCCTTCAAAGCCAGCAAAC</td>
<td></td>
</tr>
<tr>
<td>HIST1H2AC</td>
<td>GGCCTCCTTCTCATACTCCA</td>
<td>GGCCTCCTTCTCATACTCCA</td>
<td></td>
</tr>
<tr>
<td>DDX52</td>
<td>GGTGTAACAGTTGATCAA</td>
<td>GGTGTAACAGTTGATCAA</td>
<td></td>
</tr>
<tr>
<td>TMEM214</td>
<td>GGGAGCTACGGGGATCAGC</td>
<td>GGGAGCTACGGGGATCAGC</td>
<td></td>
</tr>
<tr>
<td>GABPB1</td>
<td>AGATGGTGCAAGACGACGAC</td>
<td>AGATGGTGCAAGACGACGAC</td>
<td></td>
</tr>
<tr>
<td>EIF3B</td>
<td>CGGTGACGGCTGGCTGGCTGG</td>
<td>CGGTGACGGCTGGCTGGCTGG</td>
<td></td>
</tr>
<tr>
<td>LAMP1</td>
<td>AAAAGAGTGAGGAGGAAGCCA</td>
<td>AAAAGAGTGAGGAGGAAGCCA</td>
<td></td>
</tr>
<tr>
<td>LMO7</td>
<td>CCACCACGCTCTCTCTGAA</td>
<td>CCACCACGCTCTCTCTGAA</td>
<td></td>
</tr>
<tr>
<td>SCAMP3</td>
<td>CCATGCTAAGAATGACCTGG</td>
<td>CCATGCTAAGAATGACCTGG</td>
<td></td>
</tr>
<tr>
<td>UHRF2</td>
<td>GCCCTAAAATGTGAGCTCAG</td>
<td>GCCCTAAAATGTGAGCTCAG</td>
<td></td>
</tr>
<tr>
<td>COPS3</td>
<td>CCATCAGCTAAGAATGACCTGG</td>
<td>CCATCAGCTAAGAATGACCTGG</td>
<td></td>
</tr>
<tr>
<td>SLSC9A8</td>
<td>CCATCAGCTAAGAATGACCTGG</td>
<td>CCATCAGCTAAGAATGACCTGG</td>
<td></td>
</tr>
<tr>
<td>FLNA</td>
<td>CGGACATCCCTCAACATCC</td>
<td>CGGACATCCCTCAACATCC</td>
<td></td>
</tr>
<tr>
<td>C21orf58</td>
<td>GCCGATACGAGCAACGAGA</td>
<td>GCCGATACGAGCAACGAGA</td>
<td></td>
</tr>
<tr>
<td>CALU</td>
<td>CCTCAGCTCACTACATCCTGT</td>
<td>CCTCAGCTCACTACATCCTGT</td>
<td></td>
</tr>
</tbody>
</table>

87
### APPENDIX A. PRIMERS SRSF1 EXPERIMENT

#### Table A.1: Primers used in the PCR validation of the detected splicing events using EventPointer on the SRSF1 dataset

<table>
<thead>
<tr>
<th></th>
<th>Reverse 2</th>
<th>ACCCATCTTTATCCGCGTCT</th>
<th>CTGTGTGGTTCGGGTGCT AGCTGTTTTCTTACAAATCGTT</th>
<th>171 / 275</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCT6P1</td>
<td>Forward Reverse</td>
<td>ST20</td>
<td>CTCTTTCCTCCGCTCTTCC GGAGACAGAGGTTGACGATGA</td>
<td>159 / 250</td>
</tr>
<tr>
<td>ST20</td>
<td>Forward Reverse</td>
<td>ACO2</td>
<td>CCTACTGGGTACTCGGGCTG TTGCTCGCTCAATTTCCTG</td>
<td>243 / 432</td>
</tr>
<tr>
<td>ACO2</td>
<td>Forward Reverse</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Primer used for the validation of the detected alternative splicing events using Event-Pointer with the SRSF1 dataset. The table includes: gene name, direction of the primers, primer sequence and expected product sizes of the splicing variants.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Direction</th>
<th>Sequence (5' → 3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donson</td>
<td>Forward</td>
<td>CCATTTGCAGGCAGATGTAA</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>CAACCTGCAGCATGTTTA</td>
</tr>
<tr>
<td>TOR1AIP2</td>
<td>Forward 1</td>
<td>TCGGCCATGTTTGTGTCTTA</td>
</tr>
<tr>
<td></td>
<td>Forward 2</td>
<td>CCATCTTCCTGCGCTACCT</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>AGCCAAGTGCCTCTGTGT</td>
</tr>
<tr>
<td>PSMG1</td>
<td>Forward</td>
<td>GACAGGGAGGTGCGCTCTG</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>ACCAACTTCTCTCCAGACTC</td>
</tr>
<tr>
<td>COPS7A</td>
<td>Forward</td>
<td>CCTGGTGCTACGTTGTGG</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>CAGGTTGACAACTGAGAGGT</td>
</tr>
<tr>
<td>POLR3C</td>
<td>Forward</td>
<td>GCTTCTTTCTTTGTGCGTTC</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>TTGCCGTTCAACAGAAGCTC</td>
</tr>
</tbody>
</table>

**Top 5 RNA-Seq**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Direction</th>
<th>Sequence (5' → 3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>DONSON</td>
<td>Forward</td>
<td>CCATTTGCAGGCAGATGTAA</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>CAACCTGCAGCATGTTTA</td>
</tr>
<tr>
<td>Gene</td>
<td>Direction</td>
<td>Sequence (5’ → 3’)</td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
<td>--------------------</td>
</tr>
<tr>
<td>OSGIN2</td>
<td>Forward</td>
<td>GGCGBAAAGTTAAAGGCTAGG</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>CAAAGTACCAGTTCAGTGGAA</td>
</tr>
<tr>
<td>ACTR10</td>
<td>Forward</td>
<td>GCGAGGACTAAAATCCAGGC</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>AACCAGGCCACACAATAGC</td>
</tr>
<tr>
<td>KIF20B</td>
<td>Forward</td>
<td>TCAAGACACTCGATTCCAA</td>
</tr>
<tr>
<td></td>
<td>Reverse 1</td>
<td>TTCTGGGTGTAGCATCTTCTT</td>
</tr>
<tr>
<td></td>
<td>Reverse 2</td>
<td>CCTATGAAAGGAGGCAAT</td>
</tr>
<tr>
<td>USP9X</td>
<td>Forward</td>
<td>CGAAGTCCCATGGGTATTA</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>TGAAGGAGCACATAGCCACAG</td>
</tr>
</tbody>
</table>

**R+M+**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Direction</th>
<th>Sequence (5’ → 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DONSON</td>
<td>Forward</td>
<td>CCATTTGCAGGCCAGATGTTA</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>CAACCCCTGCAGCATTGTTTA</td>
</tr>
<tr>
<td>MELK</td>
<td>Forward</td>
<td>ATGATCACCCTCACGGTACC</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>TGCAGGTGTTCTGCAATAGG</td>
</tr>
<tr>
<td>ACADM</td>
<td>Forward</td>
<td>AGAGCCTGGGACTTTGATT</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>CCTCAGTTCATTCTCCAAA</td>
</tr>
<tr>
<td>KIF14</td>
<td>Forward</td>
<td>TTCTTTTCTCCTGCCATTTT</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>GAAGCTAGCAGGACCACCTCC</td>
</tr>
<tr>
<td>TOP2B</td>
<td>Forward</td>
<td>AATCTGGGCCCTGAGGAGTT</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>GAGTGGCCTTGGGAGATGT</td>
</tr>
</tbody>
</table>

