UNIVERSITY OF VERONA

DEPARTMENT OF

Neurosciences, Biomedicine and Movement Sciences

GRADUATE SCHOOL OF

Health and Life Sciences

DOCTORAL PROGRAM IN

Applied Life and Health Sciences

WITH THE FINANCIAL CONTRIBUTION OF

Cariverona

Cycle (1° year of attendance): 29° (2014)

TITLE OF THE DOCTORAL THESIS

A SVM-based method to classify RBM20 affected and not affected exons

S.S.D.: MED-03

Coordinator: Prof. Giovanni Malerba

Tutor: Prof. Giovanni Malerba

Doctoral Student: Dott.ssa Anna Dal Molin

A SVM-based method to classify RBM20 affected and not affected exons

Anna Dal Molin

Tesi di Dottorato

Verona, 04/07/2017

ISBN

Sommario

Mutazioni della proteina RBM20 sono state recentemente riportate causare la Cardiopatia Umana Dilatativa (DCM) (Brauch et al., 2009, Li et al., 2010). DCM è la causa principale di arresto cardiaco e mortalità nel mondo (Jefferies and Towbin, 2010). Complessivamente, il 25–50% dei casi di DCM sono mutazioni familiari e causative, descritte in più di 50 geni che codificano principalmente per componenti strutturali dei cardiomiociti.

RBM20 appartiene alla famiglia delle proteine SR e proteine associate alle proteine SR, che si assemblano nello spliceosoma prendendo parte allo splicing del pre-mRNA. RBM20 è espressa principalmente nel muscolo striato, con i livelli più alti nel cuore (Guo et al., 2012). A causa del suo coinvolgimento nella DCM, RBM20 è stata molto studiata per svelare il suo meccanismo d'azione ed i suoi target sull'RNA (Guo et al., 2012, Li et al., 2013). Guo e colleghi, tramite un'analisi del trascrittoma su ratto e uomo, hanno riportato un gruppo di 31 geni che presentano uno splicing dipendente da RBM20 (Guo et al., 2012). Più recentemente, Maatz e colleghi hanno riportato un ulteriore gruppo di 18 geni di ratto ed osservato che le sequenze di RNA riconosciute da RBM20 molto probabilmente si trovano nei 400 nucleotidi fiancheggianti gli esoni il cui splicing alternativo è regolato da RBM20 (Maatz et al., 2014). Tuttavia, sia la sequenza di RNA suggerita per essere riconosciuta da RBM20, che la sua sovra rappresentazione nelle regioni fiancheggianti degli esoni affetti, rimangono predittori deboli per identificare i geni che presentano uno splicing alternativo regolato da RBM20.

Lo scopo di questo lavoro è stato, perciò, quello di caratterizzare, attraverso un approccio bioinformatico, i motivi di sequenza degli esoni il cui splicing alternativo era regolato da RBM20, con l'intento di migliorare la predizione dei geni (esoni) influenzati da RBM20.

È stata fatta un'analisi di espressione differenziale per selezionare il gruppo di esoni regolati da RBM20; un ulteriore gruppo di esoni è stato ricavato da dati

presenti in letteratura (Maatz et al., 2014).

Una Macchina a Vettori di Supporto (SVM) è stata usata valutando più tipi di elementi genetici che possono legarsi nelle regioni fiancheggianti dei nostri esoni target. Si è scelto un approccio SVM per classificare gli esoni affetti e non affetti da RBM20, ma altri algoritmi di *machine learning* avrebbero potuto essere utilizzati; l'approccio SVM, comunque, è fra quelli più usati. Dalle analisi, il nostro modello è risultato discriminare bene gli esoni regolati da quelli non regolati da RBM20.

Da un punto di vista biologico e funzionale, questo approccio ci aiuta ad individuare nuovi geni candidati associati a malattie che dipendono da una disregolazione di RBM20.

Questo studio ha fornito informazione aggiuntiva sulla regolazione di RBM20 degli esoni target, basandosi non solo sul sito di legame all'RNA, ma anche su altri elementi genetici associati al sito di legame. Inoltre, abbiamo proposto il primo modello basato su un algoritmo di SVM per la classificazione degli esoni regolati e non regolati da RBM20.

Abstract

Mutations of RNA binding motif protein 20 (RBM20) have been recently reported to cause Human dilated cardiomyopathy **(**DCM**)** (Brauch et al., 2009, Li et al., 2010). DCM is the major cause of heart failure and mortality around the world (Jefferies and Towbin, 2010). Overall, 25–50% of DCM cases are familiar and causative mutations which have been described in more than 50 genes encoding mostly for structural components of cardiomyocytes.

RBM20 belongs to the family of the SR and SR-related RNA binding proteins which assemble in the spliceosome taking part in the splicing of pre-mRNA. RBM20 is mainly expressed in striated muscle, with the highest levels in the heart (Guo et al., 2012). Due to its involvement in DCM, RBM20 was studied a lot to unveil its mechanism of action and its RNA targets (Guo et al., 2012, Li et al., 2013). Guo and colleagues reported a set of 31 genes showing a RBM20 dependent splicing from a whole transcriptome analysis in rats and humans (Guo et al., 2012). More recently, Maatz and colleagues reported an additional set of 18 rat genes and observed that RNA sequences recognized by RBM20 are likely to be located in the 400 nucleotides flanking the exons whose alternative splicing is regulated by RBM20 (Maatz et al., 2014). However, both the suggested RNA sequence which is recognized by RBM20 and its over-representation over the flanking regions of affected exons remain poor predictors to target genes presenting splicing events regulated by RBM20.

The aim of this work was, thus, to characterize, through a bioinformatic approach, the sequence motifs of the exons whose alternative splicing was affected by RBM20**,** in order to ameliorate the prediction of the genes (exons) affected by RBM20.

A differential expression analysis was performed to select the dataset of RBM20 affected exons; a further dataset was retrieved from literature data (Maatz et al., 2014).

A Support Vector Machine (SVM) approach evaluating more kinds of genetic

elements binding in the flanking regions of our target exons was used. A SVM method was chose to classify RBM20 affected and not affected exons, but other machine learning algorithms could have been used as well; however, SVM is among the most commonly used ones. From the analyses, our model resulted to well discriminate RBM20 affected from not affected exons.

From a biological and functional point of view, this approach helps us to target novel candidate genes associated to diseases depending on a dysregulation of RBM20.

This study provided additional information about RBM20 regulation of target exons, based not only on the RNA binding site, but also on other genetic elements associated to the binding site. Furthermore, we proposed the first model based on a SVM algorithm for the classification of RBM20 affected and not affected exons.

Index

Sommario **5**

Abstract **7**

Index **9**

List of figures **12**

List of tables **13**

1 Introduction **17**

1.1 Dilated cardiomyopathy 17

1.1.1 A general overview 17

1.1.2 Causes of disease 18

1.1.3 Genetic causes of disease 18

1.2 Splicing 20

1.2.1 Introduction 20

1.2.2 The spliceosome 21

1.2.3 General splicing mechanism 21

1.2.4 Alternative splicing 23

1.2.5 Alternative splicing and disease 26

1.2.6 Alternative splicing at genome level 26

1.3 RNA binding motif protein 20 28

1.3.1 SR proteins 28

1.3.2 RBM20 28

1.4 RNA sequencing 31

1.4.1 Introduction 31

1.4.2 Transcriptome sequencing 32

1.4.3 Experimental parameters 35

1.4.4 RNA-Seq data analysis workflow 36
1.4.5 Differential gene expression 40
1.4.6 Allele-specific expression 41
1.4.7 Expression quantitative trait loci 42
1.5 Support Vector Machine 43
1.5.1 Machine learning 43
1.5.2 Supervised learning 44
1.5.3 Unsupervised learning 45
1.5.4 Reinforcement learning 45
1.5.5 Classification 46
1.5.6 Linear Support Vector Machine 47
1.5.7 Not Linear Support Vector Machine 50
1.5.8 Features selection 52
1.5.9 Testing 53
2 Materials and methods 55
2.1 Public RNA-Seq data of RBM20 mutants 56
2.2 Quality control 56
2.3 Reads alignment 57
2.4 Exons counting and differential expression 57
2.5 Differentially expressed exons and extraction of target sequences regions 58
2.6 Searching for binding sites patterns 59
2.7 Searching for additional genetic elements 60
2.8 Features selection 62
2.9 Support Vector Machine 63
3 Results 65
3.1 The dataset A1 of RBM20 affected exons 65

3.2 The dataset A2 of RBM20 affected exons 66

3.3 Searching for RBM20 RNA binding site in the target sequences 66

3.4 Searching for additional genetic elements in RBM20 affected exons 69

3.4.1 Searching for regulatory and transcriptional known motifs 69

3.4.2 Searching for transposable elements 70

3.4.3 Searching for sequence patterns more frequent in cases than in control exons 75

3.4.4 Estimating the frequency of nucleotides and dinucleotides 76

3.4.5 Getting of exons length 77

3.5 Features selection 77

3.6 Support Vector Machine analysis 79

4 Discussion 85

5 Conclusions 91

6 Bibliography 93

List of figures

Figure 1. Differences of a normal heart from an heart with DCM. 18

Figure 2. Schematic representation of the splicing process. 20

Figure 3. Schematic representation of factors implied in splicing repression and activation. 22

Figure 4. Example of alternative splicing. 24

Figure 5. The 5 main kinds of alternative splicing. 25

Figure 6. Loss-of-function mutations in the splicing factor RBM20 cause DCM via the pathological splicing of cardiac proteins. 29

Figure 7. Workflow of transcriptome sequencing for RNA-Seq experiments. 33

Figure 8. Workflow of RNA-Seq data analysis. 37

Figure 9. Graphical representation of RNA-Seq reads mapping on exon-exon junctions. 38

Figure 10. Example of linearly separable classes. 48

Figure 11. Linear classification with SVM. 49

Figure 12. Mapping in a higher dimensional features space. 51

Figure 13. Workflow of the analyses performed in this study. 55

Figure 14. Exonic and intronic regions under investigation. 59

Figure 15. Sequence logo for rat RBM20 RNA binding pattern, given at nucleotide resolution. 67

List of tables

Table 1. Public RNA-Seq sample data used for the transcriptome analyses. 56

Table 2. Datasets of rat exons retrieved. 58

Table 3. DEXSeq differential analyses significant results for rat and human samples. 65

Table 4. Descriptive statistic about substrings length and score (max, mean and median). 68

Table 5. Overall number of regulatory and transcriptional known motifs observed in the 430 or 230 nucleotides upstream and downstream regions of the target exons. 70

Table 6. Number of transposable elements (TEs) observed in the 430 nucleotides or 230 nucleotides upstream and downstream regions of the target exons. 71

Table 7. Total number of nucleotides (nt) of interspersed repeats. 72

Table 8. Number of transposable elements (TEs). 73

Table 9. Fisher exact test results on the number of nucleotides (for Total interspersed repeats) or exon flanking sequences (for SINEs) containing the selected elements, given as p-value and odds-ratio (OR) for the comparisons A1-N1 and A2-N1. 75

Table 10. Patterns from the enrichment analyses (DREME software). 76

Table 11. Descriptive statistics about exons length. 77

Table 12. Number of features found at each step of the features selection. 78

Table 13. Distribution of the 215 shared features among the different genetic elements analysed. 79

Table 14. AUC value of the model and accuracy of the prediction for the SVM classification. 80

Table 15. AUC value of the model and accuracy of the prediction for the SVM classification on 80 random exons of the dataset A1 and on the dataset A2. 81

Table 16. Accuracy of the prediction for the SVM classification on 80 random exons of the dataset A1 and on the dataset A2. 81

Table 17. Accuracy of the prediction for the SVM classification on 80 random exons of the dataset A1 and on the dataset A2. 82

1 Introduction

1.1 Dilated cardiomyopathy

1.1.1 A general overview

Dilated cardiomyopathy (DCM) is a condition in which the heart becomes enlarged and cannot pump blood efficiently. The decreased heart function can affect the lungs, liver and other body systems. DCM is one of the cardiomyopathies, a group of diseases that affect primarily the heart muscle; in particular, it is the most common form of non-ischemic cardiomyopathy. Different cardiomyopathies have different causes and affect the heart in different ways. In DCM a portion of the myocardium is dilated, often without any obvious cause. Left or right ventricular systolic pump function of the heart is impaired, leading to progressive heart enlargement via ventricular hypertrophy and ventricular dilation, a process called ventricular remodelling (Jameson et al., 2005) (Figure 1). As the heart chambers dilate, the heart muscle doesn't contract normally and can't pump well blood, the heart becomes weaker and heart failure can occur.

Common symptoms of heart failure include shortness of breath, fatigue and swelling of the ankles, feet, legs, abdomen and veins in the neck. Dilated cardiomyopathy also can lead to heart valve problems, arrhythmias (irregular heartbeats) and blood clots in the heart.

DCM is the major cause of heart failure and significant source of mortality and morbidity worldwide (Jefferies and Towbin, 2010), it occurs more frequently in men than in women, in African-Americans than in Caucasians (Coughlin et al., 1993), and it is most common from the ages of 20 to 60 years (Robbins et al., 2003), even though sometimes it occurs in children too.

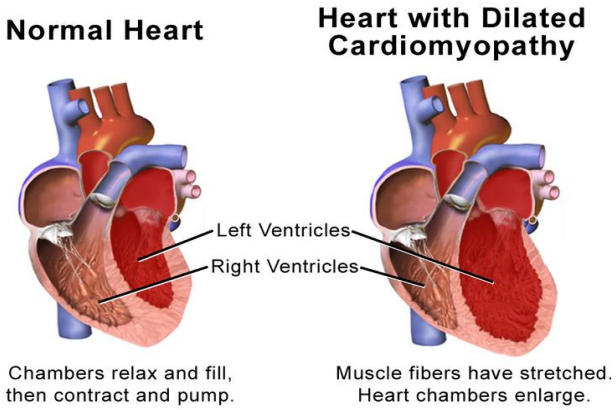

Figure 1. Differences of a normal heart from an heart with DCM (Blausen.com staff. "Blausen gallery 2014").

1.1.2 Causes of disease

Although in many cases there is no apparent cause, DCM is probably the result of the damage to the myocardium produced by a variety of toxic, metabolic or infectious agents. It may be due to fibrous change of the myocardium from a previous myocardial infarction or it may be the consequence of acute viral myocarditis (Mitchell et al., 2007), possibly mediated through an immunologic mechanism (Martino et al., 1994). Other common causes include: Chagas disease due to *Trypanosoma cruzi* (this is the most common infectious cause of DCM in Latin America), pregnancy (DCM occurs late in gestation or from several weeks to months post-partum as a peri-partum cardiomyopathy; it is reversible in half of cases (Mitchell et al., 2007)), alcohol abuse (although the cause-and-effect relationship with alcohol alone is debated (Mitchell et al., 2007)), thyroid disease, muscular dystrophy, tuberculosis (Agarwal et al., 2005), autoimmune mechanisms (San Martin et al., 2002), and recently also an extremely high occurrence of premature ventricular contractions (extrasystole) (Shiraishi et al., 2002; Belhassen, 2005).

1.1.3 Genetic causes of disease

Estimates suggest that the 25–50% of DCM cases are familiar; indeed, causative

mutations have been described in more than 50 genes encoding mostly structural components of cardiomyocytes. Among the most prevalent, there are mutations in lamin A/C and beta-myosin heavy chain, each accounting for up to 10% of cases of familiar DCM in some series (Parks et al., 2008; Villard et al., 2005).

The disease is genetically heterogeneous, but the most common form of transmission is an autosomal dominant pattern (Mitchell et al., 2007). Autosomal recessive, X-linked and mitochondrial inheritance of the disease were also found (Schönberger and Seidman, 2001). Some relatives of those affected by DCM have preclinical, asymptomatic heart-muscle changes (Mahon et al., 2005).

Recently, a genetic linkage analysis in families with autosomal dominant DCM led to the discovery of heterozygous missense mutations in an arginine/serine-rich (RS) domain of RNA binding motif protein 20 (RBM20). Subsequent profiling of a large cohort of idiopathic DCM patients identified additional missense mutations, all of which cluster within an RS-rich protein domain (Brauch et al., 2009). Mutations in RBM20 represent at least 3% of idiopathic DCM cases (Guo et al., 2012) and over 13% of those with a history of sudden cardiac death (SCD) (Marwan et al., 2012).

1.2 Splicing

1.2.1 Introduction

The most of the eukaryotic genes are formed by coding regions, called exons, and not coding regions, called introns; exons and introns are alternated. The genetic information contained in DNA is transcribed in pre-mRNA, which becomes mRNA only after post-transcriptional modifications. One of them is splicing, the process which removes the introns from the pre-mRNA and joins the exons in the mRNA (Figure 2). Only when mRNA is mature it can be translated in the correspondent protein.

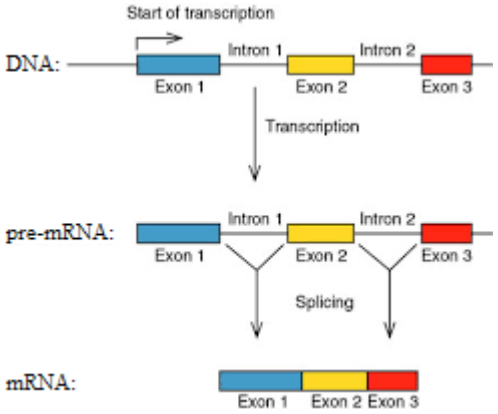

Figure 2. Schematic representation of the splicing process (modified from http://lestrangebiologist.blogspot.it/2012/05/tarea-splicing.html). Introns (black lines) are removed and exons (coloured squares) are joined in the mature transcript (mRNA).

Specific sequences are needed to let splicing happen: the boundary exon-intron at the 5'-end of the intron, called 5'-splice site or donor site, the boundary intron-exon at the 3'-end of the intron, called 3'-splice site or acceptor site, and another sequence, called branch point, which is inside the intron, near the 3'-end of the intron, and it is followed by a polypyrimidine tract. The most preserved sequences are "GU" for the 5'-splice site, "AG" for the 3'-splice site and "A" for the branch point; all of them are inside introns and specific proteins are meant to recognize them. For the very few introns which have different sequences for splicing, other proteins exist in order to recognize also these sequences.

1.2.2 The spliceosome

Splicing process can happen through reactions mediated by a molecular mechanism called spliceosome. It is composed by about 150 proteins and 5 RNA molecules (U1, U2, U4, U5 and U6; U3 is not involved in mRNA splicing). The RNAs are called small nuclear RNAs (snRNA) and form complexes RNA-proteins with proteins of the spliceosome, called small nuclear ribonuclear proteins (snRNPs). The spliceosome composition varies during the different steps of the process: different snRNPs take part in the process in different moments, each of them with its own function.

The role of snRNPs in splicing is three-fold: they recognize 5'-splice site and the branch point, bring them near, and catalyse the cut and the junction of pre-mRNA.

Interactions RNA-RNA, protein-RNA and protein-protein are all necessary, but also some proteins which don't form complexes with RNA are involved in splicing. One of these proteins is the U2 auxiliary factor (U2AF) which recognizes the polypyrimidine tract and the 3'-splice site. The other proteins help the interactions RNA-snRNP.

1.2.3 General splicing mechanism

In the splicing mechanism, U1 binds to the sequence GU at the 5'-end and U2 binds to the sequence A within the branch site with the help of the U2AF protein. The complex at this stage is called spliceosome "A complex". The formation of the **"**A complex**"** is usually the key step in determining the ends of the intron to be spliced out and defining the ends of the exon to be retained. U4, U5 and U6 join to the spliceosome, U6 replaces U1, and U1 and U4 leave. The remaining complex performs two trans-esterification reactions. In the first reaction, the 5'-end of the intron is cleaved from the upstream exon and joined to the branch point A by a 2',5'-phosphodiester bond; in the second reaction, the 3'-end of the intron is cleaved from the downstream exon and the two exons are joined by a phosphodiester bond. The intron is then released in loop form and degraded.

Splicing process is regulated by trans-acting proteins (repressors and activators) and corresponding cis-acting regulatory sites (silencers and enhancers, respectively) on the pre-mRNA. However, it was noted that the effects of a splicing factor are often position-dependent (Lim et al., 2011). The secondary structure of the pre-mRNA transcript is involved in splicing regulation too, such as by bringing splicing elements together or by masking a sequence that would otherwise serve as a binding element for a splicing factor (Warf and Berglund, 2010; Reid et al., 2009).

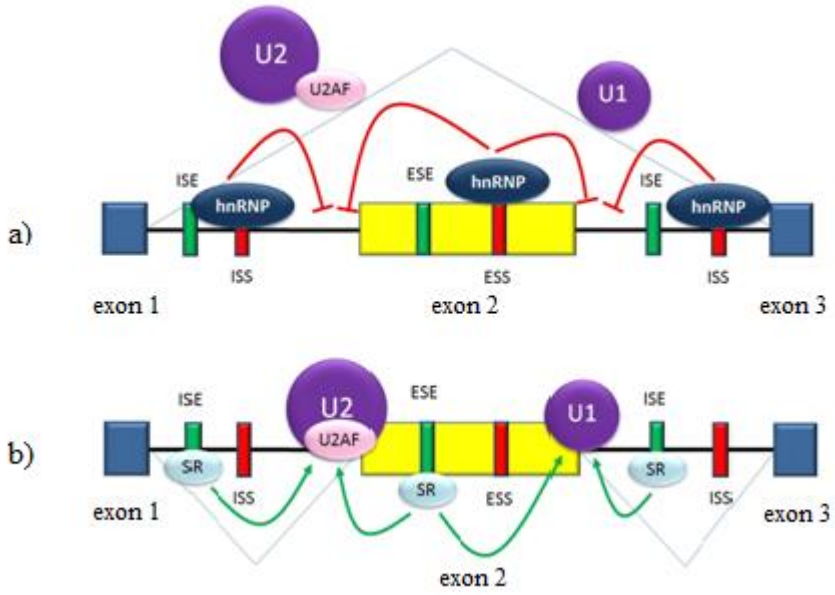

Figure 3. Schematic representation of factors implied in splicing repression and activation (modified from https://en.wikipedia.org/wiki/Alternative_splicing). a) The main splicing repressor proteins, hnRNPs, bind to ISS and ESS preventing the binding of the spliceosome proteins U2, U2AF and U1, in order to reduce the probability of a splicing event; the result is a mRNA containing only exons 1 and 3. b) The main splicing activators proteins, SR proteins, bind to ISE and ESE promoting the binding of the spliceosome proteins U2, U2AF and U1, in order to increase the probability of a splicing event; the result is a mRNA containing exons 1, 2 and 3. Blue and yellow squares: exons, black horizontal lines joining them: introns, green small squares: ISE or ESE, red small squares: ISS or ESS, red curve lines: repressor effect of hnRNPs, green curve lines: activator effect of SR proteins, blue thin lines: exons included in the final transcript.

Splicing silencers (Figure 3a) are cis-acting regulatory sites to which trans-acting repressor proteins bind, in order to reduce the probability that a nearby site will be used as a splice junction. They can be located in the intron itself (intronic splicing

silencers, ISS) or in a neighbouring exon (exonic splicing silencers, ESS). They have variable sequence, as well as the types of proteins which bind to them. The most of splicing repressors are heterogeneous nuclear ribonucleoproteins (hnRNPs), such as polypyrimidine tract binding protein (PTB). Splicing enhancers (Figure 3b) are cis-acting regulatory sites to which trans-acting activator proteins bind, in order to increase the probability that a nearby site will be used as a splice junction. They also may occur in the intron (intronic splicing enhancers, ISE) or the exon (exonic splicing enhancers, ESE). The main splicing activator proteins are members of the SR protein family. Such proteins contain RNA recognition motifs and arginine and serine-rich (RS) domains (Matlin et al., 2005; Wang and Burge, 2008).

In general, the determinants of splicing work in an inter-dependent manner that depends on the presence of other RNA sequence features and on cellular conditions.

Some cis-acting RNA sequence elements influence splicing only if multiple elements are present in the same region, and a cis-acting element can have opposite effects on splicing, depending on which proteins are expressed in the cell. The adaptive significance of splicing silencers and enhancers is attested by studies showing that there is strong selection in human genes against mutations that produce new silencers or disrupt existing enhancers (Fairbrother et al., 2004; Ke et al., 2008).

1.2.4 Alternative splicing

Alternative splicing is a kind of splicing in which not only introns are removed, but sometimes also the exon between the introns, with the aim of obtaining different mRNA from the same initial pre-mRNA. In this way, a single pre-mRNA transcript can code for more proteins, different from each other and often with different biological functions (Figure 4).

The choice of a particular splice site depends on the presence of silencers or

enhancers in the sequence and on the trans-acting proteins which bind them.

This mechanism explain why in the higher organisms there is not linear relation between the number of genes and the number of proteins in the organism; in humans, alternative splicing allows the genome to direct the synthesis of many more proteins than would be expected from its 20,000 protein-coding genes. Alternative splicing occurs as a normal phenomenon in higher eukaryotes (Black, 2003) (in humans, about 95% of multi-exonic genes are alternatively spliced (Pan et al., 2008)) and it is involved in evolution too (Keren, 2010).

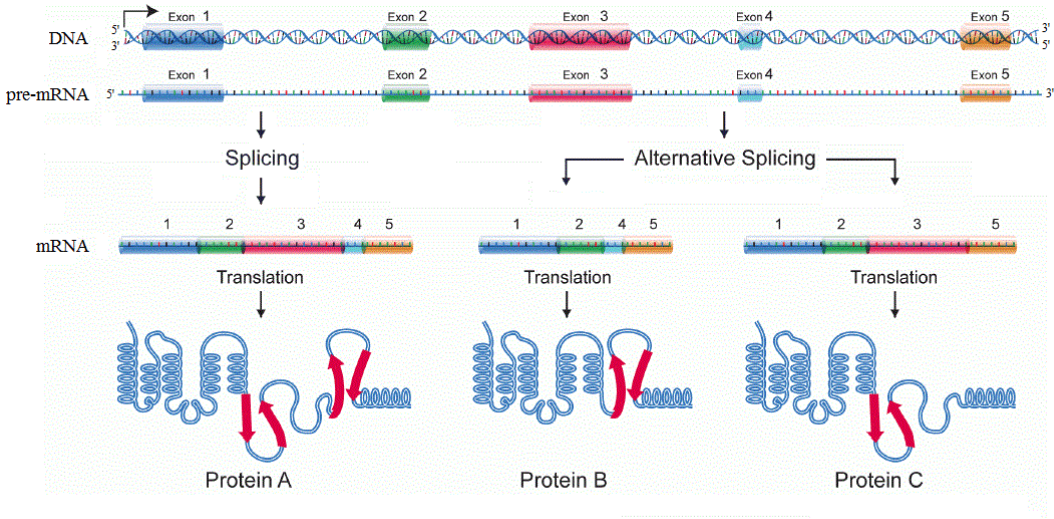

Figure 4. Example of alternative splicing (modified from https://en.wikipedia.org/wiki/Alternative_splicing). Protein A is formed by all exons, protein B lacks exon 3 and protein C lacks exon 4. The different mRNAs code for proteins with different domains and so with different functions too.

There are 5 main kinds of alternative splicing (Black, 2003; Pan et al., 2008; Matlin et al., 2005; Sammeth, 2008) (Figure 5):

1. exon skipping or cassette exon: in this case, an exon may be spliced out of or retained in the primary transcript; this is the most common model in mammalian pre-mRNAs;
2. mutually exclusive exons: one of two exons is retained in mRNAs after splicing, but not both;
3. alternative donor site: an alternative 5'-splice site (donor site) is used,

changing the 3'-boundary of the upstream exon;

4. alternative acceptor site: an alternative 3'-splice site (acceptor site) is used, changing the 5'-boundary of the downstream exon;
5. intron retention: a sequence may be spliced out as an intron or retained; it is distinguished from exon skipping because the retained sequence is not flanked by introns. This is the rarest model in mammals.

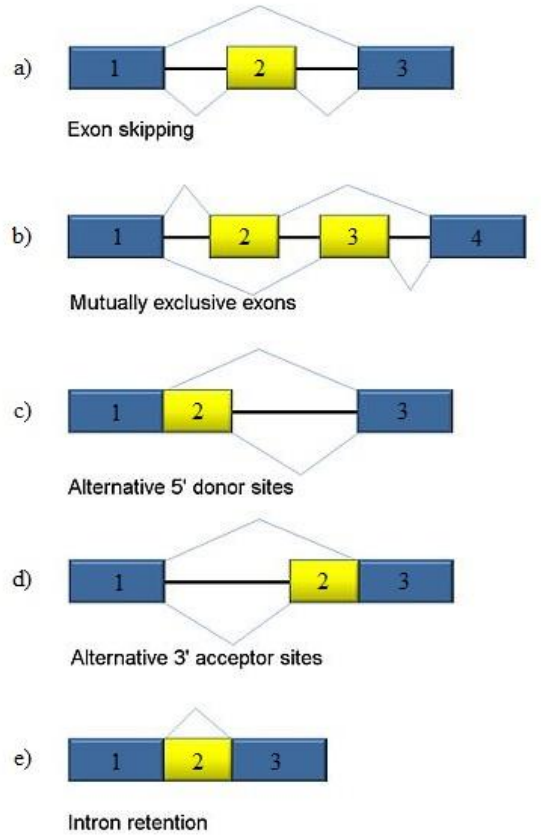

Figure 5. The 5 main kinds of alternative splicing (modified from https://en.wikipedia.org/wiki/Alternative_splicing). Blue and yellow squares: exons, black lines joining them: introns, blue thin lines: exons included in the final transcript. The final transcript will include exons: a) 1-2-3 or 1-3, b) 1-2-4 or 1-3-4, c) 1-3 or 2-3, d) 1-2 or 1-3, e) 1-2-3 or 1-3.

In addition to these main kinds of alternative splicing, there are two other main mechanisms by which different mRNAs may be generated from the same gene:

6. multiple promoters: is properly described as a transcriptional regulation

mechanism rather than alternative splicing; by starting transcription at different points, transcripts with different 5'-starting exons can be generated;

7. multiple polyadenylation sites: provide different 3'-end points for the transcript.

Both of these mechanisms are found in combination with alternative splicing and provide additional variety for the mRNAs derived from the same gene.

1.2.5 Alternative splicing and disease

Changes in the RNA processing machinery may lead to mis-splicing of multiple transcripts, while single-nucleotide alterations in splice sites or in cis-acting splicing regulatory sites may lead to differences in splicing of a single gene and thus in the mRNA produced from a mutant genic transcript.

A number of splicing-related diseases do exist (Ward and Cooper, 2010). A study in 2005 found that more than 60% of mutations which cause human diseases affect splicing rather than directly affecting coding sequences (Lopez-Bigas et al., 2005); a more recent study indicates that one-third of all hereditary diseases are likely to have a splicing component (Lim et al., 2011). Also chromatin structure and histones modifications could have a key role in the regulation of alternative splicing, so epigenetic regulation could determine not only what parts of the genome are expressed but also how they are spliced (Luco et al., 2011).

1.2.6 Alternative splicing at genome level

Genome-wide analysis of alternative splicing is a challenging task. Typically, alternatively spliced transcripts were found by comparing EST sequences, but this required sequencing of very large numbers of ESTs. Thus, high-throughput approaches to investigate splicing have been developed: DNA microarray-based analyses, RNA-binding assays and deep sequencing. These methods can be used to screen for polymorphisms or mutations in or around splicing elements which

affect protein binding.

Results from use of deep sequencing indicate that, in humans, an estimated 95% of transcripts from multi-exon genes undergo alternative splicing, with a number of pre-mRNA transcripts spliced in a tissue-specific manner (Pan et al., 2008). In order to predict functions for alternatively spliced isoforms, functional genomics and computational approaches based on multiple instance learning have also been developed integrating RNA-seq data (Eksi et al., 2013). Deep sequencing has also aided in the *in vivo* detection of the transient loops which are released during splicing, in the determination of branch site sequences, and the large-scale mapping of branch points in human pre-mRNA transcripts (Taggart et al., 2012).

1.3 RNA binding motif protein 20

1.3.1 SR proteins

SR proteins are required for constitutive pre-mRNA splicing and also regulate alternative splice site selection in a concentration-dependent manner (Cáceres et al., 1997), being involved in the assembly of the spliceosome machinery. The choice of splice sites depends on the relationship between the particular SR protein and the cis-acting factors within the exonic or intronic sequences. These protein contain an RS domain, which is a region of variable length rich in repetitive arginine-serine dipeptides, and one or two N-terminal RNA recognition motifs (RRMs), which enable their interaction with a particular pre-mRNA and can be highly phosphorylated on their serine residues (Sahebi et al., 2016).

1.3.2 RBM20

RNA binging motif protein 20 (RBM20) belongs to the family of the SR and SR-related RNA binding proteins. RBM20 is mainly expressed in striated muscle, with the highest levels in the heart (Guo et al., 2012). Recently, mutations in RBM20 have been shown to cause human dilated cardiomyopathy (DCM) (Brauch et al., 2009, Li et a. 2010). RBM20-DCM is a novel example of heart failure owing to a global defect in post-transcriptional regulation (Wyles et al., 2016). Due to the involvement of RBM20 in DCM, this protein was studied a lot in order to unveil its mechanism of action and its RNA targets (Guo et al., 2012; Li et al., 2013).

In vitro and *in vivo* animal studies have significantly contributed to characterize the structural and functional pathogenesis of RBM20 deficiency. Disease pathology recapitulated in RBM20 loss-of-function and deletion models involves altered splicing of numerous genes including several linked to cardiomyopathy, ion transport and contractile function (Guo et al., 2012; Guo et al., 2013; Beraldi et al., 2014). A naturally occurring loss-of-function RBM20 deletion in rats demonstrated that RBM20 plays a major role in alternative splicing in cardiac

adaptive responses mediated by Titin (Ttn) and Calcium/Calmodulin-dependent protein kinase II (Camk2δ) (Guo et al., 2012; Beraldi et al., 2014). Deficient RBM20 resulted in impaired both sarcomere organization and ion transport in the sarcoplasmic reticulum, and premature cell death. Additionally, knock-down studies in murine cells revealed RBM20-deficient cardiogenesis attributable to early disruption of RNA processing and sarcomere remodeling, establishing its pathogenesis as a developmental disorder (Beraldi et al., 2014). A general workflow of RBM20 regulation is displayed in Figure 6.

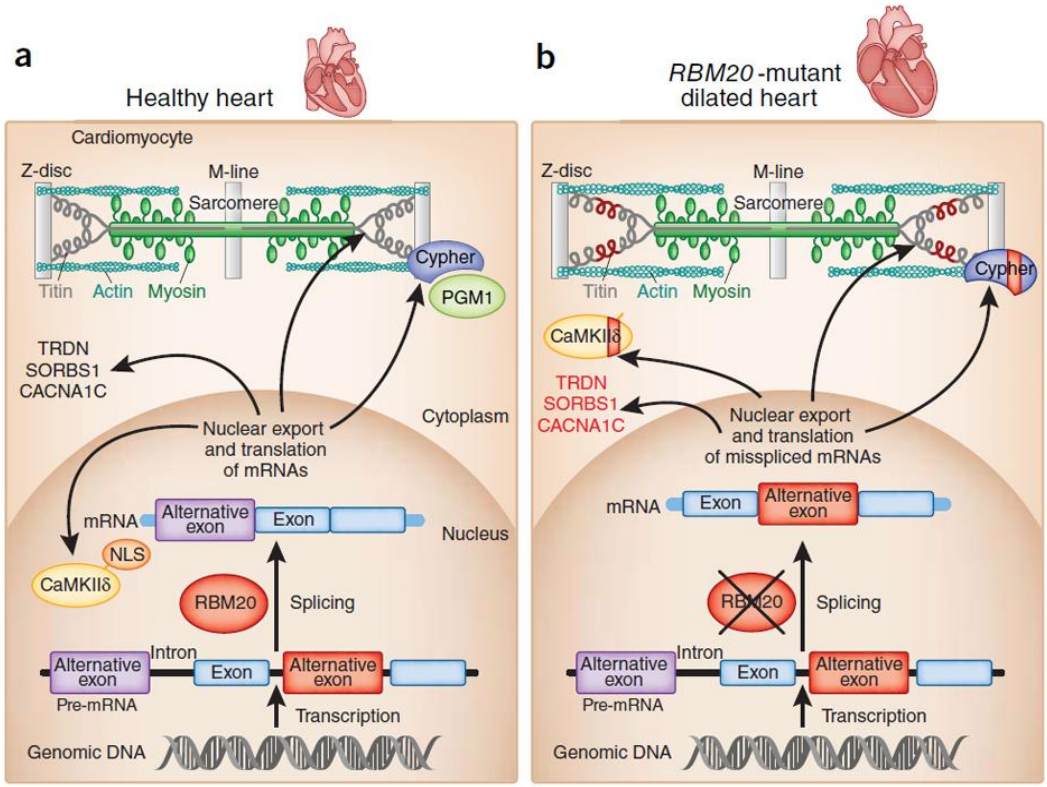

Figure 6. Loss-of-function mutations in the splicing factor RBM20 cause DCM via the pathological splicing of cardiac proteins (Linke and Bücker, 2012). From botton to up of the figure: DNA is transcribed into pre-mRNA, from which introns are removed and exons reconnected in multiple ways to generate mature mRNA; the mature mRNA is then exported from the nucleus and translated into protein. a) In a healthy heart, splice factors such as RBM20 are part of the machinery which controls mRNA splicing patterns. b) Loss-of-function mutations in RBM20 result in aberrant mRNA splicing patterns of cardiac proteins, for example by the inclusion of a different alternative exon (shown in red). Mis-spliced mRNAs are exported from the nucleus, leading to the expression of pathological protein isoforms in cardiomyocytes. Among the proteins with aberrant exons splicing patterns in RBM20-mutant hearts were found Titin (TTN), protein kinase CaMKIIδ, the sarcomere-associated protein Cypher (PGM1 can no longer bind Cypher), triadin (TRDN), sorbin and SH3 domain–containing protein-1 (SORBS1), and the α1C subunit of the L-type calcium channel (CACNA1C). The result is a dilated heart.