**R+M∅**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Direction</th>
<th>Sequence (5’ → 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSGIN2</td>
<td>Forward</td>
<td>GGCGBAAAGTTAAAGGCTAGG</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>CAAAGTACCAGTTCAGTGGAA</td>
</tr>
<tr>
<td>ACTR10</td>
<td>Forward</td>
<td>GCGAGGACTAAAATCCAGGC</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>AACCAGGCCACACAATAGC</td>
</tr>
<tr>
<td>KIF20B</td>
<td>Forward</td>
<td>TCAAGACACTCGATTCCAA</td>
</tr>
<tr>
<td></td>
<td>Reverse 1</td>
<td>TTCTGGGTGTAGCATCTTCTT</td>
</tr>
<tr>
<td></td>
<td>Reverse 2</td>
<td>CCTATGAAAGGAGGCAAT</td>
</tr>
<tr>
<td>USP9X</td>
<td>Forward</td>
<td>CGAAGTCCCATGGGTATTA</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>TGAAGGAGCACATAGCCACAG</td>
</tr>
<tr>
<td>TOP2B</td>
<td>Forward</td>
<td>CCAGATTGATGGTGGCTGT</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>AACCAGGCCACACAATAGC</td>
</tr>
<tr>
<td>Gene</td>
<td>Direction</td>
<td>Sequence (5' → 3')</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>KDM1B</td>
<td>Forward</td>
<td>AACCCAGTAGCATAATGTGTGAA</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>CGGGAGTTAGCAGAGTGTGG</td>
</tr>
<tr>
<td>C1orf43</td>
<td>Forward</td>
<td>AGGAGCTGGGGGCTCATACTTT</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>GTGAATGTCGTGCTGGTGAT</td>
</tr>
<tr>
<td>MRFAP1</td>
<td>Forward 1</td>
<td>TTGGTTCCGGCTTTACCTCTGC</td>
</tr>
<tr>
<td></td>
<td>Forward 2</td>
<td>CTTGTCCGTTGGCTTCTCTG</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>GCAGAAACTGCTGGAAATCC</td>
</tr>
<tr>
<td>EXOSC10</td>
<td>Forward</td>
<td>CCACATCTCCAGCCTCATTT</td>
</tr>
<tr>
<td></td>
<td>Reverse 1</td>
<td>CCAGAGAAGCCTGGGTGTTGA</td>
</tr>
<tr>
<td></td>
<td>Reverse 2</td>
<td>ACAGACCCAGCCATCGTTA</td>
</tr>
<tr>
<td>UBQLN1</td>
<td>Forward</td>
<td>TGCATTGCCTCTAGGTTTTGA</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>GTCTGCCCCCTACATGAGAA</td>
</tr>
<tr>
<td>R(\otimes)M(^{+})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAYSD1</td>
<td>Forward 1</td>
<td>GAGGAATGGGCTGTTCCAT</td>
</tr>
<tr>
<td></td>
<td>Forward 2</td>
<td>CAGAGCAATTAACAGAAACTGG</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>AGCAGCGAGCAGCTCTAAAGG</td>
</tr>
<tr>
<td>RNF185</td>
<td>Forward</td>
<td>GACAGCAGTCTTGAGTGTGGA</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>TGTGGCATAATCCCCAAAGG</td>
</tr>
<tr>
<td>GPBP1</td>
<td>Forward</td>
<td>CTTTAGCTGCAGGTTGTGGA</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>AGGTGGAGCAGGCTTTTAG</td>
</tr>
<tr>
<td>CASP6</td>
<td>Forward</td>
<td>GGTTTTCCACCCAGGCTGTG</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>GCGTTTGGCTGCAATGAG</td>
</tr>
<tr>
<td>ZNF692</td>
<td>Forward</td>
<td>GGAGGAGCTGTGGAACCTCAA</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>AGGACTCAAGAGGCGAGGTT</td>
</tr>
<tr>
<td>R(\otimes)M(^{-})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UBIAD1</td>
<td>Forward</td>
<td>CCGCTTGCCCTACTACCTG</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>TTTCTGGGAGCTTGAGCCT</td>
</tr>
<tr>
<td>CTC1</td>
<td>Forward</td>
<td>ATGGACAGTGGTGACAG</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>TGGGCACTTTTTTCTGTTC</td>
</tr>
<tr>
<td>POLR2J4</td>
<td>Forward 1</td>
<td>CAGGGACAGCCTACTGATGA</td>
</tr>
<tr>
<td>R(^{+})M(^{-})</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

91
APPENDIX B. PRIMERS CX4945 EXPERIMENT

<table>
<thead>
<tr>
<th>Gene</th>
<th>Forward 2</th>
<th>Reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AGGCCTCAGGAAACTGACAA</td>
<td>GGACACCAAGGTACCCAATG</td>
</tr>
<tr>
<td>LRRC37BP1</td>
<td>Forward</td>
<td>AGCAGCGCCTCAAGCTCCAC</td>
</tr>
<tr>
<td></td>
<td>Reverse</td>
<td>GAGTCAGCCTCAAGCTCCAC</td>
</tr>
</tbody>
</table>

**Table B.1:** Primers used in the PCR validation of the detected splicing events for the comparison between microarrays and RNA-Seq
ADDITIONAL FIGURES MICROARRAYS VS RNA-SEQ

Additional figures for the comparison between microarrays and RNA-Seq. The figures correspond to the Z-value of the matched events in both technologies and \( \Psi \) estimates obtained with EventPointer.

Figure C.1: Scatterplot of the splicing z-value provided by EventPointer for RNA-Seq and HTA 2.0. The colors represent the groups: \( R^- M^- \) (black), \( R^- M^+ \) (blue), \( R^+ M^- \) (green) and \( R^+ M^+ \) (red). The correlation is 0.89.
Figure C.2: Estimated PSI (for RNA-Seq -red-, microarrays -blue- and PCR image analysis -green-), PCR bands, the reference HTA transcriptome and the alternative paths of the ACADM, TOP2B and KIF4. These genes were found to be significant by both technologies. All of them were considered validated. The arrows below the alternative paths depict the location of the primers and the numbers shown are the expected lengths of the PCR bands with the selected primers.
Figure C.3: Estimated PSI (for RNA-Seq -red- and PCR image analysis -green-), PCR bands, the reference HTA transcriptome and the alternative paths of the OSGIN2, ACTR10, USP9X, KIF20B and CPSF2. These genes were detected only by RNA-seq. All of them were considered validated (the PCR for ACTR10 was noisy though).
Figure C.4: Estimated PSI (for RNA-Seq -red-, microarrays -blue- and PCR image analysis -green-), PCR bands, the reference HTA transcriptome and the alternative paths of the SAYSD1, RNF185, ZNF692, CASP6 and GPBP1 genes. These genes were found to be significant only by RNA-seq but were detected also by HTA. All of them were considered validated (i.e. the prediction by RNA-seq is correct). CASP6 includes some additional bands non-detected by EventPointer.
Figure C.5: Estimated PSI (for microarrays -blue- and PCR image analysis -green-), PCR bands, the reference HTA transcriptome and the alternative paths of the KDM1B, C1orf41, MRFAP1, EXOSC10 and UBQLN1 genes. These genes were detected only by HTA. In EXOSC10 a new unpredicted band appears. All of them were considered validated (i.e. the prediction by HTA is correct).
APPENDIX C. ADDITIONAL FIGURES MICROARRAYS VS RNA-SEQ

Figure C.6: Estimated PSI (for RNA-Seq -red-, microarrays -blue- and PCR image analysis -green-), PCR bands, the reference HTA transcriptome and the alternative paths of the UBIAD1, CTC1, POLR2J4, LRRC37BP1 and WNK1 genes. These genes detected by both technologies and found to be significant only by HTA. UBIAD1 and POLR2J4 bands does not distinguish treatment and control as clearly as other samples. In two of the genes (UBIAD1, LRRC37BP1) unpredicted long bands (probably intron retentions) appeared in the PCRs. CTC1, LRRC37BP1 and WNK1 were considered to be validated.
EventPointer: an effective identification of alternative splicing events using junction arrays

Juan P. Romero¹,², Ander Muniategui¹,², Fernando J. De Miguel³, Ander Aramburu¹,², Luis Montuenga³,⁴,⁵, Ruben Pio⁵,⁶ and Angel Rubio¹,²*

Abstract

Background: Alternative splicing (AS) is a major source of variability in the transcriptome of eukaryotes. There is an increasing interest in its role in different pathologies. Before sequencing technology appeared, AS was measured with specific arrays. However, these arrays did not perform well in the detection of AS events and provided very large false discovery rates (FDR). Recently the Human Transcriptome Array 2.0 (HTA 2.0) has been deployed. It includes junction probes. However, the interpretation software provided by its vendor (TAC 3.0) does not fully exploit its potential (does not study jointly the exons and junctions involved in a splicing event) and can only be applied to case–control studies. New statistical algorithms and software must be developed in order to exploit the HTA 2.0 array for event detection.

Results: We have developed EventPointer, an R package (built under the aroma.affymetrix framework) to search and analyze Alternative Splicing events using HTA 2.0 arrays. This software uses a linear model that broadens its application from plain case–control studies to complex experimental designs. Given the CEL files and the design and contrast matrices, the software retrieves a list of all the detected events indicating: 1) the type of event (exon cassette, alternative 3′, etc.), 2) its fold change and its statistical significance, and 3) the potential protein domains affected by the AS events and the statistical significance of the possible enrichment. Our tests have shown that EventPointer has an extremely low FDR value (only 1 false positive within the tested top-200 events). This software is publicly available and it has been uploaded to GitHub.

Conclusions: This software empowers the HTA 2.0 arrays for AS event detection as an alternative to RNA-seq: simplifying considerably the required analysis, speeding it up and reducing the required computational power.

Keyword: Alternative splicing, Junction microarrays, Protein domains

Background

Alternative Splicing (AS) has been shown to be a key factor in cellular processes such as development and differentiation as well as in different pathologies, including cancer [1–3]. AS has been studied using Exon arrays and, more recently, using RNA-seq and junction arrays [4].

The first array that made use of junction probes was based on Agilent technology and included approximately 125,000 junction probes, but lacked exon probes [5]. These were included in a later version of the array [6]. In 2010, Oryzon Genomics, in collaboration with our group, introduced an array based on Agilent technology, covering 7,958 genes with a total number of 115,318 exon probes and 105,141 junction probes [7]. This new array made use of a massive number of control probes (as much as 20 % of the array) to ensure the proper normalization of the measurements.