For this study, we focused on a set of 18 rat genes whose alternative splicing was regulated by RBM20 (Maatz et al., 2014). Furthermore, RBM20 RNA binding site sequence pattern at nucleotide resolution was revealed and a significant over-representation of the site was found to map in the 400 nucleotides flanking the exons whose alternative splicing is regulated by RBM20 (Maatz et al., 2014). A detailed rule about RBM20 regulation on target exons is, nevertheless, so far unknown.

1.4 RNA sequencing

1.4.1 Introduction

Initial gene expression studies relied on low-throughput methods, such as northern blots and quantitative polymerase chain reaction (qPCR), that are limited to measuring single transcripts. The initial transcriptomics studies were performed using hybridization-based microarray technologies which provide a high-throughput option at relatively low cost (Schena et al., 1995). However, these methods present several limitations: they need to know a priori the sequences being queried, problematic cross-hybridization artefacts in the analysis could happen due to possible highly similar expressed sequences (i. e. paralogue genes), and quantification of both lowly and highly expressed genes could result problematic (reduced dynamic range) (Casneuf et al., 2007; Shendure, 2008).

In contrast to hybridization-based methods, sequence-based approaches have been developed in order to elucidate the transcriptome by directly determining the transcript sequence. Initially, the generation of expressed sequence tag (EST) libraries by Sanger sequencing of complementary DNA (cDNA) was used in gene expression studies, but this approach is relatively low-throughput and not ideal for transcripts quantification (Adams et al., 1991; Itoh et al., 1994; Adams et al., 1995).

To overcome these technical constraints, tag-based methods, such as serial analysis of gene expression (SAGE) and cap analysis of gene expression (CAGE), were developed with the aim to enable higher throughput and more precise quantification of expression levels. The fact that these tag-based methods quantify the number of tagged sequences, which directly corresponded to the number of mRNA transcripts, provides advantages over measuring analogue intensities as in array-based methods (Velculescu et al., 1995; Shiraki et al., 2003). However, these assays are insensitive to measuring expression levels of splice isoforms and can't be used for novel gene discovery. Their use was further limited due to the laborious cloning of sequence tags, the high cost of automated Sanger sequencing and the requirement for large amounts of input RNA (Kukurba and Montgomery, 2015).

The development of high-throughput next-generation sequencing (NGS) revolutionized transcriptomics by enabling RNA analysis through the direct sequencing of complementary DNA (cDNA) (Wang et al., 2009). Global gene expression measurement based on NGS (RNA-Seq) is able to identify and quantify all the individual transcripts (known and unknown) in any cell type at any time (for instance, drug time course experiments). It has distinct advantages over previous approaches. Specifically, RNA-Seq facilitates the ability to look at alternative gene spliced transcripts, post-transcriptional modifications, gene fusion, mutations and changes in gene expression over time, differences in gene expression in different groups or treatments, and allele-specific expression (Maher et al., 2009). In addition to precursor messenger RNA (pre-mRNA) and messenger RNA (mRNA) transcripts, RNA-Seq can look at different populations of RNAs including total RNA, ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), and many others (Ingolia et al., 2012). RNA-Seq can also be used to determine exon-intron boundaries and verify or correct previously annotated 5'- and 3'-gene boundaries. Recently, advances in the RNA-Seq workflow, from sample preparation to sequencing platforms to bioinformatic data analysis, have enabled deep profiling of the transcriptome and the opportunity to elucidate different physiological and pathological conditions (Kukurba and Montgomery, 2015).

1.4.2 Transcriptome sequencing

The introduction of high-throughput next-generation sequencing (NGS) technologies revolutionized transcriptomics. A typical RNA-Seq experiment consists of isolating RNA, converting it to complementary DNA (cDNA), preparing the sequencing library, and sequencing it on an NGS platform (Kukurba and Montgomery, 2015) (Figure 7).

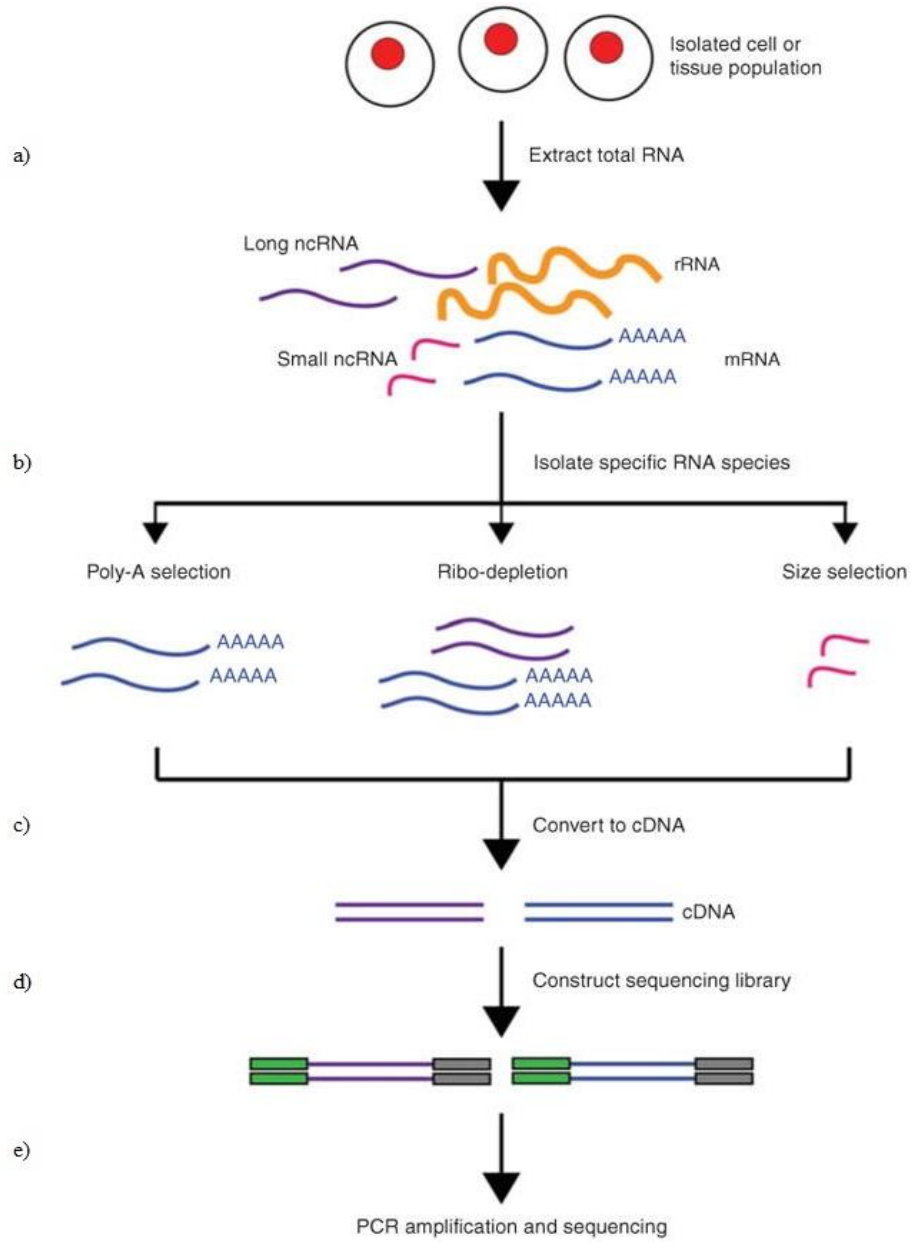

Figure 7. Workflow of transcriptome sequencing for RNA-Seq experiments (modified from Kukurba and Montgomery, 2015). a) RNA is extracted from the biological material of choice (e.g., cells, tissues). b) Second, subsets of RNA molecules are isolated using the specific protocol for each RNA specie, such as the Poly-A selection protocol to enrich for polyadenylated mRNA transcripts, or a Ribo-depletion protocol to remove rRNAs, or a selection based on RNAs size. c) RNA is then converted to complementary DNA (cDNA) by reverse transcription and d) sequencing adaptors are ligated to the ends of the cDNA fragments. e) Following amplification by PCR, the RNA-Seq library is ready for sequencing.

The first step in transcriptome sequencing is the isolation of RNA from a biological sample. To ensure a successful RNA-Seq experiment, the RNA should be of sufficient quality to produce a library for sequencing. Low-quality RNA can

substantially affect the sequencing results and lead to erroneous biological conclusions. Therefore, high-quality RNA is essential for successful RNA-Seq experiments (Tomita et al., 2004; Thompson et al., 2007; Rudloff et al., 2010).

Following RNA isolation, the next step in transcriptome sequencing is the creation of an RNA-Seq library, which can vary by the selection of RNA species and between NGS platforms. The construction of sequencing libraries principally involves isolating the RNA molecules of interest, reverse transcribing the RNA to cDNA, fragmenting or amplifying randomly primed cDNA molecules and ligating sequencing adaptors. Within these basic steps, there are several choices in library construction and experimental design that must be carefully made (Kukurba and Montgomery, 2015).

Reverse transcription results in loss of strandedness, which can be avoided with chemical labelling that let distinguish the second strand from the fisrt strand of cDNA. RNA, cDNA, or both are fragmented. Fragmentation and size selection are performed to select sequences that are of the appropriate length for the experiment. Fragmentation of the RNA reduces 5’ bias of randomly primed reverse transcription and the influence of primer binding sites (Mortazavi et al., 2008), with the downside that the 5’- and 3’-ends are converted to DNA less efficiently. Another consideration for constructing cost-effective RNA-Seq libraries is testing multiple indexed samples in a single sequencing lane. The large number of reads (i.e. short sequences) that can be generated per sequencing run permits the analysis of increasingly complex samples. Indexing cDNA with barcodes enables the pooling and sequencing of multiple samples in the same sequencing reaction, because the barcodes identify which sample the read originated from. For any given study, it is further important to consider the level of sequencing depth required to answer experimental questions with confidence (Kukurba and Montgomery, 2015).

The majority of high-throughput sequencing platforms use a sequencing-by-synthesis method to sequence tens of millions of sequence clusters in parallel. The NGS platforms can often be categorized as either ensemble-based (i.e. sequencing many identical copies of a DNA molecule) or single-molecule-based (i.e.

sequencing a single DNA molecule). The differences between these sequencing techniques and platforms can affect downstream analysis and interpretation of the sequencing data. An important consideration for choosing a sequencing platform is transcriptome assembly. Transcriptome assembly is necessary to transform a collection of short sequencing reads into a set of full-length transcripts. In general, longer sequencing reads make it simpler to accurately and unambiguously assemble transcripts, as well as identify splicing isoforms, which may not be discovered with too short reads. The extremely long reads are ideal for *de novo* transcriptome assembly, in which the reads are not aligned to a genome or transcriptome reference (Kukurba and Montgomery, 2015).

The quality of the raw sequence data should be evaluated to ensure high-quality reads. User-friendly software tools designed to generate quality overviews include the FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Parameters which should be evaluated include the sequence diversity of reads, adaptor contamination, base qualities, nucleotide composition and percentage of called bases. These technical artifacts can arise at the sequencing stage or during the construction of the RNA-Seq. If possible, actions to correct for these biases should be performed, such as trimming the ends of reads, to expedite the speed and improve the quality of the read alignments (Kukurba and Montgomery, 2015).

1.4.3 Experimental parameters

A variety of parameters are considered when designing and conducting RNA-Seq experiments:

- tissue specificity: gene expression varies within and between tissues and RNA-Seq measures this mix of cell types. This can make it difficult to isolate the biological mechanism of interest;
- time dependence: gene expression changes over time and RNA-Seq only takes a snapshot. Time course experiments can be performed to observe changes in the transcriptome;
- coverage or depth: number of reads which include a given nucleotide in the reconstructed transcript. RNA harbours the same mutations observed in

DNA and their detection requires deeper coverage. To detect transcripts of moderate to high abundance, about 30-40 million reads are required to accurately quantify gene expression; to obtain coverage over the full-sequence diversity of complex transcript libraries, including rare and lowly-expressed transcripts, up to 500 million reads is required (Fu et al., 2014). With high enough coverage, RNA-Seq can be used to estimate the expression of each allele too. The depth of sequencing required for specific applications can be extrapolated from a pilot experiment (Li et al., 2008);

- data generation artifacts (or technical variance): the reagents (i. e. library preparation kit), personnel involved and kind of sequencer can result in technical artifacts that might be mis-interpreted as meaningful results. As with any scientific experiment, it is prudent to conduct RNA-Seq in a well controlled setting. If this is not possible, another solution is to detect technical artifacts by inferring latent variables and subsequently correcting for these variables (Stegle et al., 2012);
- data management: a single RNA-Seq experiment in humans is usually on the order of 1 Gb (Kingsford and Patro, 2015); this large volume of data can pose storage issues. One solution is compressing the data using multi-purpose computational schemas or genomics-specific schemas.

1.4.4 RNA-Seq data analysis workflow

Gene expression profiling by RNA-Seq provides an unprecedented high-resolution view of the global transcriptional landscape. As the sequencing technologies and protocol methodologies continually evolve, new informatics challenges and applications develop. Beyond surveying gene expression levels, RNA-Seq can also be applied to discover novel gene structures, alternatively spliced isoforms and allele-specific expression (ASE). In addition, genetic studies of gene expression using RNA-Seq have observed genetically correlated variability in expression, splicing, and ASE (Montgomery et al., 2010; Pickrell et al., 2010; Battle et al., 2013; Lappalainen et al., 2013).

The conventional pipeline for RNA-Seq data includes generating files containing

reads sequenced from an NGS platform, aligning these reads to an annotated reference genome or transcriptome, and quantifying expression of genes (Figure 8). RNA-Seq analysis presents unique computational challenges not encountered in other sequencing-based analyses and requires specific consideration to the biases inherent in expression data (Kukurba and Montgomery, 2015).

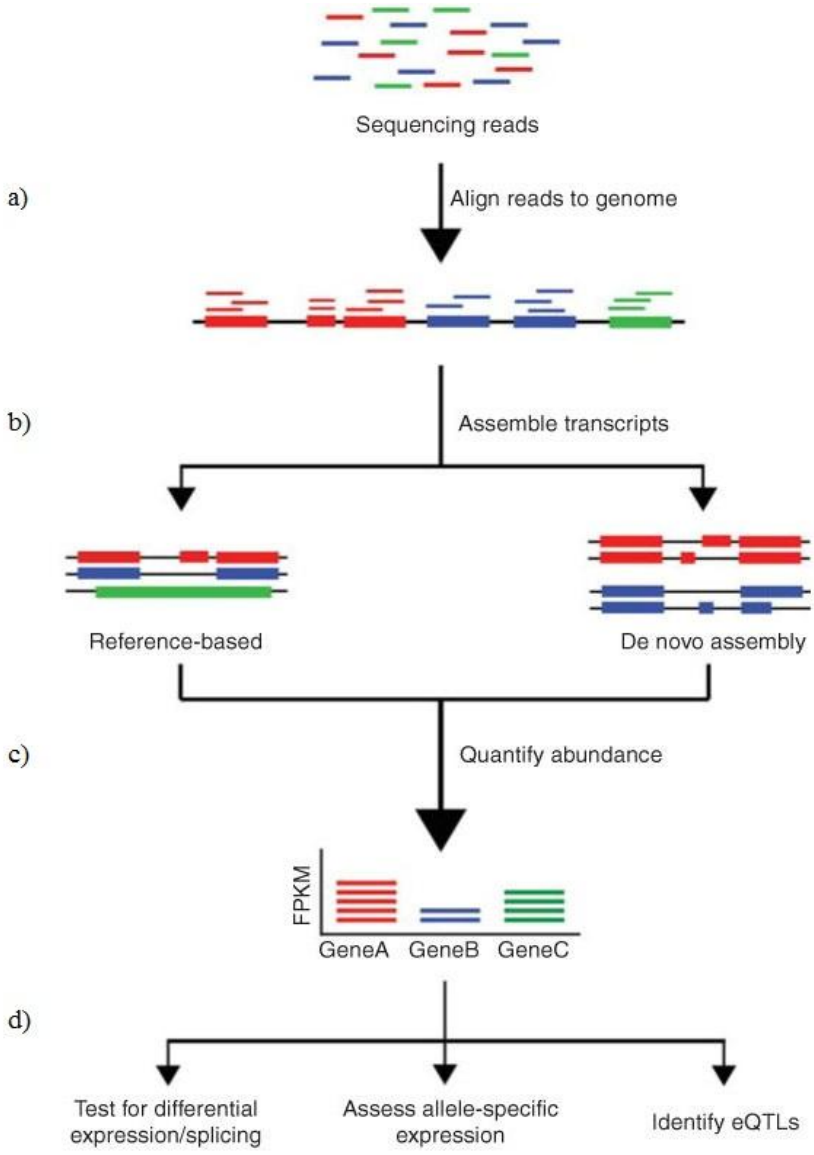

Figure 8. Workflow of RNA-Seq data analysis (modified from Kukurba and Montgomery, 2015). a) Sequenced reads are aligned to a reference genome. b) Reads may be assembled into transcripts using reference transcript annotations or *de novo* assembly approaches. c) The expression level of each gene is estimated by counting the number of reads which align to each exon or full-length transcript. d) Downstream analyses with RNA-Seq data include testing for genes or exons differential expression between samples, detecting allele-specific expression, and identifying expression quantitative trait loci (eQTLs).

When we map RNA-Seq reads (in FASTQ format) to a genome reference (Figure 8a), we have to pay attention because many reads map across splice junctions (Figure 9). In fact, conventional read mapping algorithms are not recommended for mapping RNA-Seq reads to the reference genome because of their inability to handle spliced transcripts (Kukurba and Montgomery, 2015). One approach to resolve this problem is to supplement the reference genome with sequences derived from exon–exon splice junctions acquired from known gene annotations (Mortazavi et al., 2008). A preferred strategy is to map reads with a "splicing-aware" aligner which can recognize the difference between a read aligning across an exon–intron boundary and a read with a short insertion. The more commonly used RNA-Seq alignment tools include GSNAP (Wu and Nacu, 2010), MapSplice (Wang et al., 2010a), RUM (Grant GR et al., 2011), STAR (Dobin et al., 2013) and TopHat (Trapnell et al., 2009). Each aligner has different advantages in terms of performance, speed, and memory utilization. Selecting the best aligner to use depends on these metrics and the overall objectives of the RNA-Seq study (Kukurba and Montgomery, 2015).