In 2008, Affymetrix presented the Human Junction Array (HJAY). It was their first experimental array with
exon and junction probes [8]. This microarray included approximately 6 million probes, comprising ~315,000 exons and ~260,000 junctions. Each exon and junction was interrogated by 8 different probes. The probes of this array were selected using RefSeq, ExonWalk and Ensembl annotations. Two years earlier, ExonHit introduced the Splicearray, also using the Affymetrix technology (although they also provide now a version using Agilent technology). However, the use of ExonHit arrays is not as widespread as the use of standard ones from Agilent or Affymetrix.

In 2011, Affymetrix, together with Stanford University, designed the custom Glue Grant Human Transcriptome Array (GG-H array) [4]. And in 2013, Affymetrix launched the GeneChip® Human Transcriptome Array 2.0 (HTA 2.0), a more up to date catalogue of the HGAY and GG-H arrays. The HTA 2.0 array interrogates a total of 1,048,904 exons or exon clusters with more than 6.3 million probes (approximately 10 probes per exon) and more than 339,000 exon-exon junctions with more than 1.3 million probes (around 4 probes per junction).

In a previous work [9], we developed an algorithm to detect AS cassette events. It was applied to both the HGAY and the Oryzon arrays. HGAY array clearly outperformed the other platform and had a validation rate for top-ranked events of nearly 100%. The results proved that the Affymetrix platform is a good option to detect AS events. The main problem of this array is that it was unsupported upon their release.

Only the Transcriptome Analysis Console (TAC) 3.0 software offered by Affymetrix and AltAnalyze [10] are the available options to analyze HTA 2.0 and HGAY arrays. FIRMA [11], using CDFs generated by Brainarray [12], can be applied to extract and summarize expression but the junction probes would be missing from this analysis pipeline.

The main drawback of the TAC and FIRMA approaches, is that neither of them combines the information provided by the junction probes with the corresponding exon probes in the event under study. For example, in the detection of a cassette event, it is not sufficient to detect the altered expression values of exon probes. In addition to that, the flanking junctions must behave coherently and the skipped junction must have a negative correlation. A similar argument can be made for other AS events. On the other hand, AltAnalyze using the ASPIRE algorithm [13] combines the information of two probesets to get a figure of merit for each event. For example, in a cassette exon, AltAnalyze would provide three figures of merit: one corresponding to the probeset of the exon and the junction that skips it and two more combining the flanking junctions with the junction that skips the exon. Even though, this approach is intrinsically better, it would be still desirable to have a single figure of merit per event.

There is an algorithm (MADS+) developed to exploit combined information from exons and junctions [14]. However, its development has been discontinued and it cannot be applied to the HTA 2.0 platform. Furthermore, most of the algorithms developed to detect AS events (including MADS+) are limited to the analysis of case-control studies. Its extension to more complex experimental designs, such as case-control studies with paired samples or time-course studies, is non-trivial.

Here, we present EventPointer, an algorithm to detect AS events using the HTA 2.0 platform. It can be applied to any experiment using appropriately configured design and contrast matrices. EventPointer is based on the limma framework in Bioconductor [15].

### Results

#### Implementation

Since EventPointer is described in-depth in the Methods section, here we briefly describe its main characteristics. Using Affymetrix junction arrays (the software accepts either HGAY and HTA 2.0), after mapping the probes against the Ensembl transcriptome (Ensembl v. 74) [16], the splicing graph for each gene is generated and EventPointer identifies and classifies the different AS events that can be detected with these arrays. The different classes are alternative 3’, alternative 5’, alternative first exon, alternative last exon, cassette exon, retained intron, mutually exclusive exons and complex events (none of the above). These steps are specific for the arrays and do not need to be repeated for each experiment. The output of this pipeline is a CDF file that groups the probes into probesets that reflect the splicing events. These different steps are depicted in Figs. 1 and 2.

For each experiment, the output of the hybridization (i.e. the CEL files returned by the scanner) must be summarized using any standard pipeline (in our case, RMA). Following probeset summarization (according to the previously prepared CDF file), EventPointer uses the provided design and contrast matrices to compute the statistical significance of AS events. By construction, the AS events can always be validated using standard PCR with at most three primers. We describe here the results of applying the EventPointer algorithm to the HTA 2.0 arrays on an experiment with 27 samples.

The R package EventPointer is available for download at Github. It includes the CDF file needed for aroma.affymetrix pre-processing pipeline and the necessary functions to obtain the statistical results.

EventPointer also enables visualization using IGV. EventPointer generates an output file that can be loaded to IGV to display the events (the reference and both paths) as well as the location of the probes for each of
the paths. Figure 3 includes a screen capture of an event displayed in IGV.

Within the vignettes that accompany EventPointer we have included some secondary analyses (clustering and functional enrichment analysis of the identified targets) that illustrates the potential of the provided tools.

Mapping annotations
The number of events that can be theoretically detected by EventPointer for the HTA 2.0 and HJAY arrays are 70,886 and 37,069, respectively. Figure 4 shows the number of events using the different canonical categories.

As Fig. 4 shows, HTA 2.0 includes more splicing events than HJAY for all event types. The majority of
Fig. 2 Description of the detection of events from the splicing graph using EventPointer. The SG is extended (every single node is split into two to ensure that the probesets are mapped only to edges) and corrected to force. The splicing graph is interrogated to find events and detected events are classified into one of the three possible groups (event a is a cassette and b an alternative last event). Each of these groups are further subdivided into the different event types checking the length of the junctions.

Fig. 3 Visualization of EventPointer prediction in IGV. Image generated by EventPointer as displayed in IGV.
detected event types correspond to complex events (i.e. events that cannot be included in any of the standard categories). As the transcriptome annotation improves, it also becomes more complex. Previously we developed ExonPointer, an algorithm to detect AS cassette exon events [17]. EventPointer extends ExonPointer more than ten-fold, taking the 6515 cassette events in the HTA 2.0 array to more than 70,000 splicing events of any type. In addition to that, EventPointer provides a more rigorous definition of a cassette event and can be run on the HTA 2.0 arrays. We will focus on the results with HTA 2.0 since it interrogates more exons and junctions than HJAY and is a more recent and stable development of Affymetrix.

Transcriptome data
The performance of the developed algorithms was tested in an experiment where the splicing factor SRSF1 was knocked down using siRNA on the A549 lung adenocarcinoma cell line. This cell line was obtained from the American Type Culture Collection (ATCC). The experiment included three conditions: cells treated only with the vehicle of the transfection (Lipofectamine 2000, Invitrogen), cells treated with scramble siRNA (i.e. a sequence that will not lead to the specific degradation of any cellular mRNA) and cells transfected with a siRNA that targets SRSF1. Each condition had three biological replicates that, in turn, were hybridized three times (9 hybridizations). The efficiency of this inhibition has been stated elsewhere [9].

The splicing factor SRSF1 [18] has a pleiotropic effect: it regulates splicing (enhancing or decreasing certain isoforms), regulates nonsense-mediated mRNA decay, has a role on RNA metabolism (translation), RNA protein binding, has a potential oncogenic role in cancer, regulates the mitosis among other processes [9, 19–21]. We have performed a differential expression analysis of this experiment. We have included the main results in the Additional file 1. The categories of gene ontology enriched are concordant with the aforementioned functions (Additional file 1).

Determining differential splicing events
An AS event is considered to be differentially spliced if the concentration of the isoforms mapped to either paths of the event (for example, in a cassette event the path that skips the path and the path that includes the exon) are differentially expressed in opposite directions (i.e. if one of the isoforms is overexpressed, the other must be underexpressed). For this particular experiment, the contrast matrix compares the knock down samples using siRNA with the samples of the cells transfected with scramble siRNA. In the comparison of these two conditions (with siRNA and with scramble siRNA), out of the theoretical 70886 events, 3718 showed a \( p \)-value < 10\(^{-3}\) (approx. 5 % of the events).

Each of the paths in an event (the word path is used here to refer to any of the two configurations of each event; see Figs. 1 and 2, in a cassette event one path skips the exon and the other one includes it) is annotated with the domains (if any) included in the Ensembl database. In some cases, the domain is disrupted in one of the paths compared with the other. We performed a statistical analysis of the enrichment of the domains that may affected by the AS events. The statistics of this analysis are described in the Methods section. In brief, we performed a Wilcoxon test between the isoforms that contain a domain. The results of this analysis are summarized in Table 1.

The enrichment analysis illustrates one of the potentials of this analysis. The laminins are proteins of the extracellular matrix. The modification of their domains is known to induce a pro-invasive phenotype [22].
Several of the motifs are related to RNA binding: RNA recognition motif (a.k.a. RRM, RBD, or RNP domain). These results are coherent with the GO enrichment analysis (Additional files 1 and 2) in which several of the enriched categories are related with RNA modification.

To our knowledge, the relation between SRSF1 and the WD40 domain was unknown. The underlying common function of all WD40-repeat proteins is the coordination of multi-protein complex assemblies, where the repeating units serve as a rigid scaffold for protein interactions [23]. In addition to this, the RNA domains are also targets of PRPF8, another splicing factor. IBR, as well as WD40, domains are related to ubiquitin ligase complexes [24].