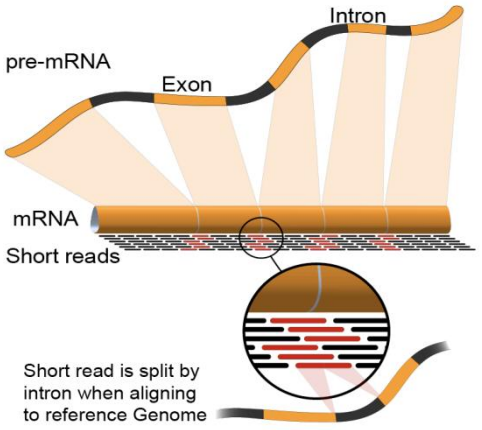

Figure 9. Graphical representation of RNA-Seq reads mapping on exon-exon junctions (https://en.wikipedia.org/wiki/RNA-Seq). After pre-mRNA is converted into mRNA by removing all introns, and mRNA is fragmented in short reads, converted to cDNA, amplified and sequenced, reads are usually aligned to a genome reference in order to assemble transcripts. Reads mapping on the exon-exon junctions are difficult to be found, so specific splicing-aware algorithms are used.

After RNA-Seq reads are aligned, we obtain SAM or BAM format files reads

which can be assembled into transcripts (Figure 8b). Two methods are used to infer transcript models: a reference-based assembly and a *de novo* assembly. In the first approach, a genome or transcriptome (based on the use we need) reference is used, on which reads map in order to reconstruct transcripts. This is particularly helpful when reads don't cover all exons and gaps could be introduced in the reconstructed transcriptome sequence. Typically, alignment algorithms have two steps: align short portions of the read and use dynamic programming to find an optimal alignment, sometimes in combination with known annotations. Software tools that use reference-guided alignment include Bowtie (Langmead et al., 2009), TopHat (Trapnell et al., 2009), STAR (Dobin et al., 2013) and GMAP (Wu and Watanabe, 2005).

De novo assembly doesn't require a genome or transcriptome reference to reconstruct the transcriptome and it is typically used when the genome is unknown, incomplete or altered compared to the reference (Grabherr et al., 2011). Challenges when using short reads for *de novo* assembly include determining which reads should be joined together into contiguous sequences (contigs), robustness to sequencing errors and other artifacts, and computational efficiency. The early algorithm used for *de novo* assembly transitioned from overlap graphs, which identify all pair-wise overlaps between reads, to de Bruijn graphs, which break reads into sequences of length *k* and collapse all *k*-mers into a hash table (Illumina, 2016). Overlap graphs were used with Sanger sequencing, but don't scale well to the millions of reads generated with RNA-Seq. Paired-end sequencing (sequencing of both ends of a fragment) and long read sequencing of the same sample can mitigate the deficits in short read sequencing by serving as a template or skeleton.

A note on assembly quality: assembly quality can vary depending on which metric is used, assemblies which scored well in one specie don't necessarily perform well in all other species, and combining different approaches might be the most reliable thing to do (Lu et al., 2013). The nature of the transcriptome (i. e. gene complexity, degree of polymorphisms, alternative splicing, dynamic range of expression), common technological challenges (i. e. sequencing errors) and steps

of the bioinformatics workflow (i. e. gene annotation, inference of transcripts) can substantially affect transcriptome assembly quality (Kukurba and Montgomery, 2015).

A common downstream characteristic of transcripts reconstruction softwares is the estimation of gene expression levels (Figure 8c). Computational tools such as Cufflinks (Trapnell et al., 2010) and MISO (Katz et al., 2010) quantify expression by counting the number of reads that map to full-length transcripts. Alternative approaches, such as HTSeq (Anders et al., 2014), can quantify expression without assembling transcripts, but counting the number of reads that map to an exon (Anders et al., 2013). To accurately estimate gene expression, reads count must be normalized to correct for systematic variability, such as library fragments size, sequence composition bias and reads depth (Oshlack and Wakefield, 2009; Roberts et al., 2011b). To account for these sources of variability, RPKM (reads per kilobase of transcripts per million mapped reads) metric was introduced, which normalizes a transcript reads count by both the gene length and the total number of mapped reads in the sample (Kukurba and Montgomery, 2015). In the case of paired-ends reads, RPKM is replaced by FPKM (paired fragments per kilobase of transcript per million mapped reads) metric, which accounts for the dependency between paired-end reads in the RPKM estimate (Trapnell et al., 2010). Another technical challenge for transcript quantification is the mapping of reads to multiple transcripts which are the result of genes with multiple isoforms or close paralogs. One solution to correct for this "read assignment uncertainty" is to exclude all reads that do not map uniquely; however, this strategy is far from ideal for genes lacking unique exons. An alternative strategy is to construct a likelihood function which models the sequencing experiment and estimates the maximum likelihood that a read maps to a particular isoform (Kukurba and Montgomery, 2015).

1.4.5 Differential gene expression

Downstream analyses with RNA-Seq data include testing for genes or exons differential expression between samples, detecting allele-specific expression

(ASE), and identifying expression quantitative trait loci (eQTLs) (Figure 8d).

The primary objective of many gene expression experiments is to detect transcripts or exons showing differential expression across various conditions, such as healthy vs. unhealthy, drug-treatment vs. not-treated, and to find which of them are up- or down-regulated. Hence, negative binomial distribution models which take into account over-dispersion or extra-Poisson variation are used to fit the distribution of read counts across biological replicates (Kukurba and Montgomery, 2015). A variety of statistical methods have been designed specifically to detect differential expression for RNA-Seq data. Among them, we mention Cuffdiff (which is part of the Tuxedo suite of tools: Bowtie, Tophat and Cufflinks) (Trapnell et al., 2013), edgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010), DEGseq (Wang et al., 2010b), and the one that I used, DEXSeq (Anders et al., 2012). Although these packages can assign significance to differentially expressed transcripts or exons, the biological observations should be carefully interpreted (Kukurba and Montgomery, 2015).

1.4.6 Allele-specific expression

A major advantage of RNA-Seq is the ability to profile transcriptome dynamics at a single-nucleotide resolution. Therefore, the sequenced transcript reads can provide coverage across heterozygous sites, representing transcription from both the maternal and paternal alleles. If a sufficient number of reads cover a heterozygous site within a gene, the null hypothesis is that the ratio of maternal to paternal alleles is balanced. Significant deviation from this expectation suggests allele-specific expression (ASE). Potential mechanisms for ASE include genetic variation (i. e. single-nucleotide polymorphism in a cis-regulatory region upstream of a gene) and epigenetic effects (i. e. genomic imprinting, methylation, histone modifications) (Kukurba and Montgomery, 2015). Early studies showed that ASE differences can affect up to 30% of loci within an individual (Ge et al., 2009) and that are caused by both common and rare genetic variants (Pastinen, 2010). Studies have also applied ASE to identify expression modifiers of protein-coding variation (Lappalainen et al., 2011; Montgomery et al., 2011), effects of

loss-of-function variation (MacArthur et al., 2012) and differences between pathogenic and healthy tissues (Tuch et al., 2010).

1.4.7 Expression quantitative trait loci

Another prominent direction of RNA-Seq studies has been the integration of expression data with other types of biological information, such as genotyping data. The combination of RNA-Seq with genetic variation data has enabled the identification of genetic loci correlated with gene expression variation, also known as expression quantitative trait loci (eQTLs) (Kukurba and Montgomery, 2015). This expression variation, caused by both common and rare variants, is postulated to contribute to phenotypic variation and susceptibility to complex disease across individuals (Majewski and Pastinen, 2011). The goal of eQTL analysis is to identify associations that will uncover underlying biological processes, discover genetic variants causing disease and determine causal pathways. Most of the eQTLs identified directly influenced gene expression in an allele-specific manner and were located near transcriptional start sites (TSS), indicating that eQTLs could modulate expression directly (cis-eQTL) (Kukurba and Montgomery, 2015). Later studies identified trans-eQTLs, which are variants that affect the expression of a distant gene (>1 Mb) by modifying the activity or expression of upstream factors that regulate the gene (Fehrmann et al., 2011; Battle et al., 2013; Westra et al., 2013).

1.5 Support Vector Machine

1.5.1 Machine learning

Significant advances in biotechnology and more specifically high-throughput sequencing result incessantly in an easy and inexpensive data production, thereby ushering the science of applied biology into the area of big data (Marx, 2013; Mattmann, 2013).

Up to date, besides high performance sequencing methods, there is a plethora of digital machines and sensors from various research fields generating data, including super-resolution digital microscopy, mass spectrometry, Magnetic Resonance Imaging (MRI). Although these technologies produce a big amount of data, they do not provide any kind of analysis, interpretation or extraction of knowledge. Thus, the area of Biological Data Mining or otherwise Knowledge Discovery in Biological Data, is more than ever necessary and important. The primary objective is to investigate the rapidly accruing body of biological data and set the basis potentiating answers to fundamental questions in biology and medicine (Kavakiotis et al., 2017).

The power and effectiveness of these approaches are derived from the ability of methods to extract patterns and create models from data, particularly significant ability in the Big Data era, when the dataset can reach Terabytes or Petabytes of data. Consequently, the abundance of data has strengthened considerably data-oriented research in biology. In such a hybrid field, one of the most important research applications is prognosis and diagnosis related to human-threatening and/or life quality reducing diseases. Applying machine learning and data mining methods in research is a key approach to utilizing large volumes of available data for extracting knowledge (Kavakiotis et al., 2017).

Machine learning is the scientific field dealing with the ways in which machines learn from experience. The purpose of machine learning is the construction of computer systems that can adapt and learn from their experience (Wilson and Keil, 1999). It represents a set of methods and techniques recently developed which explore the study and construction of algorithms that can learn from and

make predictions on data (Kohavi and Provost, 1998). Such algorithms overcome following strictly static program instructions by making data driven predictions or decisions, building a model from input samples.

A core objective of a learner is to generalize from its experience (Bishop, 2006; Mohri et al., 2012). Generalization in this context is the ability of a learning machine to perform accurately on new, unseen examples/tasks after having experienced a learning data set. The training examples come from some generally unknown probability distribution (considered representative of the space of occurrences) and the learner has to build a general model about this space that enables it to produce sufficiently accurate predictions in new cases.

Machine learning tasks are typically classified into three broad categories (Russell and Norvig, 2003): supervised learning (in which the system infers a function from labelled training data), unsupervised learning (in which the system tries to infer the structure of unlabelled data), and reinforcement learning (in which the system interacts with a dynamic environment). However, not all learning tasks can be associated uniquely to these categories.

Some of the most common techniques of supervised learning are Decision Trees (DT), Random Forest (RF), and Support Vector Machines (SVM); a technique of unsupervised learning is k-Nearest Neighbors (k-NN).

1.5.2 Supervised learning

In supervised learning, the system must learn inductively a function called target function, which is an expression of a model describing the data. The function is used to predict the value of a variable, called dependent variable or output variable or label or class, from a set of variables, called independent variables or input variables or characteristics or features. The set of possible input values of the function, i.e. its domain, are called instances. Each element is described by a set of characteristics (features). In order to train the model, the set of all elements, for which the output variable value is known, is divided in two subsets. A subset

is the training set and is used to create the model, based on the feature of its elements. The other subset is the testing set and is used to test the model on elements with known output variable value.

A lot of different approaches can be applied to select the training and testing sets; usually, the 75-80% of the elements are considered as training, the rest as testing. The training set can further be divided in subsets, in which some of them act as training subsets and some as testing subsets (cross-validation). The most used approach is the k-fold cross-validation, in which the training set is divided in *k* subsets, of which *k*-1 are training subsets and 1 is the testing subset, in turns. In our analysis, we used the 10-fold cross-validation with 5 repeats.

In order to infer the best target function, the learning system, given a training set, takes into consideration alternative functions, called hypothesis (Kavakiotis et al., 2017). Once found the best function able to represent our data, it can be applied on a new set of data, for which the output variable value is unknown, in order to predict it.

1.5.3 Unsupervised learning

In unsupervised learning, the system tries to discover the hidden structure of our data or associations between variables. In that case, the dataset consists of instances without any corresponding label. The system doesn't have any information about the correctness of the predicted output variable value or about the target function to be approximated, but it takes information through environment experimentation. The system itself has to identify regularities from the set of all elements. Once identified, it has the capability of generalize on unknown new elements. An example of unsupervised learning is clustering.

1.5.4 Reinforcement learning

The term Reinforcement Learning is a general term given to a family of

techniques, in which the system attempts to learn through direct interaction with the environment so as to maximize some notion of cumulative reward (Alpaydin, 2004). It is important to mention that the system has no prior knowledge about the behaviour of the environment and the only way to find out is through trial and failure (trial and error) (Kavakiotis et al., 2017). It can be associative and not associative, immediate and with delay, direct and indirect. Reinforcement learning is mainly applied to autonomous systems, due to its independence in relation to its environment.

1.5.5 Classification

Classification is a theme of extraordinary importance for the resolution of problems in the real world and it can be applied to very different situations. There are two main kinds of classification: a supervised classification, in which each element of the dataset is assigned to one specific class already known, or a unsupervised classification, in which classes or clusters are found among the dataset of all elements.

From the point of view of the statistical learning theory (Vapnik, 1995), the classification consists in the building of the function f: $R^n \times R \rightarrow R$ which, if applied to a element x in the space R^n to whom is associated a real parameter α, it can predict the class to whom it belongs (usually, label $y \in R$):

$$y_{pred} = f(x, \alpha) \tag{1.1}$$

Given a set of pre-classified samples taken from an unknown probability distribution $P(x, y)$, it is important to choose between all functions $f(x, \alpha)$, the one which minimize the theoretical error:

$$R(\alpha) = \int L(y, f(x, \alpha))\, dP(x, y) \tag{1.2}$$

where L is the loss function and is defined as:

$$L(y, f(x, \alpha)) = |y_{real} - f(x, \alpha)| \tag{1.3}$$

indicating the discrepancy between the real and the predicted class of the element *x*.

$R(\alpha)$, also called expected risk, is a measure of how much the hypothesis performed on the prediction of *y* for an element *x* is far from its real value.

In the case of binary classification, when the set of all elements can be divided only in 2 classes, both y_{real} and α can assume only values {0, 1}; thus, the loss function can be defined as follows:

$$L(y, f(x, \alpha)) = \begin{cases} 0 \text{ if } y = f(x, \alpha), \\ 1 \text{ if } y \neq f(x, \alpha). \end{cases} \quad (1.4)$$

For this loss function, $R(\alpha)$ determines the probability of a wrong classification: the case of differences between the real and the predicted class is called classification error. The aim is, thus, to find the function *f* which minimizes $R(\alpha)$, i. e. the probability of error, in a given dataset.

Furthermore, it is necessary also to minimize the VC dimension (from the name of its creators Vapnik-Chervonenkis), defined as the number of vectors which the system can separate in two classes; it is a sort of complexity of the classificator.

Besides the classification problem, other situations, in which the output variable value corresponds to continuous variable values, exist in the machine learning field. These problems are called regression problems and the function to be approximated is just the regression function.

1.5.6 Linear Support Vector Machine

An algorithm which can solve the classification problem is the Support Vector Machine (SVM) algorithm. SVM is a new and powerful machine learning approach developed by Vapnik (Vapnik, 1995), which can minimize both the number of errors of the prediction and the VC dimension.

Now I will focus on binary classification, in which the set of all elements is divided in two classes and each element can belong only to a class.

Suppose the elements of our training set are linearly separable, i. e. it exists an hyperplane which can separate the elements which belong to a class from the elements which belong to the other class (Figure 10).

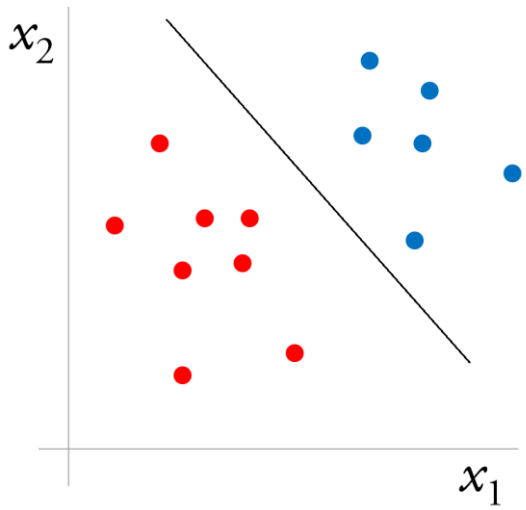

Figure 10. Example of linearly separable classes (modified from http://efavdb.com/svm-classification). The elements of the red class can be linearly separated from the elements of the blue class through a separator hyperplane (black line).

If we call w the vector normal to the hyperplane and b the intercept in the origin, all points laying on the separator hyperplane have to satisfy the equation:

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 0 \tag{1.5}$$

Considering x_i the i-th element of the training set and y_i the class of the i-th element, in binary classification in which the two possible classes are {-1,+1}, all elements of the training set satisfy the following conditions:

$$\boldsymbol{w} \cdot \boldsymbol{x}_i + b \geq 1 \quad \text{if} \quad y_i = 1 \tag{1.6}$$

$$\boldsymbol{w} \cdot \boldsymbol{x}_i + b \leq -1 \quad \text{if} \quad y_i = -1 \tag{1.7}$$

which we can join in a single disequation:

$$y_i\,(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) - 1 \geq 0\ \forall\,\mathrm{i} \qquad (1.8)$$

Infinite hyperplanes exist which can satisfy these disequations, but in order to obtain a good classification it is necessary to determine the parameters *w* and *b* of the hyperplane which best separates the two subsets of elements, minimizing the classification error.

The elements for which $\boldsymbol{w}\cdot\boldsymbol{x}_i$ + b = 1 all lay on the hyperplane H1, the elements for which $\boldsymbol{w}\cdot\boldsymbol{x}_i$ + b = -1 all lay on the hyperplane H2. H1 and H2 are parallel between them and also parallel with the separator hyperplane; any element can be found in the space between them. The distance between them is called margin and it measures 2/||***w***|| (Figure 11).

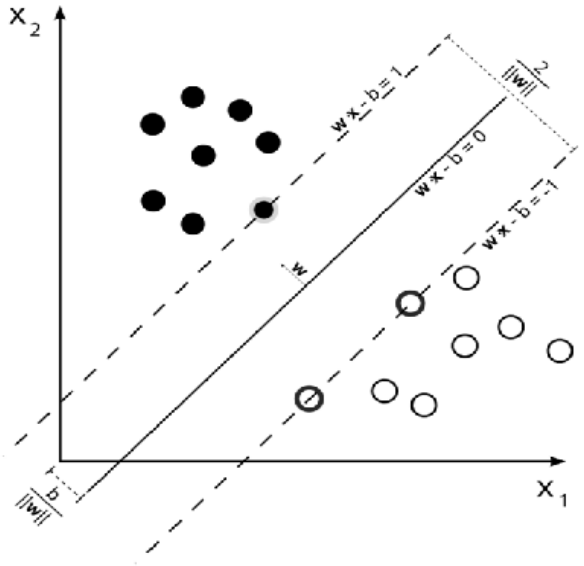

Figure 11. Linear classification with SVM (https://it.wikipedia.org/wiki/Macchine_a_vettori_di_supporto). A hyperplane (solid line) was found that separates the elements in two classes (black circles and empty circles). No elements were observed between the hyperplanes H1 and H2 (dashed lines).

All elements laying on H1 or H2 are called support vectors and are the critical elements for SVM. All other elements, instead, don't affect SVM in any way; if they would be removed or moved without going over the hyperplanes H1 or H2, and the algorithm would be repeated, we will obtain exactly the same separator

hyperplane.