We have included a few of the domains that were underexpressed after the knock-down of SRSF1. It is important to point out that the statistical significance is much smaller. An intriguing result is that the tRNA synthetases domains are overexpressed and underexpressed depending on their class.

**Table 1** Enrichment of domains in the list of AS events regulated by SRSF1

<table>
<thead>
<tr>
<th>Domains (PFAM)</th>
<th>P-value</th>
<th>Z-value</th>
<th>Description</th>
<th>Main characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF000052 PF00008 PF06009 PF000053</td>
<td>[4.15E-14, 4.81E-08]</td>
<td>[7.56, 5.46]</td>
<td>Laminin B (Domain IV), Laminin Domain I and II, Laminin EGF-like (Domains III and V)</td>
<td>Laminin family</td>
</tr>
<tr>
<td>PF08373 PF00076</td>
<td>[4.61E-10, 1.11E-07]</td>
<td>[6.23, 5.31]</td>
<td>RAP domain, RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)</td>
<td>RNA binding</td>
</tr>
<tr>
<td>PF000400</td>
<td>1.19E-09</td>
<td>6.08</td>
<td>WD domain, G-beta repeat</td>
<td>WD domain</td>
</tr>
<tr>
<td>PF02780 PF02779</td>
<td>[2.30E-09, 2.05E-06]</td>
<td>[5.97, 4.75]</td>
<td>Transketolase, C-terminal domain and pyrimidine binding domain</td>
<td>Transketolase</td>
</tr>
<tr>
<td>PF04836 PF05004</td>
<td>[2.38E-09, 1.66E-08]</td>
<td>[5.97, 5.64]</td>
<td>Interferon-related protein conserved region and interferon-related developmental regulator (IFRD)</td>
<td>Interferon related</td>
</tr>
<tr>
<td>PF00038 PF04732</td>
<td>[5.97E-08, 9.48E-08]</td>
<td>[5.42, 5.34]</td>
<td>Intermediate filament protein and head (DNA binding) region</td>
<td>Intermediate filament</td>
</tr>
<tr>
<td>PF02893</td>
<td>9.48E-08</td>
<td>5.34</td>
<td>GRAM domain</td>
<td>GRAM domain</td>
</tr>
<tr>
<td>PF08337</td>
<td>1.35E-07</td>
<td>5.27</td>
<td>Plexin cytoplasmic RasGAP domain</td>
<td>Plexin</td>
</tr>
<tr>
<td>PF02096</td>
<td>2.18E-07</td>
<td>5.18</td>
<td>60Kd inner membrane protein</td>
<td>Cytochrome oxidase</td>
</tr>
<tr>
<td>PF00152</td>
<td>6.05E-07</td>
<td>4.99</td>
<td>tRNA synthetases class II (K and N)</td>
<td>tRNA synthetases</td>
</tr>
<tr>
<td>PF00884</td>
<td>3.38E-06</td>
<td>4.65</td>
<td>Sulfatase</td>
<td>Sulfatase</td>
</tr>
<tr>
<td>PF01485</td>
<td>3.63E-06</td>
<td>4.63</td>
<td>IBR domain</td>
<td>Protein Quality control</td>
</tr>
<tr>
<td>PF03950 PF00749</td>
<td>2.85E-04, 1.37E-03</td>
<td>[-3.63, -3.46]</td>
<td>tRNA synthetases class I (E and Q), anti-codon binding domain and catalytic domain</td>
<td>tRNA synthetases</td>
</tr>
<tr>
<td>PF08725</td>
<td>2.95E-04</td>
<td>-3.62</td>
<td>Integrin beta cytoplasmic domain</td>
<td>Cell-cell receptors</td>
</tr>
</tbody>
</table>

The domains are sorted by its statistical significance. If several domains share similar properties, are grouped into a single row. The upper part of the table shows overexpressed domains. The second part of the table (divided by a thick line) shows underexpressed domains. The statistical significance was larger for overexpressed than for underexpressed domains.
ranked according to their p-values. In turn, the p-value represents the "differential opposite expression" of each of the isoforms interrogated by an event. The statistical significance of the events (and their ranking) was very different across the different AS types. All the five top-ranked "cassette" and "complex" events were within the overall top-20 ranked events. In contrast, only one out of the five top-ranked "alternative 3' site" was found within the top-150 ranked events and the top ranked "mutually exclusive exons" was in position 500. These results are summarized in Table 2.

The 17 top ranked events appeared within the top-5 of any of the categories and were validated. In all of them the validation was positive although, in a few cases, the PCR-band analysis (using ImageJ) did not pass statistical significance. In Fig. 5 we show some of the PCR results. Additional file 3 shows the results for all the events. Primer sequences are also included in the additional material (Additional file 4).

Comparison with other software
Affymetrix Transcriptome Analysis Console (TAC 3.0) and AltAnalyze are the only available software to detect alternative splicing events using HTA 2.0. The main features from each of them are briefly explained below and the comparison with EventPointer is discussed in the latter paragraphs.

TAC 3.0
The software from Affymetrix is publicly available for download and provides the user with different tools to go beyond simple differential gene expression analysis. Some of the options are gene pathway networks, miRNA and target gene interactions and alternative splicing events identification. It works only in Windows operating system.

Given the corresponding cel files, the software automatically runs the analysis based on the options provided by the user.

TAC uses Splicing Index (SI) [25] as a measure to detect alternative splicing events. Briefly, the SI of a probe-set compares two ratios:

$$SI = \left\{ \frac{\text{Probeset signal Cond. 1}}{\text{Overall signal of the gene in Cond. 1}} , \frac{\text{Probeset signal Cond. 2}}{\text{Overall signal of the gene in Cond. 2}} \right\}$$

If the SI is close to one, then the behavior of the probe-set is coherent with the behavior of the gene. If it is much larger or much smaller than 1, then the probe-set signal may indicate the presence of alternative splicing.

TAC applies several filters based on expression levels (adjustable by the user) prior to the calculation of the splicing index for any given PSR (probe selection region, Table 2

<table>
<thead>
<tr>
<th>Ranking</th>
<th>HGNC symbol</th>
<th>Type</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MYCBP2</td>
<td>Complex Event</td>
<td>OK</td>
</tr>
<tr>
<td>2</td>
<td>KIF23</td>
<td>Cassette Exon</td>
<td>OK</td>
</tr>
<tr>
<td>3</td>
<td>AC024560.3</td>
<td>Cassette Exon</td>
<td>OK</td>
</tr>
<tr>
<td>4</td>
<td>FBX022</td>
<td>Cassette Exon</td>
<td>OK</td>
</tr>
<tr>
<td>5</td>
<td>SRSF3</td>
<td>Complex Event</td>
<td>–</td>
</tr>
<tr>
<td>6</td>
<td>SUPT16H</td>
<td>Alternative Last Exon</td>
<td>OK</td>
</tr>
<tr>
<td>7</td>
<td>HMBOX1</td>
<td>Alternative First Exon</td>
<td>OK</td>
</tr>
<tr>
<td>8</td>
<td>ACAD11</td>
<td>Complex Event</td>
<td>OK</td>
</tr>
<tr>
<td>9</td>
<td>NCOR1</td>
<td>Cassette Exon</td>
<td>OK</td>
</tr>
<tr>
<td>10</td>
<td>AUP1</td>
<td>Retained Intron</td>
<td>OK</td>
</tr>
<tr>
<td>11</td>
<td>IFT27</td>
<td>Alternative Last Exon</td>
<td>OK</td>
</tr>
<tr>
<td>12</td>
<td>GALNT10</td>
<td>Alternative Last Exon</td>
<td>–</td>
</tr>
<tr>
<td>13</td>
<td>PARD3</td>
<td>Alternative Last Exon</td>
<td>–</td>
</tr>
<tr>
<td>14</td>
<td>PRMT2</td>
<td>Complex Event</td>
<td>OK</td>
</tr>
<tr>
<td>15</td>
<td>HORMAD1</td>
<td>Cassette Exon</td>
<td>OK</td>
</tr>
<tr>
<td>16</td>
<td>ANAPC7</td>
<td>Alternative Last Exon</td>
<td>OK</td>
</tr>
<tr>
<td>17</td>
<td>OGT</td>
<td>Complex Event</td>
<td>OK</td>
</tr>
<tr>
<td>18</td>
<td>MSL3</td>
<td>Alternative First Exon</td>
<td>OK</td>
</tr>
<tr>
<td>19</td>
<td>NTSC</td>
<td>Alternative 5'</td>
<td>OK</td>
</tr>
<tr>
<td>20</td>
<td>ALG2</td>
<td>Alternative First Exon</td>
<td>OK</td>
</tr>
<tr>
<td>21</td>
<td>MSL3</td>
<td>Alternative First Exon</td>
<td>OK</td>
</tr>
<tr>
<td>22</td>
<td>BAIAP2L1</td>
<td>Alternative First Exon</td>
<td>OK</td>
</tr>
<tr>
<td>23</td>
<td>HIST1H2AC</td>
<td>Alternative 5'</td>
<td>–</td>
</tr>
<tr>
<td>24</td>
<td>DDX52</td>
<td>Retained Intron</td>
<td>–</td>
</tr>
<tr>
<td>25</td>
<td>TMEM214</td>
<td>Retained Intron</td>
<td>–</td>
</tr>
<tr>
<td>26</td>
<td>GASPB1</td>
<td>Alternative 5'</td>
<td>OK</td>
</tr>
<tr>
<td>27</td>
<td>EIF3B</td>
<td>Retained Intron</td>
<td>–</td>
</tr>
<tr>
<td>28</td>
<td>LAMP1</td>
<td>Retained Intron</td>
<td>–</td>
</tr>
<tr>
<td>29</td>
<td>LM07</td>
<td>Alternative 3'</td>
<td>OK</td>
</tr>
<tr>
<td>30</td>
<td>SCAMP3</td>
<td>Alternative 5'</td>
<td>X</td>
</tr>
<tr>
<td>31</td>
<td>UHRF2</td>
<td>Alternative 5'</td>
<td>X</td>
</tr>
<tr>
<td>32</td>
<td>COP53</td>
<td>Alternative 3'</td>
<td>X</td>
</tr>
<tr>
<td>33</td>
<td>SLS9A8</td>
<td>Alternative 3'</td>
<td>–</td>
</tr>
<tr>
<td>34</td>
<td>FLNA</td>
<td>Alternative 3'</td>
<td>X</td>
</tr>
<tr>
<td>35</td>
<td>C2Iorf58</td>
<td>Alternative 3'</td>
<td>OK</td>
</tr>
<tr>
<td>36</td>
<td>CALU</td>
<td>Mutually Exclusive</td>
<td>X</td>
</tr>
<tr>
<td>37</td>
<td>CCT6P1</td>
<td>Mutually Exclusive</td>
<td>–</td>
</tr>
<tr>
<td>38</td>
<td>ST20</td>
<td>Mutually Exclusive</td>
<td>OK</td>
</tr>
<tr>
<td>39</td>
<td>ACO2</td>
<td>Mutually Exclusive</td>
<td>X</td>
</tr>
<tr>
<td>40</td>
<td>KIAA0100</td>
<td>Mutually Exclusive</td>
<td>X</td>
</tr>
</tbody>
</table>