The algorithm which SVM uses in order to find the two parameters of the optimal separator hyperplane consists in maximizing the margin between the two classes and minimizing the difference intra-classes.

1.5.7 Not Linear Support Vector Machine

In the case in which it doesn't exist an hyperplane which linearly separate the elements of our dataset, i. e. some elements lay on the wrong semiplane, we can say that our dataset is not linearly separable.

Thus, the elements of the training set satisfy the following conditions:

$$\boldsymbol{w} \cdot \boldsymbol{x}_i + b \geq 1 - \xi_i \quad \text{if} \quad y_i = 1 \tag{1.9}$$

$$\boldsymbol{w} \cdot \boldsymbol{x}_i + b \leq -1 + \xi_i \quad \text{if} \quad y_i = -1 \tag{1.10}$$

which we can join in a single disequation:

$$y_i \, (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \forall \, \mathrm{i} \tag{1.11}$$

where $\xi_i \geq 0$ is the slack variable and is as higher as farer the *i*-th element is from its correct class.

The separator hyperplane, determined through support vectors, is far $-b/\|\boldsymbol{w}\|$ from the origin and every misclassified element is far $-\xi/\|\boldsymbol{w}\|$ from its correct class.

The function which separates the two classes of elements is, thus, not linear (it could be quadratic or cubic or another function). Nevertheless, if we want to separate through an hyperplane even not linearly separable datasets, we have to map our elements in a higher dimensional space, called features space. If we consider $m > n$, the function which we use to map the elements in the new space is:

$$\Phi : \mathrm{R}^{\mathrm{n}} \rightarrow \mathrm{R}^{\mathrm{m}} \tag{1.12}$$

The two classes which are not linearly separable in the input space became

linearly separable in the feature space through the mapping function Φ (Figure 12).

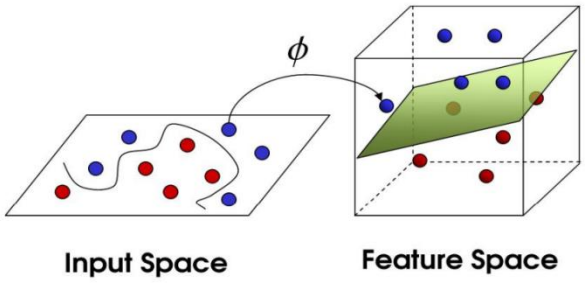

Figure 12. Mapping in a higher dimensional features space (https://www.linkedin.com/pulse/support-vector-machine-srinivas-kulkarni). Features not linearly separable in the input space (on the left of the figure) becomes linearly separable in the features space (on the right of the figure) through the mapping function Φ.

Thus, now, we should replace each element x of our dataset with the function $\Phi(x)$. Since the function Φ can be very expensive to be calculated, we use the kernel trick, i. e. an inexpensive kernel function K which directly calculates the inner product of two elements, without explicitly calculating Φ:

$$K(\boldsymbol{x}_i , \boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j) \tag{1.13}$$

In that way, we can get SVM to learn in the high dimensional features space given by Φ, without explicitly calculating Φ, but calculating only K.

The most used kernels are:

- linear kernel: $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y}$
- polynomial kernel: $K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} \cdot \boldsymbol{y})^d$ or $K(\boldsymbol{x}, \boldsymbol{y}) = (1 + \boldsymbol{x} \cdot \boldsymbol{y})^d$
- Gaussian Radial Basis Function (RBF) kernel: $K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-|\boldsymbol{x}\text{-}\boldsymbol{y}|^2)/(2\sigma^2)$
- sigmoid kernel: $K(\boldsymbol{x}, \boldsymbol{y}) = \tanh(\kappa \boldsymbol{x} \cdot \boldsymbol{y} - \delta)$

For this work, I used a SVM with the RBF kernel, which is computationally cost-effective.

1.5.8 Features selection

Features selection is one of the most important processes. Especially when we have a lot of features in our features space, many of them could be irrelevant or redundant features, so the features selection analysis is applied to select a subset of features from the features space which is more relevant to and informative for the construction of a model. The aims of a features selection analysis are three: 1) improving the overall prediction performance, 2) providing faster and more cost-effective analysis, and 3) providing a better understanding of the underlying process that generated the observations (Guyon and Elisseeff, 2003).

Various features selection algorithms have been published and they may be grouped into three main classes, based on how they determine the selected features (Dash and Liu, 1997; Guyon and Elisseeff, 2003; Liu and Yu, 2005). The first class contains the wrapper algorithms, which use a machine learning algorithm to evaluate different subsets of features and finally select the one with the best performance on classification accuracy; it uses heuristic rules to find locally optimal solutions.

The second class has the filter algorithms, which measure the association of each feature or features subset with the labels of the dataset, and order all the features or features subsets based on this measure. The most of the filter algorithms evaluate individual features.

The third class contains the hybrid algorithms, which aim to automatically generate an optimally selected features subset by integrating the wrapper and the filter strategies within different heuristic features selection steps.

In the present work, we first assessed features individually through a single feature univariate association analysis, to understand their influence on the system, and then in pairs through a correlation test, to find high correlated features. Perfectly correlated features are truly redundant in the sense that no additional information is gained by adding them (Guyon and Elisseeff, 2003).

1.5.9 Testing

After the features selection analysis and the training phase in which the SVM gets the parameters of the optimal separator hyperplane, based on the selected features, it can proceed classifying a new set of elements. This phase is called testing and it consists in assigning the correct class to a new element *x*.

As concerning binary classification, the basis decision function is:

$$f(x) = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x} + b) \tag{1.14}$$

which assigns the label 1 or -1 to each element, based on its predicted class. If we use a kernel function in our model, this function have to be slightly modified, but the principle is the same.

Usually, before testing new unknown elements, the SVM is applied on a subset of elements of our dataset with known class (usually the 20-25% of all elements) in order to evaluate the performance of the prediction of the SVM on known data. There are a lot of measures of performance that we can use for this propose: the accuracy, sensibility, specificity, precision, recall, F1 score, 95% of the confident interval (CI), and the area under the curve (AUC) of the receiver operating characteristics (ROC) curve.

Both the AUC of the model ($0 \leq \text{AUC} \leq 1$) and the accuracy of the prediction ($0 \leq \text{accuracy} \leq 1$) were calculated to evaluate the performance of this work.

Once tested SVM on known data with good results, it can be used to predict new unknown data.

2 Materials and methods

All analyses performed in this study are summarized in Figure 13.

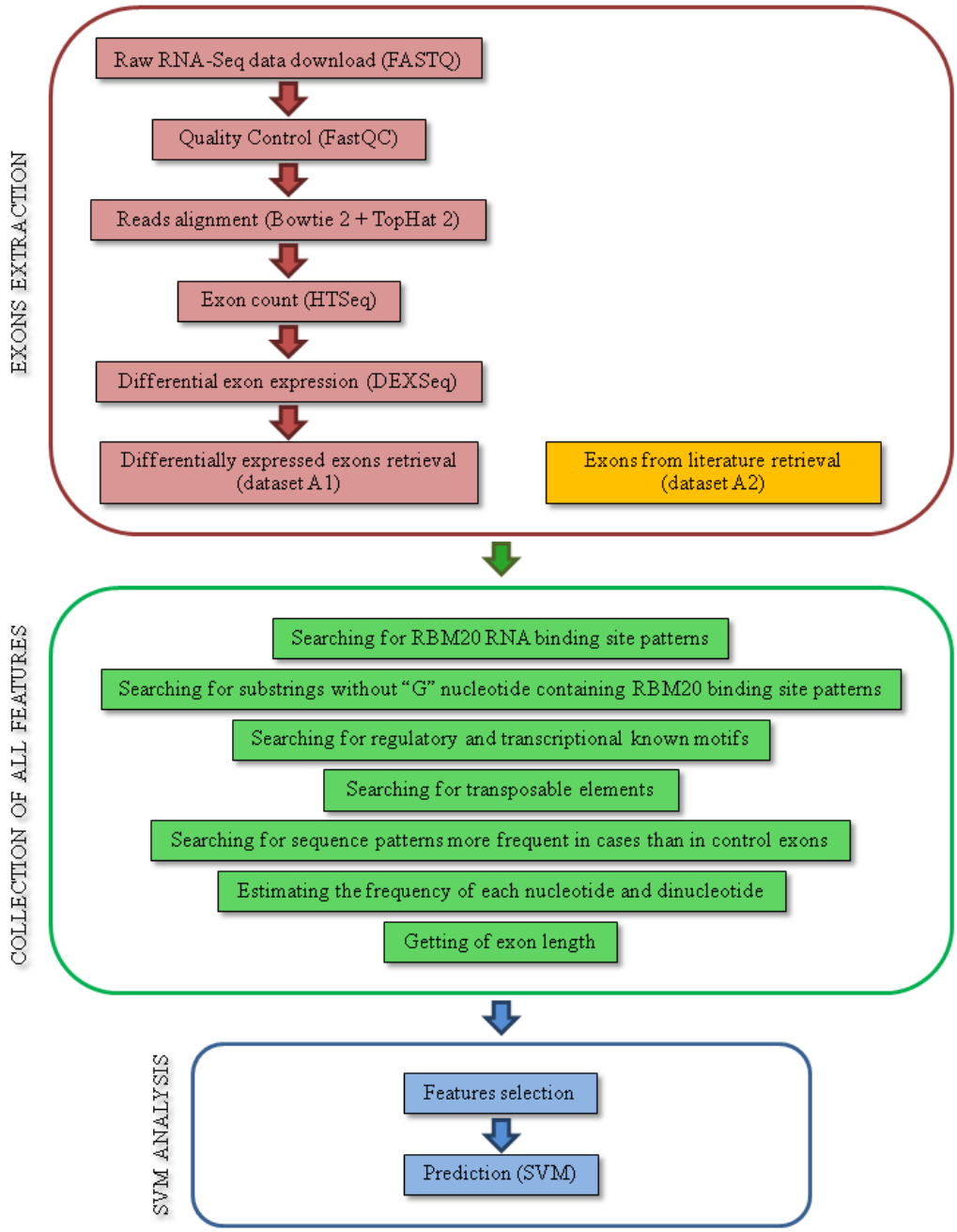

Figure 13. Workflow of the analyses performed in this study. The whole procedure can be divided in 3 main sections: exons extraction, collection of all features, and SVM analysis.

2.1 Public RNA-Seq data of RBM20 mutants

Gene expression profiles of $RBM20^{+}$ and $RBM20^{-}$ cells were estimated by analysing published RNA-Seq data (Guo et al., 2012) from human and rat cardiomyocites. Three human samples (2 $RBM20^{+/+}$ samples with dilated cardiomyopathy without mutations in RBM20 and 1 $RBM20^{-/-}$ sample with dilated cardiomyopathy with mutations in RBM20) and nine rat samples (3 $RBM20^{+/+}$ wild-type, 3 $RBM20^{+/-}$ heterozygous and 3 $RBM20^{-/-}$ knock-out samples) were downloaded from the Sequencing Archive of the European Nucleotide Archive (accession code: ERP001301) (Table 1).

Table 1 – Public RNA-Seq sample data used for the transcriptome analyses. Three human samples and nine rat samples were downloaded from ENA Sequencing Archive. Each FASTQ file had 100bp paired-ends sequences (100 PE). The mean number of sequences for sample was calculated summing the number of sequences of each sample of that group of samples, and then dividing for the number of samples of that group. "SD": standard deviation. "*": no SD value because only one sample was analysed.

Specie	Genotype	Number of samples	Mean number and SD of 100 PE sequences for sample (Millions)
Homo sapiens	$RBM20^{+/+}$	2	82.5 ±4.2
	$RBM20^{-/-}$	1	83.1 *
Rattus norvegicus	$RBM20^{+/+}$	3	73.4 ±0.8
	$RBM20^{+/-}$	3	78.5 ±4.9
	$RBM20^{-/-}$	3	76.0 ±1.7

2.2 Quality control

Before alignment, FASTQ files underwent a quality control analysis performed with FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). To remove adaptors sequences from the reads (i.e. sequences) and to trim poor quality bases, Scythe and Sickle tools were used. Scythe removes 3'-end adapters, with a Naive Bayesian approach; Sickle removes 3'-ends and 5'-ends of reads with low quality, using sliding windows along with quality and length thresholds (Joshi and Fass, 2011).

2.3 Reads alignment

Reference genome sequences (RGSC version 3.4 for *Rattus norvegicus*, and GRCh version 37 for *Homo sapiens*) were retrieved from Ensembl database (www.ensembl.org). Reads (100bp paired-ends - 100 PE) were mapped against the proper genome reference using Bowtie version 2 (Langmead and Salzberg, 2012) and TopHat version 2.0.12 (Kim et al., 2013) with the additional information of a GTF file.

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences; it is particularly good at aligning reads of about 50 up to 100bp or 1000bp, and at aligning to relatively long genomes (i. e. mammalian). It supports gapped, local, and paired-end alignment modes.

TopHat 2 is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to transcriptome, to genome, and to novel/known splice sites using the aligner Bowtie 2, and then it analyses the mapping results to identify splice junctions between exons.

2.4 Exons counting and differential expression

The expression of each exon was quantified counting the number of reads that map to it, and normalizing the reads count to correct for the systematic variability.

Reads count was performed through HTSeq package (Anders et al., 2014). Package DEXSeq (Anders et al., 2012) was used to normalize data, estimate the number of reads per exon and the dispersion of such measurement, and then to test for differential exon expression among groups. DEXSeq uses generalized linear models (GLMs) to model read counts, and offers reliable control of false discoveries by taking biological variation into account. It detects with high sensitivity genes and exons which are subjected to differential exon usage.

We tested the 2 human RBM20$^{+/+}$ samples against the human RBM20$^{-/-}$ sample, and the 3 groups of rat samples (RBM20$^{+/+}$, RBM20$^{+/-}$, RBM20$^{-/-}$) in pairs to

detect differentially spliced exons. An exon was defined to be differentially included/skipped when associated with a DEXSeq analysis showing an adjusted p-value < 0.05.

2.5 Differentially expressed exons and extraction of target sequences regions

According to DEXSeq results, a number of rat exons were retrieved for the following analyses, grouped as follows (Table 2): 1 dataset of differentially spliced exons (*A1* of 232 exons), 1 dataset of exons found to be regulated by RBM20 (Maatz et al., 2014) (*A2* of 80 exons), and 1 dataset of exons not regulated by RBM20 (*N1* of 80 exons).

Table 2 – Datasets of rat exons retrieved. Two datasets containing exons whose splicing was affected by RBM20 (A1 and A2), and a dataset containing exons whose splicing was not affected by RBM20 (N1) were retrieved. The number of exons contained in each dataset is indicated.

Kind of retrieved exons	Dataset	Number of exons
differentially spliced exons	A1	232
exons found to be regulated by RBM20	A2	80
exons not regulated by RBM20	N1	80

The datasets retrieved are arranged as follows:

- the dataset A1 contains all the differentially expressed rat exons;
- the dataset A2 includes the exons detected by Maatz (Maatz et al., 2014) analyzing the same raw data of Guo (Guo et al., 2012), and searching for RBM20 binding sites patterns in the differentially expressed exons; the dataset A2 includes 80 exons belonging to 18 genes;
- the dataset N1 contains random exons retrieved from Ensembl, in order to obtain half exons in the forward strand and half in the reverse strand, one exon for gene.

Exons sequences of each dataset were extracted from the genome reference sequence (RGSC3.4 for *Rattus norvegicus*) using the Samtools program (http://samtools.sourceforge.net/), according to exons coordinates reported by different sources:

- by DEXSeq differential analysis (for the dataset A1);
- by literature (for the dataset A2);
- by Ensembl (for the dataset N1).

Only the top strand exon sequence (i. e. the transcribed strand sequence) was extracted.

Every exon was studied by focusing on either the first or last 30 exonic nucleotides and the preceding or following 400 intronic nucleotides, respectively, as suggested from literature (Maatz et al., 2014), for a total of 430+430 nucleotides analysed for each exon. The intronic nucleotide regions were reduced to the first and last 200 nucleotides when necessary (230+230 nucleotides analysed for each exon) (Figure 14).

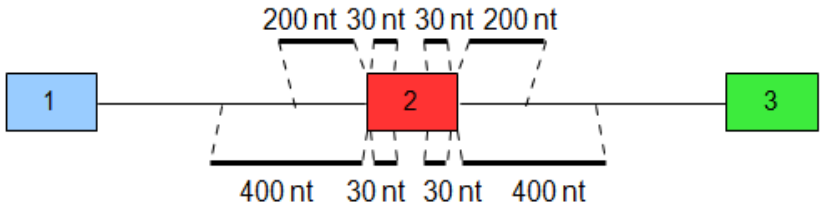

Figure 14. Exonic and intronic regions under investigation. For each exon the 430 or 230 flanking nucleotides were extracted, 30 of which were in the exon and the remaining in the flanking introns. Coloured squared: exons, black horizontal lines between exons: introns, dashed lines: regions investigated.

2.6 Searching for binding sites patterns

An in house R algorithm was employed to search the genomic sequences of the exons of the datasets A1, A2 and N1 for specific RNA binding sites patterns.

By the R program, every single binding site pattern was searched along the target sequences using a nucleotide probability matrix (npm), in which the probability of each position to be “A”, “C”, “G” and “T” is fixed. For each sequence position, the overall probability to be the starting point of that binding site was estimated by a score based on the sum of the negative logarithm of the occurrence frequency of each nucleotide in the npm. Positions with the higher scores were likely to match the binding site. A threshold allowing up to 1 mismatch with the binding site pattern was chosen in order to tag the most likely binding sites. For each putative binding site, the number of single sites and cluster of sites (sites close to each other less than 10 nucleotides) in the 430 nucleotides flanking the selected exons and in sub-regions made of 100 nucleotides each (exons upstream and downstream regions studied separately) were calculated. A descriptive statistic about the counts, the size of the clusters of sites and the distribution of single sites and clusters in the flanking regions was then performed.

2.7 Searching for additional genetic elements

A further R program was written to investigate additional genetic elements in the flanking regions of the 3 datasets (exons upstream and downstream regions were studied separately). First of all, we searched for known genetic elements (substrings without “G” nucleotide containing RBM20 binding site patterns, regulatory and transcriptional known motifs, and transposable elements), then for new genetic elements (enrichment analysis), and finally for sequence characteristics (frequency of nucleotides and dinucleotides, exons length):

- substrings without “G” nucleotide containing RBM20 binding site patterns. As RBM20 RNA binding site logo (Maatz et al., 2014) doesn't contain any “G” nucleotide, each target sequence was first searched for substrings without “Gs”, and then the presence of RBM20 binding site pattern/patterns was investigated on these substrings. This analysis was more accurate with respect to the simple search for RBM20 binding site. Only substrings longer than 15 nucleotides were selected for the following

analyses. Each exon was statistically described through 7 features, related to the number of substrings and the length of the longer substring of that exon, the highest score among all the scores of the putative binding sites of the longer substring, the number of putative binding sites with scores higher than a chosen threshold in the longer substring, and the maximum, mean and median scores among all scores of the putative binding sites found in all substrings of that exon. The analysis was performed on the 430 nucleotides regions flanking the target exons, subsequently reducing the region of interest on the 230 flanking nucleotides;

- regulatory and transcriptional known motifs. For each sequence, the algorithm sent automatically a query to RegRNA2.0 website (Chang et al., 2013) to indicate the motifs to search for; all the 16 kinds of motifs available in the website were selected. Then, it extracted the result of the search, counting the number of all motifs found in that sequence. The analysis was performed on the 430 nucleotides regions flanking the target exons, subsequently reducing the region of interest on the 230 flanking nucleotides;

- transposable elements. Each dataset of exons was loaded in RepeatMasker web server (Smit et al.,) and investigated for all the 20 types of transposable elements available; the result of the analysis was extracted and the number and type of transposable elements found in each sequence were counted. The analysis was performed on the 430 nucleotides regions flanking the target exons, subsequently reducing the region of interest on the 230 flanking nucleotides. A subset of transposable elements (SINEs, Simple repeats and LTRs) and the total number of interspersed repeats were studied deeply, and Fisher exact test was performed to evaluate a possible enrichment of each of these selected elements in cases or in control exons (p-value < 0.05);

- additional sequence patterns more frequent in cases than in control exons (enrichment analysis). Software DREME of MEME package was used (E-value < 0.05), searching for patterns in the top strand and in both strands

of the 430 nucleotides regions flanking the selected exons. Patterns with p-value < 10^{-7} and E-value < 10^{-3} were selected;

- frequency of each nucleotide and dinucleotide in the target sequences. The 430 nucleotides regions flanking the selected exons and sub-regions made of 100 nucleotides each were analysed;

- length of the exons analyzed. A descriptive statistic was applied, based on the length of the exons of each dataset.

2.8 Features selection

All information collected before (sections 2.6 and 2.7) were stored in a single R object, in which every count/size/characteristic/statistic/frequency was represented as numeric value (feature). Every exon flanking region was searched for the same genetic elements, so all exons were described by the same number of features.

As when there are a lot of features many of them could be irrelevant or redundant, a single feature univariate association analysis was performed in order to select the subset of features able to distinguish between case and control exons. Every feature was analysed with one out of three possible tests based on the feature numerical type, comparing a group of case exons to a group of control exons. Features indicating the presence of a particular element (i. e. features about the presence of regulatory and transcriptional known motifs, or of transposable elements) were analysed with a Fisher exact test, features regarding counts (i. e. features about the search for binding site patterns and for substrings without "G" nucleotide) were analysed with General Linear Models with family Poisson, the remaining features were analysed with a Kruskal-Wallis test.

Features with an associated p-value < 0.2 were selected and underwent a pair-wise correlation test to discard highly correlated features (p-value < 0.0000005 and r coefficient > |0.9|). For each pair of highly correlated features, the feature with the lowest p-value in the single feature association analysis was retained and the other feature was wiped out from the following analyses.

2.9 Support Vector Machine

Support Vector Machine (SVM) (Vapnik, 1995) was used to classify RBM20 affected exons. The 75% of the exons (affected and not) was randomly chosen to train the SVM, based on the selected features. The model was built with a radial kernel function, and a 10-fold cross-validation with 5 repeats was carried out, in order to evaluate the best model to predict our data. The model with higher area under the curve (AUC) value of the receiver operating characteristics (ROC) curve was selected ($0 \leq AUC \leq 1$) and used to predict the class (RBM20 affected or RBM20 not affected) of the remaining 25% of the exons. To evaluate the performance of the prediction, the accuracy value was estimated ($0 \leq accuracy \leq 1$). The accuracy is defined as the proportion of true results (both true positives and true negatives) among the total number of cases examined.

An additional analysis was performed by permuting the features value before SVM analysis. Through the permutation process, the values of RBM20 affected and not affected exons were scrambled within the same feature, to disrupt any possible correlation between exon class and feature value.

Analyses were performed on both balanced (i. e. an equal number of case and control exons) and unbalanced (i. e. a different number of case and control exons) training and testing sets, to evaluate SVM performances in these contexts.

3 Results

3.1 The dataset A1 of RBM20 affected exons

In order to study the different genomic characteristics of exons whose splicing is regulated by RBM20, a group of RBM20 affected exons was identified by performing a differential analysis on public RNA-Seq data (raw sequence data in FASTQ format) of RMB20 mutants.

Existing human and rat sequencing data of RNA (Guo et al., 2012) were downloaded and the reads underwent a Quality Control analysis performed with FastQC tool before aligning them to the genome reference. Poor quality bases were trimmed using Scythe and Sickle tools and the resulting human and rat reads were aligned to the respective genome reference (GRCh37 for human reads, RGSC3.4 for rat reads) using Bowtie2 and TopHat2. Every exon was quantified based on the number of reads mapping on it, and differential analyses were performed between the 2 groups of human samples (2 $RBM20^{+/+}$ and 1 $RBM20^{-/-}$ samples) and the 3 groups of rat samples (3 $RBM20^{+/+}$, 3 $RBM20^{+/-}$ and 3 $RBM20^{-/-}$ samples) in pairs, through DEXSeq.

The number of significant exons and genes resulted from the differential analyses is shown in Table 3.

Table 3 – DEXSeq differential analyses significant results for rat and human samples. Expression profiles were determined from RNA-Seq trimmed data. Exons were considered to be significantly differentially expressed when showing an adjusted p-value < 0.05; genes were considered to be significant having at least a significant exon.

	Rat			**Human**
	$RBM20^{+/+}$ vs. $RBM20^{-/-}$	**$RBM20^{+/-}$ vs. $RBM20^{-/-}$**	**$RBM20^{+/+}$ vs. $RBM20^{+/-}$**	**$RBM20^{+/+}$ vs. $RBM20^{-/-}$**
exons	232	88	41	1452
genes	128	41	14	590

128 significant differentially expressed rat genes and 232 significant exons were observed when comparing $RBM20^{+/+}$ vs. $RBM20^{-/-}$ samples, while in the other two comparisons of rat samples the number of significant genes and exons decreased (41 and 14 significant genes, 88 and 41 significant exons, for $RBM20^{+/-}$ vs. $RBM20^{-/-}$ and $RBM20^{+/+}$ vs. $RBM20^{+/-}$ comparisons, respectively). When comparing the human samples, an increased number of significant genes and exons (590 and 1452, respectively) was found; this is probably due to the higher number of human exons analysed with respect to rat exons (644,354 and 236,327, respectively), and to the smaller samples amount (3 human samples in total) which might have introduced a higher number of false positive results.

Because of the small number of human samples, we decided to focus only on the slightly larger number of rat samples; in particular, the 232 significant exons resulted from the comparison $RBM20^{+/+}$ vs. $RBM20^{-/-}$ were used as dataset of RBM20 affected exons (dataset A1).

3.2 The dataset A2 of RBM20 affected exons

A second dataset of RBM20 affected exons and a dataset of control exons were also studied.

A group of 97 differentially spliced rat exons and with RBM20 binding sites mapping in the flanking regions of the exons was retrieved by Maatz (Maatz et al., 2014) analysing the same Guo data I analysed (Guo et al., 2012). The 80 out of 97 exons with clusters of RBM20 binding sites in the exons flanking regions were selected and their genomic sequences were downloaded (dataset A2).

As control dataset, 80 exons were randomly selected from rat genome, so far not known to be regulated by RBM20 (dataset N1).

3.3 Searching for RBM20 RNA binding site in the target sequences

RBM20 binding sites (Figure 15) were searched in the flanking regions of datasets

A1, A2 and N1 through the algorithm based on a nucleotide probability matrix (see section 2.6).

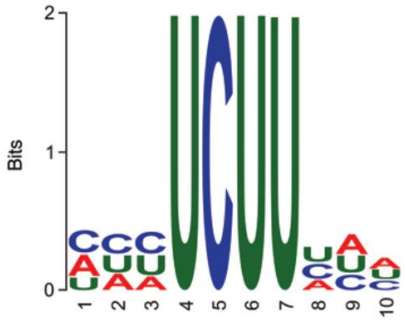

Figure 15. Sequence logo for rat RBM20 RNA binding pattern, given at nucleotide resolution (Maatz et al., 2014).

As RBM20 RNA binding site logo doesn't contain any “G” nucleotide, each target sequence was further searched for substrings without “Gs”, and then RBM20 binding site pattern/patterns were investigated only in these substrings. This analysis was more accurate with respect to the simple search for RBM20 binding site.

First of all, the flanking regions of the exons of the datasets A1, A2 and N1 were searched for all substrings which don't contain any “G” nucleotide, through an in house program (see section 2.7). Two analyses were performed by investigating exon flanking regions both of 430 and of 230 nucleotides, to evaluate a possible enrichment of substrings nearer the exons.

As through a descriptive statistic about substrings length, about 2/3 of unique substrings resulted to be shorter than 15 nucleotides, we decided to discard them and to focus our attention only on substrings longer than 15 nucleotides. RBM20 binding sites were so searched in these substrings, through the in house program (see section 2.6), analysing both 430 and 230 nucleotides exon flanking regions, to evaluate possible higher scores nearer the exons.

A descriptive statistic about the length of the selected substrings and the score related to the similarity or dissimilarity of a pattern to RBM20 binding site pattern is shown in Table 4.

Table 4 – Descriptive statistic about substrings length and score (max, mean and median). Exon flanking regions of 430 nucleotides and 230 nucleotides were analysed. "Exon side": exon upstream (u) or downstream (d) region analysed. Three parameters were evaluated: mean, median and variance of each of the 4 measures calculated for each exon flanking region of that dataset. Score is related to the similarity (high value) or dissimilarity (low value) of a sequence pattern to RBM20 binding site pattern.

Length of region of interest	Dataset	Exon side (u/d)	Parameter	Length of substrings	Max score	Mean score	Median score
430 nt	A1	u	mean	21.03	-6.34	-15.57	-15.37
			median	19	-4.02	-15.36	-14.67
			variance	58.02	13.77	9.04	12.15
		d	mean	20.6	-6.92	-15.9	-15.72
			median	18	-8.27	-15.77	-14.73
			variance	56.07	16.03	10.21	13.79
	A2	u	mean	20.63	-8.69	-17.68	-17.91
			median	18	-8.73	-17.55	-17.91
			variance	43.68	20.06	6.79	11.11
		d	mean	20.24	-9.16	-18.08	-18.55
			median	18	-8.8	-17.87	-20.73
			variance	46.49	22.32	8.44	11.51
	N1	u	mean	21.41	-9.37	-17.79	-18.03
			median	19	-8.8	-17.91	-17.95
			variance	76.45	21.89	10.3	14.37
		d	mean	20.62	-9.55	-18.08	-18.18
			median	18	-8.8	-18.04	-20.73
			variance	60.13	18.62	7.37	12.19
230 nt	A1	u	mean	20.74	-5.85	-15.09	-14.88
			median	18	-3.59	-14.63	-14.56
			variance	61.15	12.72	8.68	12.42
		d	mean	21.05	-6.96	-15.91	-15.85
			median	18	-8.25	-15.72	-14.74
			variance	65.02	18.32	11.34	14.63
	A2	u	mean	21.3	-8.05	-17.44	-17.67
			median	19	-8.73	-17.44	-17.88
			variance	51.42	18.75	6.13	10.1
		d	mean	20.17	-8.84	-17.96	-18.56
			median	18	-8.8	-17.87	-20.73
			variance	45.65	24.58	7.71	11.32
	N1	u	mean	22.08	-9.05	-17.58	-17.78
			median	19	-8.8	-17.33	-17.91
			variance	104.77	21.83	10.28	14.67
		d	mean	20.03	-9.74	-17.87	-17.84
			median	17	-8.8	-17.93	-17.93
			variance	56.69	17.83	7.18	12.11

The mean length of substrings was about 21 nucleotides and the median length was about 18 nucleotides for all datasets, for both exons sides and for both regions investigated. The control dataset N1 had higher variance of substrings length than the cases datasets A1 and A2 when focusing on 430 nucleotides flanking regions, while the variance on 230 nucleotides flanking regions resulted more variable among datasets.

Mean and median of max score are higher for dataset A1 than the other two datasets (especially in the exon upstream region), corresponding to an exact match with RBM20 binding pattern, while A2 and N1 best patterns allow 1 mismatch in the site sequence. Mean and median of mean and median scores are higher for A1 too. The variance of the three scores resulted to be higher in the exon downstream than in the exon upstream region of cases exons (datasets A1 and A2), and in the exon upstream than in the exon downstream region of control exons (dataset N1). No significant difference was noticed between the scores observed analyzing 430 or 230 nucleotides exon flanking regions.

3.4 Searching for additional genetic elements in RBM20 affected exons

Since the reported RBM20 RNA binding site is not sufficient to clearly distinguish between RBM20 affected and not affected exons, the presence of additional genetic elements which might be associated with RBM20 RNA binding site in the regulation of splicing events was investigated. First of all, target exons were searched for known genetic elements (substrings without “G” nucleotide containing RBM20 binding site patterns (see section 3.3), regulatory and transcriptional known motifs, and transposable elements), then for new genetic elements (enrichment analysis), and finally for sequence characteristics (frequency of nucleotides and dinucleotides, exons length).

3.4.1 Searching for regulatory and transcriptional known motifs

Many functionally important regions of the genome can be recognized by

searching for sequence patterns, or motifs, corresponding to binding sites for transcription factors. Differential expression of genes and exons depends on these regulatory proteins. So, identifying the motifs bound by other transcription factors than RBM20 can provide useful insights in the regulation.

Datasets A1, A2 and N1 were analysed using the online resource RegRNA2.0, through an in house program (see section 2.7), which automatically submitted the input sequence and searched for all available known motifs for *Rattus norvegicus*. Results were parsed in order to discover different single regulatory and transcriptional motifs. Analyses were performed either on 430 or 230 nucleotides regions, to evaluate a possible overall or individual enrichment of motifs nearby the target exons.

Table 5 – Overall number of regulatory and transcriptional known motifs observed in the 430 or 230 nucleotides upstream and downstream regions of the target exons.

Length of region of interest	Dataset	Num. of motifs in the upstream region	Num. of motifs in the downstream region
430 nt	A1	5064	4796
	A2	1518	1525
	N1	1618	1811
230 nt	A1	2582	2684
	A2	806	842
	N1	840	1037

Overall, 1105 different regulatory and transcriptional motifs were observed in the three dataset of exons, distributed as shown in Table 5. No indication was given so far about the kind of motifs recurring in the different regions, but only on their number.

3.4.2 Searching for transposable elements

Transposable elements are sequences of DNA which can move from a position to another in the genome, creating mutations often cause of genetic diseases. They

might play some kind of regulatory role, determining which genes are turned on and when this activation takes place (McClintock, 1965). Furthermore, specific proteins are specialised in masking cryptic splice sites created by transposable elements, in order to protect the human transcriptome from the aberrant exonization of these elements, through the binding to specific sequence patterns (Zarnack et al., 2013). Thus, target exons were searched for transposable elements to evaluate a possible enrichment of these elements in the exons whose alternative splicing is regulated by RBM20.

The datasets A1, A2 and N1 were analysed with RepeatMasker web server. Each group of exons was analysed separately and searched for all transposable elements known by the program. Through a in house program (see section 2.7) the results were extracted, and the number and the kind of transposable elements present in the flanking regions of each exon were obtained. As for the search for regulatory and transcriptional motifs, two analyses were performed: one on regions of 430 nucleotides and one on regions of 230 nucleotides, to evaluate a possible enrichment of transposable elements nearby the exons.

Table 6 – Number of transposable elements (TEs) observed in the 430 nucleotides or 230 nucleotides upstream and downstream regions of the target exons.

Length of region of interest	Dataset	Num. of TEs in the upstream region	Num. of TEs in the downstream region
430 nt	A1	96	116
	A2	12	24
	N1	46	41
230 nt	A1	38	56
	A2	3	10
	N1	19	16

The number of transposable elements observed in the three datasets (A1, A2 and N1), when considering 430 and 230 nucleotides exon flanking regions, is displayed in Table 6. As shown, many exons don’t contain any transposable element.

After observing the occurrences of each element in the target sequences, we decided to focus on 3 kinds of transposable elements (SINEs, Simple repeats and LTRs) and on the total number of interspersed repeats, evaluating the number of sequences containing or not each element.

Table 7 – Total number of nucleotides (nt) of interspersed repeats. The total number of interspersed repeats was investigated in exon upstream and downstream regions of both 430 and 230 nucleotides. "+": number of nucleotides with interspersed repeats; "-": number of nucleotides without interspersed repeats.

Transposable element	Length of region of interest	Dataset	+/-	Num. of nt in the upstream region	Num. of nt in the downstream region
Total interspersed repeats	430 nt	A1	+	6125	7797
			-	93635	91963
		A2	+	502	983
			-	33898	33417
		N1	+	3411	3503
			-	30989	30897
	230 nt	A1	+	945	2271
			-	52415	51089
		A2	+	0	156
			-	18400	18244
		N1	+	1089	824
			-	17311	17576

All the three datasets of exons presented more interspersed repeats (Table 7) in the exon downstream region than the exon upstream region, for both 430 and 230 nucleotides regions, except for the dataset N1 in the 230 nucleotides exon flanking regions (1089 nucleotides of interspersed repeats in the exon upstream regions and 824 nucleotides of interspersed repeats in the exon downstream regions).

Table 8 – Number of transposable elements (TEs). The total number of SINEs, Simple repeats and LTRs was investigated in exon upstream and downstream regions of both 430 and 230 nucleotides. "+": number of exon flanking regions with at least a SINE, a Simple repeat or a LTR; "-": number of exon flanking regions without SINEs, Simple repeats or LTRs.

Transposable element	Length of region of interest	Dataset	+/-	Num. of TEs in the upstream region	Num. of TEs in the downstream region
SINEs	430 nt	A1	+	33	36
			-	199	196
		A2	+	3	7
			-	77	73
		N1	+	22	14
			-	58	66
	230 nt	A1	+	8	14
			-	224	218
		A2	+	0	1
			-	80	79
		N1	+	9	6
			-	71	74
Simple repeats	430 nt	A1	+	33	40
			-	199	192
		A2	+	6	10
			-	74	70
		N1	+	11	13
			-	69	67
	230 nt	A1	+	19	27
			-	213	205
		A2	+	2	6
			-	78	74
		N1	+	6	6
			-	74	74
LTRs	430 nt	A1	+	6	5
			-	226	227
		A2	+	1	0
			-	79	80
		N1	+	3	4
			-	77	76
	230 nt	A1	+	1	3
			-	231	229
		A2	+	0	0
			-	80	80
		N1	+	1	2
			-	79	78

Concerning SINEs elements (Table 8), more SINEs in downstream than in upstream regions of cases exons were observed, and the opposite situation was found for control exons, for both 430 and 230 nucleotides regions.

Simple repeats distribution (Table 8) reflected interspersed repeats distribution, showing more simple repeats in the exon downstream region than the exon upstream region, both for 430 and 230 nucleotides regions, except for the control dataset in the 230 nucleotides exon flanking regions (6 simple repeats in both exon upstream and downstream regions).