The first column shows the ranking of the event in EventPointer according to the p-value. The last column states if the validation was positive: a (OK) mark is shown if the validation is significant in PCR, a (~) mark is shown if differences in the PCR can be observed but the statistical significance is above 0.05. A (X) mark is shown if the event was not validated either because there was no differential splicing, no alternative splicing or no expression at all.

The statistical significance of the events (and their ranking) was very different across the different AS types. All the five top-ranked "cassette" and "complex" events were within the overall top-20 ranked events. In contrast, only one out of the five top-ranked "alternative 3' site" was found within the top-150 ranked events and the top ranked "mutually exclusive exons" was in position 500. These results are summarized in Table 2.

The 17 top ranked events appeared within the top-5 of any of the categories and were validated. In all of them the validation was positive although, in a few cases, the PCR-band analysis (using ImageJ) did not pass statistical significance. In Fig. 5 we show some of the PCR results. Additional file 3 shows the results for all the events. Primer sequences are also included in the additional material (Additional file 4).

Comparison with other software
Affymetrix Transcriptome Analysis Console (TAC 3.0) and AltAnalyze are the only available software to detect alternative splicing events using HTA 2.0. The main features from each of them are briefly explained below and the comparison with EventPointer is discussed in the latter paragraphs.

TAC 3.0
The software from Affymetrix is publicly available for download and provides the user with different tools to go beyond simple differential gene expression analysis. Some of the options are gene pathway networks, miRNA and target gene interactions and alternative splicing events identification. It works only in Windows operating system.

Given the corresponding cel files, the software automatically runs the analysis based on the options provided by the user.

TAC uses Splicing Index (SI) [25] as a measure to detect alternative splicing events. Briefly, the SI of a probe-set compares two ratios:

$$SI = \left\{ \frac{\text{Probeset signal Cond. 1}}{\text{Overall signal of the gene in Cond. 1}} , \frac{\text{Probeset signal Cond. 2}}{\text{Overall signal of the gene in Cond. 2}} \right\}$$

If the SI is close to one, then the behavior of the probe-set is coherent with the behavior of the gene. If it is much larger or much smaller than 1, then the probe-set signal may indicate the presence of alternative splicing.

TAC applies several filters based on expression levels (adjustable by the user) prior to the calculation of the splicing index for any given PSR (probe selection region,
i.e. subexon) or junction. It also classifies (some) of the events according to the standard categories (exon cassette, intro retention, etc.). And, for the ones that are classified, TAC includes a “splicing score” (SS). This value is based on “how well the data fits into pre-defined splicing patterns”. Besides, both PSRs and their related junctions all contribute to an event score” that is bounded between 0 and 1 (from the TAC manual). No additional information is provided on how this score is computed or how the events are classified and can be considered as an experimental method. Only around half of the events are given a SS. In our case, 23/40 (57 %) of the validated events included the SS.

We sorted all the events found by TAC according to the absolute value of the logarithm of the SI (negative values indicate the lack of the exon in the case samples).
See Additional file 5 with a list of top TAC predicted events.

AltAnalyze

This is an open-source software developed in the Nathan Salomonis lab at Cincinnati Children’s Hospital Medical Center and the University of Cincinnati. This project began in the laboratory of Bruce Conklin at the Gladstone Institutes. It can be downloaded from their webpage (http://www.altanalyze.org/) and it can be run in different operating systems such as Windows, Mac OS and Ubuntu. As stated by the developers: “requires no advanced knowledge of bioinformatics programs or scripting”.

The software enables analysis of data produced by both conventional and splicing sensitive microarrays (e.g. exon and junction arrays) as well as RNA-Seq data and the pipeline enables a complete analysis that includes identification of alternative splicing events and differential expression as well as different functional annotations of the genomic regions identified as alternatively spliced.

For the detection of alternative splicing events, AltAnalyze uses two different algorithms to measure splicing events: Splicing Index (as in TAC) and analysis of splicing by isoform reprocity (ASPIRE). See Additional file 5 with a list of top AltAnalyze predicted events. AltAnalyze provides the user the option to set different thresholds to filter genes and AS events depending on the expression levels.

The ASPIRE algorithm is used when two probesets (A and B) measure the exclusion and inclusion of an exon respectively. It provides a score similar to a fold change, bounded between -1 and 1, where negative values indicate that the expression in the probeset (A or B) in experimental group is higher than the control group. A single splicing event can have (and usually does) several inclusion indexes per event. Each of them correspond to the pairwise comparisons between the probesets that include and exclude the event respectively.

In order to identify an event as statistically significant, AltAnalyze relies on three different values: ratio of inclusion, ratio of exclusion and $\delta I$. The first ratios measure the proportion of the inclusion or exclusion of an exon with the mean expression of the whole gene. Both ratios must be in opposite directions to continue the analysis. The $\delta I$ value measures the difference between both the inclusion and exclusion ratios. As a default value, any event must have a $\delta I$ above 0.2. A detailed explanation of the algorithm can be found in the AltAnalyze user manual. EventPointer takes into account all the exons and junctions involved in the alternative splicing event to give its statistical significance. The results for EventPointer are normalized using the probeset in the reference path and using the whole gene for AltAnalyze and TAC. On the other side, TAC (using SI) and AltAnalyze (using ASPIRE) provide up to 3 statistical values for each event (skipping junction vs flanking junction 1, skipping junction vs exon and skipping junctions vs flanking junction 2). Both AltAnalyze and TAC are focused on case–control studies. EventPointer can be applied to any experimental design that can be described with a design and a contrast matrix.

Using the “GRanges” R package [26], we matched the events expressed and detected by the three algorithms (EventPointer, AltAnalyze and TAC). As explained before, an event detected by EventPointer can be matched to more than one element of either TAC or Analyze, as a result we kept the unique events that where matched in order to create the Venn diagram depicted in Fig. 6. This diagram shows the expressed events for all of them. These events do not necessarily show differential usage of the variants.

As already mentioned, TAC provides a different SI for each of the probesets that interrogate an event. We summarized these values taking the most significant SI for each event. Table 3 provides the ranking of the top ten events detected by EventPointer in both AltAnalyze and TAC.

Some comparison examples of top-ranked events

The events that can be clustered into three groups: events with high ranking in the three algorithms, events

---

**Fig. 6** Venn diagram of common events identified by EventPointer, AltAnalyze and TAC. The diagram displays the total number of events detected by each algorithm (in parenthesis) and the corresponding values for the intersections.
with high ranking in EventPointer and low ranking in AltAnalyze and TAC and, finally, events low ranked to EventPointer and high ranked in the other two algorithms. The compared events, displayed in TAC, are shown in Fig. 7.

Within the first group, KIF23 was ranked 2nd, 11th and 3rd in EventPointer, AltAnalyze and TAC, respectively. The corresponding statistical values in each algorithm are: 1.00e-58 (EventPointer pvalue), 0.5523 (ASPIRE δ) and 16.67 (TAC SI). The three methods identify this event as high-ranked. Figure 7a shows the TAC window at the particular region for this event.

HMBOX1 is also an event identified as high-ranked by the three algorithms. The corresponding rankings are 7th, 6th and 98th in EventPointer, AltAnalyze and TAC, respectively. Even though the ranking in TAC is not as low as in the other two methods, the PCR validation confirms the alternative splicing event. Figure 7b shows the TAC window for this event.

In the second group, NCOR1 was ranked 9th in EventPointer while AltAnalyze and TAC rank the event in positions 365 and 5757, respectively. This event shows either the highest SI or δ value when compared to the other events found for the same gene. The event was validated by PCR and shows a significant p-value (5.69e–38) in EventPointer. Figure 7c shows the event in TAC.

ACAD11 is ranked as the 8th event in EventPointer while AltAnalyze and TAC rank the event in positions 396 and 3371, respectively. It shows a similar behavior as NCOR1. Figure 7d displays TAC window for this event.

In the second group, NAA50 was ranked 10926th, 6th and 148th in EventPointer, TAC and AltAnalyze, respectively. The reason is that this event is backed up by only one junction probe-set. Since EventPointer finds no coherence with the other paths of the event, the ranking is low (Fig. 7e). This event was validated using PCR. The PCR results are included in the supplementary material (Additional file 6).

Table 3  Ranking of the top-ranked events according to EventPointer and their ranking positions in the other algorithms (EP: EventPointer, AA: AltAnalyze)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Event type</th>
<th>Genomic position</th>
<th>EP</th>
<th>AA</th>
<th>TAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYCBP2</td>
<td>Complex Event</td>
<td>13:77673020-77695630</td>
<td>1</td>
<td>12</td>
<td>383</td>
</tr>
<tr>
<td>KIF23</td>
<td>Cassette Exon</td>
<td>15:69713986-69714774</td>
<td>2</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>AC024580.3</td>
<td>Cassette Exon</td>
<td>3:197348575-197350253</td>
<td>3</td>
<td>9</td>
<td>113</td>
</tr>
<tr>
<td>FBOX22</td>
<td>Cassette Exon</td>
<td>15:76196323-76205608</td>
<td>4</td>
<td>55</td>
<td>1392</td>
</tr>
<tr>
<td>SRSF3</td>
<td>Complex Event</td>
<td>6:36566626-36568967</td>
<td>5</td>
<td>2095</td>
<td>280</td>
</tr>
<tr>
<td>SUPT16H</td>
<td>Alternative Last Exon</td>
<td>14:21837979-21852105</td>
<td>6</td>
<td>NA</td>
<td>1026</td>
</tr>
<tr>
<td>HMBOX1</td>
<td>Alternative First Exon</td>
<td>8:28902878-28904970</td>
<td>7</td>
<td>6</td>
<td>98</td>
</tr>
<tr>
<td>ACAD11</td>
<td>Complex Event</td>
<td>3:132297677-132298402</td>
<td>8</td>
<td>396</td>
<td>3371</td>
</tr>
<tr>
<td>SUPT16H</td>
<td>Alternative Last Exon</td>
<td>17:16052765-16055312</td>
<td>9</td>
<td>365</td>
<td>5757</td>
</tr>
<tr>
<td>AUP1</td>
<td>Retained Intron</td>
<td>2:74754863-74755133</td>
<td>10</td>
<td>NA</td>
<td>10501</td>
</tr>
</tbody>
</table>

Discussion

We have presented a complete pipeline to detect AS events using HJAY and HTA 2.0 arrays. The main advantages of this method over the Splicing Index or ASPIRE are that: 1) it can be applied to any experimental design (by providing the corresponding design and contrast matrices) and not only to case–control studies, 2) it exploits the redundancy of all the junction probes involved in an alternative splicing event and 3) it labels all the events according to the different categories. All the suggested events can be validated using standard PCR by construction.

EventPointer is event-focused instead of transcript-focused [27, 28]. It estimates the statistical significance
Fig. 7 (See legend on next page.)
of AS events without estimating the concentration of the underlying transcripts. For each event, EventPointer identifies the type of event (cassette, alternative start site, alternative donor, etc.) and provides its statistical significance according to the design and contrast matrices given by the user. EventPointer also generates a graphical output using IGV [29] to make the identification and validation of the events easier.

There are, however, some events that cannot be identified with EventPointer. For example, a couple of isoforms with different transcription start sites in which one of them is included in the other. These types of events are also very difficult to detect using RNAseq or PCR since there is no a specific sequence in the second isoform to design a primer.

In this work, we consider that there is a differential splicing event if the isoforms in the associated paths change their expression in opposite directions. Although this statistical test can miss some AS events, the selected events have a clear change on their expression. These changes, usually, have more biological relevance than other subtler cases in which only one isoform – usually weakly expressed - changes its expression across the conditions.

Once the Affymetrix CEL files are analyzed (i.e. background corrected, normalized, and summarized using the standard procedures), the statistical analysis to detect the AS events is very fast by using the limma package. Using a standard Intel i5 processor, the analysis requires about 10 s. The hardware requirements are modest (a low-end desktop computer with 4GB of RAM is sufficient). The whole enrichment analysis takes only fractions of a second. This is an advantage compared to the requirements on storage, computational power and memory of RNA-seq analysis. In addition to that, the proposed methodology is very reliable: only one false positive was found within the top-200 tested events.

The validation rate is quite high. Our thought is that the reasons for this are on the one hand EventPointer gets a single statistics that combines the information within all the probesets interrogating the event and, on the other hand, uses a proper reference region for each event: most of the methods select a number of probesets (all of them, only the one of constitutive exons, core or quasi-constitutive exons, etc.) that are used for all the events in the gene. However, in our case, this reference is different for each event.

One of the key parts of the analysis of alternative splicing is to provide a biological interpretation of the splicing events, i.e. what is the difference between the isoforms expressed in a condition specific manner. EventPointer provides the protein family domains that are affected on each of the splicing events. It also performs an enrichment analysis (event-based) to identify which are the domains that are significantly over or underexpressed in the condition under study. Even though in its present form, EventPointer only provides information on the PFAM domains we are actively developing it and we have some alpha versions of the software that provide annotation for other domain databases such us Pirsf, Superfamily, Smart, Prosite, or Interpro. In addition to protein domains, there are other interesting biological data that could be inferred. For example in [30], Ray et al. identify the binding motifs of several hundreds of RNA binding proteins and the potential binding sites in the human genome. Using this information is possible to predict which are the splicing factors that are driving the differential usage of isoforms. Another potential improvement would be to identify miRNA binding sites and check if the splicing pattern causes skipping these binding sites and therefore, the corresponding miRNAs may be no longer regulating the expressed isoforms. This functionality is already offered by AltAnalyze and we expect to include it in the near future. Finally, in its present form, EventPointer works for the HJAY and the HTA2 arrays. We are actively developing the annotation to apply EventPointer to Mouse and Rat junction arrays (MTA-1 and RTA-1).

EventPointer could be extended to RNA-seq by building up the corresponding splicing graph and, in fact, we are currently working on this extension. AltAnalyze can also be applied to RNAseq and has been used for example in [31]. In order to apply EventPointer to RNAseq data, the splicing graph must be constructed based on the sequencing reads and/or the annotation (for microarrays we only use the annotation since the sequences of the probes are fixed). The statistics to perform the analysis must also be changed: instead of using a linear model on the log of the data signal, other methods such as voom [32], or edgeR [33] should be used to take into account the discrete nature of the reads.

There is one potential advantage of EventPointer when extending it to RNAseq experiments. In microarrays, the affinities of the probes are difficult to predict and, usually, they are considered to be unknown. Any algorithm (including EventPointer) get results by implicitly estimating the affinities given the data. The role of affinities...
in RNAseq is played by the length of the regions that originated the reads (the length of the exon where they came from, for example). EventPointer can be adapted to use the length of these regions to perform the statistics or leave them as unknown and guess them (as it happens with the arrays). This second approach does not require the general assumption of considering uniform coverage of the reads (that is known not to be true).

Conclusions
We have developed EventPointer software to detect AS events using Affymetrix arrays. It has a high validation rate and shows its effectiveness to detect AS events using Affymetrix splicing-sensitive arrays. In addition to that, its connection with IGV makes it very convenient to validate the results using PCR.

This technique can be used on its own, but also to cross-validate RNA-seq experiments. In addition to that, it provides a statistical analysis of the usage of protein domains and provides a single statistic per event that, to our knowledge, is a novel development for analysis at the transcript/event level.

Methods
Sample preparation and PCR validation
Downregulation of SRSF1, expression analyses and microarray hybridization were done as previously described [9]. Samples from two independent experiments were used for validation of splicing events by endpoint PCR. Briefly, RNA was retro-transcribed using PrimeScript RT reagent Kit (Takara). PCR was performed using PCR Master Mix (Promega) using the following program: 94 °C 2 min; 30 cycles at 94 °C 30 s, 57 °C 30 s, 72 °C 30 s; and 72 °C 10 min. Primers used are shown in Additional file 1.