LTRs elements (Table 8) showed a different behaviour: more LTRs in the upstream regions of cases exons and less LTRs in the upstream regions of control exons, than in the exon downstream regions, were observed when considering regions of 430 nucleotides; on the contrary, the opposite situation for cases exons, but the same situation for control exons, was observed when analysing regions of 230 nucleotides.

Fisher exact test was performed to evaluate a possible enrichment of each of these elements in cases or in control exons (comparisons A1-N1, A2-N1 and A1-A2).

Only the statistically significant results are shown (Table 9).

All the comparisons for the total interspersed repeats were highly significant (p-value << 0.00001), except for the comparison A1-N in the 230 nucleotides exon downstream region (p-value = 0.2). As concerning SINEs elements, significant comparisons for the exon upstream regions of both 430 and 230 nucleotides (p-value ranges from 0.000043 to 0.018, overall for both comparisons), but not for the correspondent exon downstream regions, were observed.

All the odds-ratio belonging to significant comparisons were found to be OR < 1, both for total interspersed repeats and for SINEs, for both 430 and 230 nucleotides exon flanking regions, indicating an impoverishment of these transposable elements in the exons whose alternative splicing is regulated by RBM20.

Furthermore, a higher significance, but a lower strength of association, in the 430 nucleotides regions with respect to 230 nucleotides regions, were observed.

None of the comparisons related to the presence or absence of Simple repeats or LTRs, in the flanking regions of cases and control exons, resulted to be significant; similarly, the comparison A1-A2 resulted not significant for all transposable elements analysed.

Table 9 – Fisher exact test results on the number of nucleotides (for Total interspersed repeats) or exon flanking sequences (for SINEs) containing the selected elements, given as p-value and odds-ratio (OR) for the comparisons A1-N1 and A2-N1. Regions of 430 nucleotides and 230 nucleotides were analysed. "Exon side": exon upstream (u) or downstream (d) region analysed. Only statistically significant results (p-value < 0.05) are shown. "n.s.": not significant result.

Transposable element	Length of region of interest	Exon side (u/d)	Parameter	A1-N1	A2-N1
Total interspersed repeats	430 nt	u	p-value	<<0.00001	<<0.00001
			OR	0.59	0.13
		d	p-value	<<0.00001	<<0.00001
			OR	0.75	0.26
	230 nt	u	p-value	<<0.00001	<<0.00001
			OR	0.29	0
		d	p-value	n.s.	<<0.00001
			OR	n.s.	0.18
SINEs	430 nt	u	p-value	0.01	0.000043
			OR	0.44	0.1
		d	p-value	n.s.	n.s.
			OR	n.s.	n.s.
	230 nt	u	p-value	0.018	0.0031
			OR	0.28	0
		d	p-value	n.s.	n.s.
			OR	n.s.	n.s.

3.4.3 Searching for sequence patterns more frequent in cases than in control exons

Investigating for additional possible genetic elements associated with RBM20 regulation, an enrichment analysis to search for RNA binding site patterns more frequent in RBM20 regulated than not regulated exons was performed.

Two enrichment analyses were performed, through the software DREME, between datasets A1 and A2 versus themselves with their sequences scrambled (Table 10).

Table 10 – Patterns from the enrichment analyses (DREME software). "Shuffled": dataset containing exons flanking sequences with scrambled nucleotides. Enrichment analysis was performed in the top strand or in both strands of the exons. Patterns were selected with DREME p-value < 10^{-7} and DREME E-value < 10^{-3}.

Case dataset	Control dataset	Strand analysed	Num. of patterns
A1	A1 shuffled	top strand	7
A1	A1 shuffled	both strands	14
A2	A2 shuffled	top strand	3
A2	A2 shuffled	both strands	8

Overall, 32 new binding site patterns were found from the enrichment analyses. The number of patterns observed for each analysis was proportional to the number of sequences analysed; indeed, the dataset A1 (232 exons) was enriched of more patterns than the dataset A2 (80 exons). In the same way, more patterns were observed analysing both strands of each exon flanking region, than analysing only the top strand.

Through the algorithm based on a nucleotide probability matrix (see section 2.6), these 32 new patterns were searched in the flanking regions of the exons of datasets A1, A2 and N1 (data not shown), and then used in the following analyses.

3.4.4 Estimating the frequency of nucleotides and dinucleotides

Also nucleotides and dinucleotides frequencies of target exons were calculated as additional genetic elements to be used in the following analyses (data not shown).

The datasets A1, A2 and N1 were analysed through an in house program (see section 2.7), focusing on the 430 nucleotides exon flanking regions.

3.4.5 Getting of exons length

The length of exons belonging to different datasets could be another important characteristic useful to distinguish them.

Exons length of the datasets A1, A2 and N1 was therefore calculated through an in house program (see section 2.7), and a descriptive statistic was performed for each dataset (Table 11).

Table 11 – Descriptive statistics about exons length. The 1 quartile, median, mean and 3 quartile of the length of exons of each dataset were calculated.

Dataset	Exon length			
	1 quartile	Median	Mean	3 quartile
A1	128.80	270.00	676.20	896.20
A2	87.75	168.50	317.55	279.00
N1	102.80	149.00	212.40	186.50

As displayed, the two datasets of RBM20 regulated exons (A1 and A2) contain longer exons than control exons (dataset N1), on average. Between regulated exons, dataset A1 contains averagely longer exons then dataset A2.

So, exon length could be worthy to be evaluated.

3.5 Features selection

We collected and merged all the information obtained from the previous analyses (sections 3.3 and 3.4) in a single R object for each dataset analysed (A1, A2 and N1): the results about the search for RBM20 binding site patterns and for substrings without “G” nucleotide containing RBM20 binding site patterns, the search for regulatory and transcriptional known motifs, the search for transposable elements, the search for the additional binding site patterns from enrichment analysis, the estimation of nucleotides and dinucleotides frequencies, and the getting of exons length. In particular, after the search for regulatory and

transcriptional known motifs the target regions were analysed for each different motif, while after the search for transposable elements the target regions were analysed for each kind of transposable element.

Overall, from the analyses, 9836 features (numeric characteristics) representing each exon were obtained.

Features selection was performed comparing dataset A1 with N1, and dataset A2 with N1, in order to select the subset of features able to distinguish case and control exons (see section 2.8).

The results of the features selection are shown in Table 12.

Table 12 – Number of features found at each step of the features selection. "Total features": number of features before the features selection, "Significant features": number of features after the single feature univariate association analysis, "Not redundant significant features": number of features after the correlation test, "Not redundant significant features (A1 and A2)": number of not redundant significant features found to be shared between datasets A1 and A2.

Feature kind	A1	A2
Total features	9836	9836
Significant features	478	622
Not redundant significant features	409	512
Not redundant significant features (A1 and A2)	215	

Dataset A1 showed 478 significant features and 409 not redundant significant features, while dataset A2 showed 622 significant features and 512 not redundant significant features. 215 features were observed to be shared between both sets of selected features.

In Table 13 is displayed how the 215 shared features were distributed between the genetic elements investigated.

Table 13 – Distribution of the 215 shared features among the different genetic elements analysed. The number indicates how many features related to each element were observed.

Genetic element	Number of features
RBM20 binding site patterns	11
Substrings without "G" nucleotide containing RBM20 binding site patterns	0
Regulatory and transcriptional known motifs	8
Transposable elements	1
32 binding site patterns from enrichment analysis	143
Nucleotides and dinucleotides frequencies	51
Exons length	1

As shown, the features which seemed to distinguish between case and control exons were related to RBM20 binding site patterns (11 features), to regulatory and transcriptional known motifs (8 features), to transposable elements (1 feature), to some of the 32 additional patterns (143 features overall), to nucleotides and dinucleotides frequencies (51 features), and to exons length (1 feature).

None of the features related to the presence of substrings without "G" containing RBM20 binding site was found among the shared features; this didn't imply that none of this kind of features was significant, but some features could have been not significant, some could have been significant but redundant, and others could have been significant and not redundant, but not shared.

3.6 Support Vector Machine analysis

Support Vector Machine method was used to discriminate RBM20 affected from RBM20 not affected exons.

The 215 shared features were used to predict the exons class (RBM20 affected or not affected). The 75% of the case exons of the datasets A1 and A2, one dataset at a time (RBM20 affected exons), combined together with the 75% of the control exons of the dataset N1 (RBM20 not affected exons), was used to train the SVM. The remaining 25% of the exons in A1+N1 and A2+N1 datasets was used as testing set. The performances of the models were evaluated through the AUC value, whereas the performances of the predictions were evaluated through the accuracy value (see section 2.9 for more details).

The results of SVM analyses are shown in Table 14.

Table 14 – AUC value of the model and accuracy of the prediction for the SVM classification. For each analysis, case and control datasets are indicated. The 215 shared features were used to train the SVM. $0 \leq AUC \leq 1$, $0 \leq accuracy \leq 1$.

Case dataset	Control dataset	AUC value	Accuracy
A1	N1	0.71	0.77
A2	N1	0.81	0.83

The analysis on the combined datasets A2+N1 (AUC = 0.81 and accuracy = 0.83) performed better than the analysis on the combined datasets A1+N1 (AUC = 0.71 and accuracy = 0.77).

The more probable cause of these results is the different ratio of case and control exons in the two combined datasets on which the SVM was trained. Whereas in A2+N1 the case-control ratio is 1:1 (80 affected and 80 not affected exons), the ratio for A1+N1 was about 3:1 (232 affected and 80 not affected exons). We noticed that this introduced a bias in the prediction of affected exons, classifying more than the 90% of the exons of A1+N1 as affected (result not shown), hence resulting in a lower accuracy than A2+N1.

A new SVM analysis was thus performed selecting 80 random exons out of the 232 exons of the dataset A1, in order to restore a case-control ratio of 1:1 when analyzing datasets A1+N1 (Table 15).

Table 15 – AUC value of the model and accuracy of the prediction for the SVM classification on 80 random exons of the dataset A1 and on the dataset A2. For each analysis, case and control datasets are indicated. The 215 shared features were used to train the SVM. $0 \leq AUC \leq 1$, $0 \leq accuracy \leq 1$.

Case dataset	Control dataset	AUC value	Accuracy
A1	N1	0.68	0.85
A2	N1	0.81	0.83

Despite the AUC value for the datasets A1+N1 in Table 15 was slightly lower than the respective AUC value in Table 14 (0.68 and 0.71, respectively), the accuracy value increased from 0.77 to 0.85.

In order to evaluate the predictive strength of the 215 shared features, a further analysis was performed by training and testing SVM through the permutated 215 shared features (Table 16).

Table 16 – Accuracy of the prediction for the SVM classification on 80 random exons of the dataset A1 and on the dataset A2. Case and control datasets, and the AUC value of each training set are shown. For each analysis, the different composition and the number of affected (A) and not affected (N) exons of the testing set are indicated. The composition "½ A + ½ N" is the default testing set. The 215 permutated shared features were used to train the SVM. $0 \leq AUC \leq 1$, $0 \leq accuracy \leq 1$.

Case dataset	Control dataset	AUC value	Testing set			Accuracy
			Composition	Number of A exons	Number of N exons	
A1	N1	0.65	$^1/_2$ A + $^1/_2$ N	20	20	0.55
			$^2/_3$ A + $^1/_3$ N	20	10	0.67
			$^1/_3$ A + $^2/_3$ N	10	20	0.57
			$^1/_1$ A	20	0	0.6
			$^1/_1$ N	0	20	0.5
A2	N1	0.62	$^1/_2$ A + $^1/_2$ N	20	20	0.53
			$^2/_3$ A + $^1/_3$ N	20	10	0.47
			$^1/_3$ A + $^2/_3$ N	10	20	0.63
			$^1/_1$ A	20	0	0.4
			$^1/_1$ N	0	20	0.65

The accuracy value ranged from 0.5 to 0.67 for datasets A1+N1, and from 0.4 to 0.65 for datasets A2+N1. So low values implied a small predictive capability of the 215 permutated shared features, enhancing nevertheless the original features as good predictors.

In the default setting of SVM the ratio of cases and controls is the same in the training and testing sets. As suggested by literature, the 75% of exons (60 affected and 60 not affected exons - training set) was used to train the SVM, and then the trained model was tested on the remaining 25% of exons (20 affected and 20 not affected exons - testing set). In order to evaluate how the accuracy could change according to the ratio of case and control exons in the testing set, we performed several runs changing this ratio (Table 17).

Table 17 – Accuracy of the prediction for the SVM classification on 80 random exons of the dataset A1 and on the dataset A2. Case and control datasets, and the AUC value of each training set are shown. For each analysis, the different composition and the number of affected (A) and not affected (N) exons of the testing set are indicated. The composition "½ A + ½ N" is the default testing set. The 215 shared features were used to train the SVM. $0 \leq AUC \leq 1$, $0 \leq accuracy \leq 1$.

Case dataset	Control dataset	AUC value	Testing set			Accuracy
			Composition	Number of A exons	Number of N exons	
A1	N1	0.68	$^1/_2$ A + $^1/_2$ N	20	20	0.85
			$^2/_3$ A + $^1/_3$ N	20	10	0.83
			$^1/_3$ A + $^2/_3$ N	10	20	0.83
			$^1/_1$ A	20	0	0.9
			$^1/_1$ N	0	20	0.8
A2	N1	0.81	$^1/_2$ A + $^1/_2$ N	20	20	0.83
			$^2/_3$ A + $^1/_3$ N	20	10	0.8
			$^1/_3$ A + $^2/_3$ N	10	20	0.8
			$^1/_1$ A	20	0	0.85
			$^1/_1$ N	0	20	0.8

The accuracy values for the predictions on the testing sets analysed were all high values (accuracy ≥ 0.8 for all analyses). This underlines that although an unbalanced composition of case and control exons in the training set could affect

the prediction of the testing set in favour of either case or control exons (as shown in Table 14 for datasets A1+N1), a balanced training set can predict with high accuracy also unbalanced testing sets, even when testing exons are all affected or all not affected (i. e. composition of the testing set of "$^{1}/_{1}$ A" or "$^{1}/_{1}$ N" exons).

Therefore, for all findings observed so far, these 215 shared features seemed to be sufficient to well discriminate RBM20 affected from not affected exons.

4 Discussion

Mutations of RNA binding motif protein 20 (RBM20) have been recently reported to cause Human dilated cardiomyopathy **(**DCM**)** (Brauch et al., 2009; Li et al., 2010). DCM is the major cause of heart failure and mortality around the world (Jefferies and Towbin, 2010). It is characterized by cardiac dilatation and systolic dysfunction, which is the leading cause of heart transplantation. Overall, 25–50% of DCM cases are familiar and causative mutations which have been described in more than 50 genes encoding mostly for structural components of cardiomyocytes.

RBM20 belongs to the family of the SR and SR-related RNA binding proteins which assemble in the spliceosome taking part in the splicing of pre-mRNA. RBM20 is mainly expressed in striated muscle, with the highest levels in the heart (Guo et al., 2012). Due to its involvement in DCM, RBM20 was studied to unveil its mechanism of action and its RNA targets (Guo et al., 2012; Li et al., 2013). Guo and colleagues reported a set of 31 genes showing a RBM20 dependent splicing from a whole transcriptome analysis in rats and humans (Guo et al., 2012). More recently, Maatz and colleagues reported an additional set of 18 rat genes and observed that RNA sequences recognized by RBM20 are likely to be located in the 400 nucleotides flanking the exons whose alternative splicing is regulated by RBM20 (Maatz et al., 2014). However, both the suggested RNA sequence which is recognized by RBM20 and its over-representation over the flanking regions of affected exons remain poor predictors to target genes presenting splicing events regulated by RBM20.

The aim of this work was, thus, to characterize, through a bioinformatic approach, the sequence motifs of the exons whose alternative splicing was affected by RBM20, in order to ameliorate the prediction of the genes (exons) affected by RBM20.

Public RNA-Seq data were downloaded, reads underwent a quality control and

then were aligned to a genome reference, every exon was quantified and the total expression profile was reconstructed. Through a sophisticated statistical analysis the splicing for each rat and human gene of the transcriptome was investigated, in order to obtain a dataset of RBM20 affected exons (all differentially expressed rat exons).

Our first analysis on 232 exons resulted from the differential analysis of the rat transcriptome ($RBM20^{+/+}$ versus $RBM20^{-/-}$ rats) and 80 exons retrieved from literature (Maatz at al., 2014) suggested that the consensus sequence for RBM20 RNA binding site is an important factor, but not a sufficient hallmark to specifically target RBM20 affected exons. Thus we hypothesized that other factors are needed to help RBM20 in recognizing its targets. Exons were thus queried to extract a long list of features. The feature setup involved the investigation of the nucleotide composition of the target region, the identification and counting of recurrent string patterns, and the counting of known motifs and repetitive elements reported in the databases. All this was done by writing scripts to compute the analyses and summarize the results for each exon. The more relevant features were then given as input to a Support Vector Machine (SVM). SVMs employ supervised learning algorithm which could help us to discriminate RBM20 affected from not affected exons.

The number, size and position of single patterns and clusters of patterns of RBM20 RNA binding site were, first of all, integrated as features in the SVM, altogether with the ones of the 32 new binding site patterns found to be more frequent in RBM20 regulated than not regulated exons from the enrichment analysis. Subsequently, nucleotides and dinucleotides frequencies of each flanking region, and exons length were added, having revealed on average longer RBM20 affected and shorter RBM20 not affected exons.

As RBM20 rat RNA binding site logo doesn't contain any “G” nucleotide, the flanking regions of target exons were explored to search for sequence substrings without “G” nucleotide too, as better candidates for the binding of RBM20. Focusing on the unique substrings, as the most of the unique substrings found were long less than 15 nucleotides, all substrings longer than 15 nucleotides were

selected and RBM20 binding sites were searched only in these substrings. The mean length of substrings was observed to be about 21 nucleotides and the median length about 18 nucleotides for all datasets investigated. Mean and median of max score were higher for dataset A1 than the other two datasets, corresponding to an exact match with RBM20 binding pattern, while A2 and N1 best patterns allowed 1 mismatch in the site sequence. Mean and median of mean and median scores were higher for A1 too. The variance of the three scores resulted to be higher in the downstream region than in the upstream region of cases exons, and in the upstream region than in the downstream region of control exons.

Many functionally important regions of the genome can be recognized by searching for sequence patterns, or motifs, corresponding to binding sites for transcription factors. Differential expression of genes and exons depends on these regulatory proteins. So, identifying the motifs bound by other transcription factors than RBM20 can provide useful insights in the regulation of some elements that might be associated with RBM20. From our analysis, 1105 different regulatory and transcriptional motifs were observed.

Transposable elements might play some kind of regulatory role too. Furthermore, specific proteins are specialised in masking cryptic splice sites created by transposable elements, through the binding to specific sequence patterns (Zarnack et al., 2013). Thus, target exons were searched for transposable elements, in order to evaluate a possible enrichment of these elements in the exons whose alternative splicing is regulated by RBM20.