Selection of the events
Mapping
The probes included in HTA 2.0 array from Affymetrix are mapped against the human transcriptome (Ensembl 74) using Bowtie 2.0. Multimapping probes, those that map against more than 3 genes, are removed since they are considered to be non-informative. This mapping is used to build the CDF file, i.e. to group the probes according to sets of probes and according to the events (for further information see Additional file 8).

Construction of the splicing graph
The Splicing Graph (SG) [35] is a directed graph used to represent the structure of a gene (see Fig. 5a and b). Here, nodes are denoted as subexons (contiguous region of the genome that belong to the same set of transcripts) and in the SG two nodes are connected by an edge if both subexons are contiguous in at least one isoform. Two additional nodes, start and end nodes, are added to the graph and nodes in 5′ (3′) locus of any isoform are connected to them.

The nodes of the SG are duplicated (a and b nodes are used to represent each node) so that both exon and junction probes are represented exclusively by edges in the graph (Fig. 1c). Junction probes connect ‘b’ nodes with ‘a’ nodes and probes mapped against exons connect ‘a’ nodes with ‘b’ nodes.

Pruning and recovering of the splice graph
The SGs (one per gene) are very complex and include many edges not supported by any probe on the array. The SG has been pruned to retain only edges with probes mapped to them. In some cases there are edges with no probes mapped to them to ensure the coherence of the graph, i.e. (see Fig. 2 upper panel) any node of the SG is connected with the start and end nodes. For example, although there is no probe mapping against junction E7–E8, the edge from E7 to E8 is not removed to ensure that E7 is connected with the end node. In the contrary, there is no probe mapping against junctions E3–E4 and E4–E5 nor exon E4, and thus they can be removed without affecting the coherence of the SG. For further analysis see Additional file 1.

Finding the splicing events
In this work, a Splicing Event is described as a triplet [PR, P1, P2] of subgraphs. These subgraphs are composed of sets of edges and nodes that share the following characteristics: 1) the flow traversing any of the edges of each triplet is identical, and 2) the flow traversing any edge in PR is the sum of the flows traversing P1 and P2. P1 is assigned to be the set of edges with the largest genomic length in the transcriptome and P2 to the shortest one (i.e. in a cassette event, P1 is the path that includes the skipping exon). The detection of the events can be automated using graph theory. The values of the fluxes of all the edges are calculated ensuring that two flows will not be equal by random. Then, the edges with common flow values are grouped in sets. And finally, triplets of sets with one of its flows summing up the other two are grouped together (subgraphs 1 and 2). These last correspond to events. For further information, see Additional file 1.

Labeling the type of splicing event
The splicing events can be classified into 7 major categories: cassette exons, alternative 3′ donor site, alternative 5′ donor site, intron retention, alternative last exon, alternative first exon and mutually exclusive exons. Any
event not classified into these is considered to be a complex event. The labelling of an event is carried out according to the structure of the subgraph of the event (PR, P1 and P2) (see Fig. 2 lower panel).

**Summarizing the events**

Each of the events is composed of a triplet of probesets: the probes mapped to path 1, path 2 and the reference. For a specific event, there are isoforms that are not mapped to any of the paths in the event. However, if an isoform hits the event then, by construction, the isoform can be mapped to PR and only to either P1 or P2.

Within the analysis, we consider that the signal of a probe on an Affymetrix array is the product of the affinity of the probe and the sum of the concentrations of the isoforms interrogated by the probe. Therefore, if \( y_{ij} \) is the logarithm of signal of the probe \( i \) in condition \( j \) then,

\[
y_{ij} = \log(Affy_i \tau_j) + \varepsilon_{ij} = \log(Affy_1) + \log(\tau_j) + \varepsilon_{ij}
\]

where \( Affy_i \) is the affinity of the probe, \( \tau_j \) is the sum of the concentrations of the transcripts interrogated by probe \( i \), and \( \varepsilon_{ij} \) is an error term.

In EventPointer (as in most other methods that use Affymetrix technology), the values of the probe signals within a probeset are summarized to a single value using the median polish algorithm [36]. We have assumed that the model for a single probe is also valid for the summarized value of the probeset (i.e. the signal in the probeset is proportional to the concentration of the isoforms interrogated by the probeset).

Let us consider a differential splicing event that consists of two alternative stop sites (i.e. transcript end), interrogated by 6 probes. Probes 1 and 2 belong to PR, Probes 3 and 4 to P1, and probes 5 and 6 to P2 (see Fig. 8a). Let us assume that this event is measured in two different conditions, normal (N) and tumor (T). There will be 12 different measurements that, after summarization, are converted to the 6 values \( y_{RN}, y_{1N}, y_{2N}, y_{RT}, y_{1T}, y_{2T} \) that correspond to the summarized signals of PR, P1 and P2 in the normal and tumor samples,

![Fig. 8 Steps of the statistical Analysis: a toy example](image)

- **a)** Summarize probe intensity values per path
- **b)** Extend design and contrast matrices
- **c)** Solve the model using Limma

**Statistics using the extended contrast matrix**

\[
\beta_2 + \beta_3 + \beta_5
\]
respectively (see Fig. 8a). For this example, the signal $y_{2N}$—dropping the error terms—(i.e. the probeset in the reference path in the normal sample that comprises probes 1 and 2) is

$$y_{2N} = \log(Aff_{2N}(t_{1N} + t_{2N})) = \log(Aff_{2N}) + \log(t_{1N} + t_{2N})$$

(2)

The signal of $y_{2T}$ (probeset P2 –probes 5 and 6- in the tumor sample), we have,

$$y_{2T} = \log(Aff_{2T}) + \log(t_{2T})$$

(3)

Since the probeset resides in path 2, it only measures the second isoform.

Construction of the design matrix for alternative splicing detection

A convenient formalism to describe an experiment is using the contrast and design matrices. This section describes how to build these matrices to detect differential splicing events.

Let $Q$ be the following 3x3 auxiliary matrix:

$$Q = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

(4)

Its usefulness will be shown later. Let $D$ be the design matrix of the experiment. The proposed design matrix $D_{DRS}$ to detect splicing events is:

$$D_{DRS} = [D \otimes Q],$$

(5)

where $\otimes$ is the matrix Kronecker product. We have included a simple example to illustrate the concepts. The corresponding design ($D$) and extended design matrices ($D \otimes Q$) for the aforementioned example are shown in Fig. 6b. The design matrix includes an intercept term for all the samples and an increment for the tumor sample. If this experiment only studied gene expression, the contrast will be the design and $t_{2T}$ will be the second isoform.

Values of the factors for the lineal model and their interpretation

<table>
<thead>
<tr>
<th>Value of Beta</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>$\log(Aff_{2N}) + \log(t_{1N} + t_{2N})$ No special interest</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>$\log(\frac{t_{1N}}{t_{1T}}) + \log(\frac{t_{2N}}{t_{2T}})$ No special interest</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$\log(\frac{t_{1N} + t_{1T}}{t_{1N} + t_{2N}})$ Difference of the logarithms of the fold change of isoform 2 in both conditions. AS present if different from 0.</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>$\log(\frac{t_{1N} + t_{2N}}{t_{1N} + t_{2T}})$ Difference of the logarithms of the fold change of isoform 1 in both conditions. AS present if different from 0.</td>
</tr>
<tr>
<td>$\beta_4 + \beta_5$</td>
<td>$\log(\frac{t_{2N}}{t_{1N} + t_{2N}})$ Difference of the logarithms of the fold change using relative concentrations of isoform 1 in both conditions. AS present if different from 0.</td>
</tr>
<tr>
<td>$\beta_4 + \beta_5$</td>
<td>$\log(\frac{t_{2N}}{t_{1N} + t_{2T}})$ Difference of the logarithms of the fold change using relative concentrations of isoform 2 in both conditions. AS present if different from 0.</td>
</tr>
<tr>
<td>$\beta_4 + \beta_5$</td>
<td>$\log(\frac{t_{2N}}{t_{1N}})$ Logarithm of the fold change of isoform 1.</td>
</tr>
<tr>
<td>$\beta_4 + \beta_5$</td>
<td>$\log(\frac{t_{2N}}{t_{2T}})$ Logarithm of the fold change of isoform 2.</td>
</tr>
</tbody>
</table>
In the PCR validation, the contrast that provided the best performance was the combination of the fold changes of both isoforms (i.e. $β_1 + β_2 + β_3$ and $β_1 + β_4 + β_5$ in Table 4) plus the requirement that the fold-changes have opposite directions, i.e. if isoform 1 significantly increases its expression, isoform 2 must significantly decrease its expression and visa versa. Therefore, if this test requires that the detected AS events show a significant change of the expression both paths and this change must be in opposite direction.

In order to compute this contrast, we summed up the p-values (one-tailed) for both contrasts. If the null hypothesis holds, the expected null distribution is triangular from 0 to 2 with the peak at 1, and the summation of the p-values must be close to 0 or close to 2 for genes with differential AS. Using this triangular distribution, it is possible to assign an overall p-value to their sum. We preferred this combination rather than the classical Fisher method since in the latter a single good p-value yields a good summary p-value for the event. Using this approach, both p-values must be close to zero or one in order to generate a significant overall p-value.

All the statistics had been implemented in an R package (available at Github), which depends on the limma method to get the statistical significance. Given the “standard” design matrix, $D$ and the corresponding contrast matrix, $C$, the software internally computes the design and contrast matrices of the events $D_{DRS}$ and $C_{DRS}$. Moreover, given a contrast matrix for the experiment, the contrast matrix to detect the splicing events is given by:

$$C_{DRS} = \begin{bmatrix} C_0 \otimes [1 \ 1 \ 0] \\ C_0 \otimes [1 \ 1 \ 1] \end{bmatrix}.$$

in which each row represents the contrasts $β_3 + β_4$ and $β_3 + β_4 + β_5$ i.e. for each given contrast, the differential usage of both pathways is tested and summarized.

By construction, each contrast is split into two different contrasts (to test the differential expression of both isoforms) and afterwards, they are summarized and returned to the user.

**Expression filter**

The contrast previously described is very sensitive. If one of the paths of an event is not expressed and the other one is, EventPointer would assign a significant p-value to the event. This would result in a large number of false positives due to the lack of expression.

In order to avoid this problem, EventPointer allows the user to filter the events to ensure that all the paths are expressed above a fixed threshold. For every path, the algorithm gets the maximum value of expression from all the samples. The maximum values for the references are used to set the threshold. The user provides a quantile that will set the threshold.

Once the threshold is selected, an event will be considered as expressed if the maximum value of expression for all the paths is above the threshold previously set.

**Domain expression analysis**

Using the Ensembl database is possible to relate each of the paths P1 and P2 in the events with the presence of protein domains within them. For each domain, it is possible to know in which paths P1 and P2 is included. In order to state the statistical significance of the events, we performed a Wilcoxon test paired for each event (P1 and P2). This algorithm is done using sparse matrices and turns to be very efficient (less than 1/10th of second for all the domains in the Pfam database).

### Additional files

- **Additional file 1:** Vignette of the use of the aroma.affymetrix framework to perform a gene expression analysis (SRSF1 knock-down analysis) with HTA 2.0 data. (PDF 264 kb)
- **Additional file 2:** Vignette of the use of the EventPointer on the study of SRSF1 using HTA 2.0 data. (PDF 246 kb)
- **Additional file 3:** PCR images and relative concentrations of the isoforms based on the PCR image. The red bar corresponds to the shorter isoform and the blue bar to the longer isoform. For the event in KIAA0100, it was not possible to get PCR results, and thus the image is not included. (TIF 1465 kb)
- **Additional file 4:** Table with primers used for validation and sizes of the expected splicing variants. (DOCX 24 kb)
- **Additional file 5:** Results for top-ranked events according to TAC 3.0. (DOCX 17 kb)
- **Additional file 6:** PCR images and relative concentrations of the isoforms detected by TAC 3.0. (PDF 177 kb)
- **Additional file 7:** PCR images and relative concentrations of the isoforms based on the PCR image of gene DNM2. The grey bar corresponds to the shorter isoform and the white bar to the longer isoform. (PNG 92 kb)
- **Additional file 8:** Supplementary methods. Detection and classification of events. (DOCX 63 kb)

### Abbreviations

AS, alternative splicing; ASR, analysis of splicing by isoform repository; ATCC, American type culture collection; EP, eventpointer; HJAY, human junction array; HTA, human transcriptome array; SL, splicing index; SS, splicing score; TAC, transcriptome analysis console

### Funding

This work was partially supported by Affymetrix. This funding is gratefully acknowledged. It was also funded by Instituto de Salud Carlos III-Fondo Europeo de Desarrollo Regional (P120/00661/0040 and P111.000801), and Asociación Española Contra el Cáncer (AECO). Scientific Foundation (GCB14-2178). F. de Miguel is supported by a predoctoral fellowship from the Instituto de Salud Carlos III (Ministerio de Economía y Competitividad).

### Availability of data and material

The microarray data from this study have been submitted to Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo) under the accession number GSE76602 and the corresponding CDF file under the accession number GPL21339.
Conception and design: AM, RP, AR, and LMM. Development of methodology: JPR, AA, RP, AR, and LMM. Acquisition of data (sample preparation, acquired and managed cell lines, hybridization of the samples, provided facilities, etc.): FM, RP, and LMM. Development of software: JPR, AM, AA, AR. Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): JP, AM, AA, FIA, RP, AR, and LMM. Writing, review, and/or revision of the manuscript: JPR, AM, AA, FM, RP, AR, and LMM. Study supervision: RP, AR, and LMM. All authors read and approved the final manuscript.

Competing interests
This work has been partially funded by Affymetrix.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
1. CEIT, Parque Tecnológico de San Sebastián, Paseo Mikeletegi 48, 20009 San Sebastián, Gipuzkoa, Spain.
2. Tecnun, University of Navarra, Pº de Manuel Lardizábal 13, 20018 Donostia-San Sebastián, Gipuzkoa, Spain. Program in Solid Tumors and Biomarkers, CIMA, University of Navarra, Avda. Pío XII 55, E-31008 Pamplona, Navarra, Spain.
3. Department of Histology and Pathology, University of Navarra, Pamplona, Spain.
4. IDIBAIN, Navarra Institute for Health Research, Recinto de Complejo Hospitalario de Navarra, C/Nurileama 3, 31008 Pamplona, Navarra, Spain.
5. Department of Biochemistry and Genetics, University of Navarra, Pamplona, Spain.

Received: 27 February 2016 Accepted: 7 June 2016

Published online: 17 June 2016

References


Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at www.biomedcentral.com/submit
Poster presented in the Splicing 2016 congress held in Caparica, Portugal in September 2016. The poster, along with a 5 minute shotgun presentation received the award "Excellent Shotgun Communication Award"
EventPointer: Software to identify alternative splicing events using junction arrays or RNASeq data

Juan P. Romero¹ and Angel Rubio²
¹CEIT and Tecnun, University of Navarra, San Sebastian, Spain.
²Corresponding author: email promero@ceit.es; arubio@ceit.es

Abstract
Detection of events of alternative splicing (using either microarrays or RNASeq) is the first step to uncover their relationship with a disease or pathological state. Event-based methods are appropriate for stratification and functional analysis of the protein domains (or regulatory features) involved in the splicing. A number of tools has been developed to detect event-based differential splicing (such as DESeq, rMATS, Spladder, SpliceGrapher, JuncBase in the RNASeq world and AliAnalyze for “both” worlds). Most of these tools must be applied to case-control studies. Here we present an algorithm that applied to either microarrays or microarrays, quantifies, characterizes and provides the statistical significance of splicing events.

We have validated a number of splicing events from both microarrays and RNA-sequencing using standard PCR. The PCR validations show an extremely low FDR for microarrays and RNASeq and concordance between each technologies.

The algorithm simplifies the complete analysis and provides users to run the software in an user friendly interface that can be run in a desktop computer.

Splicing Graph Generation

The Splicing Graph (SG) [2] is a directed graph used to represent the structure of a gene. Here, nodes are denoted as subexons (contiguous region of the genome that belong to the same set of transcripient) and in the SG two nodes are connected by an edge if both subexons are contiguous in at least one isoform. Two additional nodes, start and end nodes, are added to the graph and nodes in S (3, 3′ locus of any isoform are connected to them.

For RNASeq, Eventpointer uses BAM files to build the SG based on the coverages and junctions found in all the sample. [2] Eventpointer uses BAM packages [3]. For microarrays, the SG is created using the reference transcriptome and the probes in the array.

Results & Visualization

The output of the program is a table with information for every detected event:
Gene, Genomic Position, Event Type, Pvalue, Zvalue and PSI.

We used standard PCR to validate the five top-ranked AS events within each of the eight different types of events (i.e. cassette, mutually-exclusive, complex etc.). In total, 40 different events were tested.

The visualization using IGV[5] eases interpretation and allows researchers to upload prs for PCR validation.

Discussion & Conclusion

We have developed EventPointer as an event-focused algorithm to detect AS

- The algorithm can be applied to complex experimental designs
- PCR results show high validation rate
- Usable for Affymetrix Arrays (HTA v2 & Clariom) and RNASeq experiments

The visualization using IGV[5] eases interpretation and allows researchers to upload prs for PCR validation.

References


[85] MISO, *What is sashimi plot?*  
(Accessed on 17/06/2017).


http://cnx.org/contents/XrodM20j@2/Eukaryotic-Transcription.  
(Accessed on 12/03/2017).

[88] OXFORDNANOPORE, *Highlights of clive g brown's technical update*.  
(Accessed on 16/06/2017).

[89] ———, *MinION*.  
https://nanoporetech.com/publications/2015/03/26/  
(15/06/2017).


[95] N. Raghavachari, J. Barb, Y. Yang, P. Liu, K. Woodhouse, D. Levy, J. Christopher, P. J. Munson, and G. J. Kato, A systematic comparison and evaluation of high density exon arrays and rna-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease, BMC medical genomics, 5 (2012), p. 28.


134