When some of the transposable elements (SINEs, Simple repeats and LTRs) were studied in a greater detail, both cases and control exons were observed to present more interspersed repeats and simple repeats in the downstream region than the exon upstream region (both for 430 and 230 nucleotides regions). As concerning SINEs elements, more SINEs were observed in the downstream regions than in upstream regions of cases exons, and the opposite situation was discovered for control exons, both for 430 and 230 nucleotides regions. LTRs elements showed a different behaviour: when considering regions of 430 nucleotides, more LTRs in

the upstream regions of cases exons and less LTRs in the upstream regions of control exons, than in the downstream regions, were noticed; with regions of 230 nucleotides the opposite situation for cases exons, but the same situation for control exons was observed.

Total interspersed repeats were observed to be significantly less present in cases than in control exons, and SINEs elements were found to be significantly less present in the upstream region of cases exons than in the upstream region of control exons, confirming an impoverishment of these transposable elements in the exons whose alternative splicing is regulated by RBM20. Because all OR were observed to be less than 1, we hypothesized an interference made by transposable elements in the recognition of RBM20 RNA binding site and/or in the binding of RBM20 to the RNA of target exons. Comparing the two flanking regions, 430 nucleotides regions were observed to have a higher significance, but a lower strength of association, than 230 nucleotides regions.

In order to have a global view of the genetic elements binding in the flanking regions of our target exons, all the information obtained from the previous analyses were collected and merged in the form of numeric characteristics, to be used as features to train our SVM model. Overall, 9836 features representing each exon were obtained. Since many of them could be either clearly not useful or redundant features, a features analysis to select the subset of features associated with the case exons was performed. The aims of a features selection analysis are three: 1) improving the overall prediction performance, 2) providing faster and more cost-effective analysis, and 3) providing a better understanding of the underlying process that generated the observations (Guyon and Elisseeff, 2003). We decided first to assess features individually through a single feature univariate association analysis, to understand their influence on the system, and then in pairs through a correlation test, to find high correlated features.

After the features selection analysis, 409 best features comparing dataset A1 with dataset N1, and 512 best features comparing dataset A2 with dataset N1 were obtained. 215 features were found to be shared between both groups of best features. Among the shared features, features describing the RBM20 binding site

patterns, regulatory and transcriptional known motifs, transposable elements, patterns from enrichment analysis, nucleotides and dinucleotides frequencies, and exons length were observed. Any feature related to the presence of substrings without “Gs” containing RBM20 binding site patterns was found; this doesn't imply that none of this kind of features was significant or not redundant, but only that none of them was shared between the two groups of selected features. Another consideration to make is that the features of the genetic elements which are represented by more features aren’t more important than the ones of the genetic elements represented by a very few features, but every feature has the same importance in the SVM.

Once obtained the subsets of best features, they have been used to train the SVM. We chose SVM method to classify RBM20 affected and not affected exons, but other machine learning algorithms could have been used: i. e. Decision Trees (DT), Random Forest (RF), k-Nearest Neighbors (k-NN), Artificial Neural Networks (ANN), naïve Bayes (NB). However, SVM is among the most commonly used ones, not only in motif discovery, but also in neuroimaging and in diabetes research (Kavakiotis et al., 2017).

From the SVM analyses, the predictions performed with the 215 shared features resulted in high values for both the AUC of the model and the accuracy of the prediction ($0.68 \leq AUC \leq 0.81$, accuracy > 0.8). The SVM trained and tested with permutated features resulted in low values for both AUC and accuracy, thus enhancing the original features as good predictors.

Further analyses showed that although an unbalanced composition of case and control exons in the training set could affect the prediction of the testing set in favour of either case or control exons, a balanced training set can predict with good accuracy also unbalanced testing sets, even when testing exons are all affected or all not affected.

Starting from 9836 features we detected a subset of 215 features (about a half of the best features selected from each combined case-control dataset, and about 1/5 on the initial overall number of features), which represent reliable markers helping

to well discriminate RBM20 affected from not affected exons.

From a biological and functional point of view, this approach helps us to target novel candidate genes associated to diseases depending on a deregulation of RBM20.

This study, anyway, presents some limitations: our analyses were based on public RNA-Seq data, but the number of available samples was very tiny (3 human samples and 9 rat samples were analysed). Furthermore, we focused our attention mainly on rat samples (because of the greater numerosity), and thus the detection of the affected exons in the human samples remains blurred. It is also important to underline that the SVM is sensitive to the relative number of case and control samples given during the training procedure, and this may impact on the classification performance.

In the next future, we aim to reduce the number of features used to train the SVM, and to apply other machine learning algorithms for the classification of RBM20 affected exons. Additionally, we will test our SVM model on a new dataset of exons and we will investigate the human genome to find out all possible RBM20 affected exons, in order to verify the SVM model reliability.

5 Conclusions

Recently, the role of RBM20 in the cardiac function (Ma et al., 2016; Hinze et al., 2016) and its regulation of Titin (Beqqali et al., 2016; Khan et al., 2016; Jaé et al., 2016) were deepened, and other genes related to DCM were studied (Kayvanpour et al., 2017). This study provided additional information about RBM20 regulation of target exons, based not only on the RNA binding site, but also on other genetic elements associated to the binding site. Furthermore, we proposed the first model based on a SVM algorithm for the classification of RBM20 affected and not affected exons.

6 Bibliography

Adams, MD., Kelley, JM., Gocayne, J.D., Dubnick, M., et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science *252*, 1651-1656.

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., et al. (1991). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature *377*, 3-174.

Agarwal, R., Malhotra, P., Awasthi, A., Kakkar, N., and Gupta, D. (2005). Tuberculous dilated cardiomyopathy: an under-recognized entity? BMC Infect. Dis. *5*, 29.

Alpaydin E. (2004). Introduction to machine learning. Cambridge Massachusetts London England: The MIT Press.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Anders, S., McCarthy, D.J., et al. (2013). Count-based differential expression analysis of RNA sequencing data using R and bioconductor. Nature protocols *8*, 1765-1786.

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq - a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res *22*, 2008-2017.

Battle, A., Mostafavi, S., et al. (2013). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 992 individuals. Genome Res *24*, 14-24.

Belhassen, B. (April 2005). Radiofrequency ablation of "benign" right ventricular outflow tract extrasystoles: a therapy that has found its disease? J. Am. Coll. Cardiol. *45*, 1266–1268.

Beraldi, R., Li, X., Martinez Fernandez, A., Reyes, S., Secreto, F., Terzic, A., et al. (2014). Rbm20-deficient cardiogenesis reveals early disruption of RNA processing and sarcomere remodeling establishing a developmental etiology for dilated cardiomyopathy. Hum. Mol. Genet. *23*, 3779–3791.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer, ISBN 0-387-31073-8.

Black, Douglas L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annual Review of Biochemistry *72*, 291–336.

Brauch, K.M., Karst, M.L., Herron, K.J., de Andrade, M., Pellikka, P.A., Rodeheffer, R.J., Michels, V.V., and Olson, T.M. (2009). Mutations in ribonucleic acid binding protein gene cause familial dilated cardiomyopathy. J. Am. Coll. Cardiol. *54*, 930–941.

Cáceres, J.F., Misteli, T., Screaton, G.R., Spector, D.L., and Krainer, A.R. (1997). Role of the modular domains of SR proteins in subnuclear localization and alternative splicing specificity. J. Cell. Biol. *138*, 225-238.

Casneuf, T., Van de Peer, Y., and Huber, W. (2007). In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. BMC Bioinformatics *8*, 641.

Chang, T.H., Huang, H.Y., Hsu, J.B., Weng, S.L., Horng, J.T., and Huang, H.D. (2013). An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. BMC Bioinformatics *14* Suppl 2:S4.

Coughlin, S.S., Labenberg, J.R., and Tefft, M.C. (March 1993). Black-white differences in idiopathic dilated cardiomyopathy: the Washington DC dilated Cardiomyopathy Study. Epidemiology *4*, 165–172.

Dash, M., and Liu, H. (1997). Feature selection for classification. Intelligent data analysis *1*, 131-156.

Doblin, A., Davis, C.A., et al. (2013). STAR: ultrafast universal RNA-Seq aligner. Bioinformatics *29*, 15-21.

Eksi, R., Li, H.D., Menon, R., Wen, Y., Omenn, G.S., Kretzler, M., and Guan, Y. (Nov 2013). Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data. PLoS Computational Biology *9*, e1003314.

Fairbrother, W.G., Holste, D., Burge, C.B., Sharp, and P.A. (2004). Single nucleotide polymorphism–based validation of exonic splicing enhancers. PLoS Biol. *2*, e268.

Fehrmann, R.S.N., Jansen, R.C., et al. (2011). Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. *7*, e1002197.

Fu, G.K., Xu, W., et al. (2014). Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparation. Proc. Natl. Assoc. Sci. *111*, 1891-1896.

Grabherr, Manfred G., Haas, Brian J., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology *29*, 644–652.

Grant, C.E., Bailey, T.L., and Stafford, Noble W. (2011). FIMO: Scanning for occurrences of a given motif. Bioinformatics *27*, 1017-1018.

Grant, G.R., Farkas, M.H., et al. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics *27*, 2518-2528.

Guo, W., Schafer, S., Greaser, M.L., Radke, M.H., Liss, M., Govindarajan, T., Maatz, H., Schulz, H., Li, S., Parrish, A.M., et al. (2012). RBM20, a gene for hereditary cardiomyopathy, regulates titin splicing. Nat. Med. *18*, 766–773.

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. J. Mach. Learn. Res. *3*, 1157–1182.

Illumina. (2016). De Novo Assembly Using Illumina Reads (PDF).

Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (August 2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc. *7*, 1534–1550.

Itoh, K., Matsubara, K., and Okubo, K. (1994). Identification of an active gene by using large-scale cDNA sequencing. Gene *140*, 295-296.

Jameson, J.N., Kasper, D.L., Harrison, T.R., Braunwald, E., Fauci, A.S., Hauser, S.L., and Longo, D.L. (2005). Harrison's principles of internal medicine (16th ed.). New York: McGraw-Hill Medical Publishing Division.

Jefferies, J.L., and Towbin, J.A. (2010). Dilated cardiomyopathy. Lancet. *375*, 752–762.

Joshi, N.A., and Fass, J.N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. Available at https://github.com/najoshi/sickle.

Katz, Y., Wang, E.T., et al. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods *7*, 1009-1015.

Ke, S., Zhang, X.H., and Chasin, L.A. (2008). Positive selection acting on splicing motifs reflects compensatory evolution. Genome Res. *18*, 533–543.

Keren, H. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nature Reviews Genetics *11*, 345-355.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology *14*, R36.

Kingsford, C., and Patro, R. (June 2015). Reference-based compression of short-read sequences using path encoding. Bioinformatics *31*, 1920–1928.

Kohavi, R., and Provost, F. (1998). Glossary of terms. Machine Learning *30*, 271–274.

Kukurba, K.R., and Montgomery S.B. (2015). RNA sequencing and analysis. Cold Spring Harb Protoc.

Langmead, B., and Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods *9*, 357-359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology *10*, R25.

Lappalainen, T., Montgomery, S.B., et al. (2011). Epistatic selection between coding and regulatory variation in human evolution and disease. Am. J. Hum. Genetics *89*, 459-463.

Lappalainem, T., Sammeth, M., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506-511.

Li, D., et al. (2010). Identification of novel mutations in RBM20 in patients with dilated cardiomyopathy. Clin. Transl. Sci. *3*, 90–97.

Li, H., Lovci, M.T., Kwon, Y.S., Rosenfeld, M.G., Fu, X.D., and Yeo, G.W. (2008). Determination of tag density required for digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model. Proc. Natl. Acad. Sci. USA *105*, 20179–20184.

Li, S., Guo, W., Dewey, C.N., and Greaser, M.L. (2013). Rbm20 regulates titin alternative splicing as a splicing repressor. Nucleic Acids Res. *41*, 2659–2672.

Lim, K.H., Ferraris, L., Filloux, M.E., Raphael, B.J., and Fairbrother, W.G. (2011). Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc. Natl. Acad. Sci. USA *108*, 11093–11098.

Linke, W.A., and Bücker, S. (2012). King of hearts: a splicing factor rules cardiac proteins. Nature Medicine *18*, 660-661.

Liu, H., and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng. *17*, 491-502.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G., and Guigó, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? FEBS Letters *579*, 1900–1903.

Lu, B., Zeng, Z., and Shi, T. (February 2013). Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. Science China Life Sciences *56*, 143–155.

Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., and Misteli, T. (2011). Epigenetics in alternative pre-mRNA splicing. Cell *144*, 16–26.

Maatz, H., Jens, M., Liss, M., Schafer, S., Heinig, M., Kirchner, M., et al. (2014). RNA-binding protein RBM20 represses splicing to orchestrate cardiac pre-mRNA processing. J. Clin. Invest. *124*, 3419-3430.

MacArthur, D.G., Balasubramanian, S., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823-828.

Maher, C.A., Kumar-Sinha, C., Cao, X., et al. (March 2009). Transcriptome sequencing to detect gene fusions in cancer. Nature *458*, 97–101.

Mahon, N.G., Murphy, R.T., MacRae, C.A., Caforio, A.L., Elliott, P.M., and McKenna, W.J. (July 2005). Echocardiographic evaluation in asymptomatic relatives of patients with dilated cardiomyopathy reveals preclinical disease. Annals of Internal Medicine *143*, 108–115.

Majewski, J., and Pastinen, T. (2011). The study of eQTL variations by RNA-Seq: from SNPs to phenotypes. Trends Genet. *27*, 72-79.

Martino, T.A., Liu, P., and Sole, M.J. (February 1994). Viral infection and the pathogenesis of dilated cardiomyopathy. Circ Res. *74*, 182–188.

Marwan, M., Refaat, M.D., Steven, A., Lubitz, M.D., et al. (March 2012). Genetic Variation in the Alternative Splicing Regulator, RBM20, is associated with Dilated Cardiomyopathy. Heart Rhythm. *9*, 390–396.

Marx, V. (June 2013). Biology: the big challenges of big data. Nature *498*, 255–260.

Matlin, A.J., Clark, F., and Smith, C.W.J. (May 2005). Understanding alternative splicing: towards a cellular code. Nature Reviews *6*, 386–398.

Mattmann, C.A. (January 2013). Computing: a vision for data science. Nature *493*, 473–475.

McClintock, B. (1965). Components of action of the regulators Spm and Ac. Carnegie Institution of Washington Year Book *64*, 527–536.

Mitchell, R.S., Kumar, V., Abbas, A.K., and Fausto, N. (2007). Robbins Basic Pathology (8th ed.). Philadelphia: Saunders.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). Foundations of Machine Learning. MIT Press. ISBN 978-0-262-01825-8.

Montgomery, S.B., Lappalainen, T., et al. (2011). Rare and common regulatory variation in population-scale sequenced human genomes. PLoS Genet. *7*, e1002144.

Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., et al. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773-777.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods. *5*, 621–628.

Oshlack, A., and Wakefield, M.J. (2009). Transcript length bias in RNA-Seq data confounds systems biology. Biol. Direct *4*, 14.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (Dec 2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics *40*, 1413–1415.

Parks, S.B., Kushner, J.D., Nauman, D., et al. (2008). Lamin a/c mutation analysis in a cohort of 324 unrelated patients with idiopathic or familial dilated cardiomyopathy. Am. Heart. J. *156*, 161–169.

Pastinen, T. (2010). Genome-wide allele-specific analysis: insights into regulatory variation. Nat. Rev. Genetics *11*, 533-538.

Pickrell, J.K., Marioni, J.C., et al. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768-772.

Reid, D.C., Chang, B.L., Gunderson, S.I., Alpert, L., Thompson, W.A., and Fairbrother, W.G. (2009). Next-generation SELEX identifies sequence and structural determinants of splicing factor binding in human pre-mRNA sequence. RNA *15*, 2385–2397.

Robbins, S.L., Kumar, V., and Cotran, R.S. (2003). Robbins basic pathology (7th ed.). Philadelphia: Saunders.

Roberts, A., Trapnell, C., et al. (2011b). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. *12*, R22.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). Edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Rudloff, U., Bhanot, U., et al. (2010). Biobanking of human pancreas cancer tissue. Impact of ex-vivo procurement times on RNA quality. Ann. Surg. Oncol. *17*, 2229-2236.

Russell, S., and Norvig, P. (2003). Artificial Intelligence: A Modern Approach (2nd Ed.). Prentice Hall. ISBN 978-0137903955.

Sahebi, M., Hanafi, M.M., van Wijnen, A.J., Azizi, P., and Abiri, R. (2016). Towards understanding pre-mRNA splicing mechanisms and the role of SR proteins. Gene *587*, 107-119.

Sammeth, M., Foissac S., et al. (2008). A general definition and nomenclature for alternative splicing events. PLoS Comput. Biol. *4*, e1000147.

San Martín, M.A., García, A., Rodríguez, F.J., and Terol, I. (May 2002). Dilated cardiomyopathy and autoimmunity: an overview of current knowledge and perspectives. Rev. Esp. Cardiol. (in Spanish) *55*, 514–524.

Schena, M., Shalon, D., et al. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-470.

Schönberger, J., and Seidman, C.E. (August 2001). Many roads lead to a broken heart: the genetics of dilated cardiomyopathy. American Journal of Human Genetics *69*, 249–260.

Shendure, J. (2008). The beginning of the end for microarrays? Nat. Methods *5*, 585-587.

Shiraishi, H., Ishibashi, K., Urao, N., et al. (November 2002). A case of cardiomyopathy induced by premature ventricular complexes. Circ. J. *66*, 1065–1067.

Shiraki, T., Kondo, S., Katayama, S., et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. *100*, 15776-15781.

Smit, A.F.A., Hubley, R., and Green, P. RepeatMasker at http://repeatmasker.org .

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (February 2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature Protocols *7*, 500–507.

Taggart, A.J., De Simone, A.M., Shih, J.S., Filloux, M.E., and Fairbrother, W.G. (2012). Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. Nat. Struct. Mol. Biol. *19*, 719–721.

Thompson, K.L., Pine, P.S., et al. (2007). Characterization of the effect of sample quality degraded rat liver RNA. BMC Biotechnol. *7*, 57.

Tomita, H., Vawter, M.P., et al. (2004). Effect of agonal and postmortem factors on gene expression profile: quality control in microarray analyses of postmortem human brain. Biol. Psychiatry *55*, 346-352.

Trapnell, C., Hendrickson, D.G., et al. (2013). Differential analysis of gene regulation at transcript resolution with RNA-Seq. Nat. Biotechnol. *31*, 46-53.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Williams, Ba., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoforms switching during cell differentiation. Nat. Biotechnol. *28*, 511-515.

Tuch, B.B., Laborde, R.R., et al. (2010). Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. PLoS ONE *5*, e9317.

Vapnik, V.N. (1995). The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA.

Velculescu, V.E., Zhang, L., Volgelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. Science *270*, 484-487.

Villard, E., Duboscq-Bidot, L., Charron, P., et al. (2005). Mutation screening in dilated cardiomyopathy: Prominent role of the beta myosin heavy chain gene. Eur. Heart. J. *26*, 794–803.

Wang, K., Singh, D., et al. (2010a). MapSplice: accurate mapping of RNA-Seq reads for splice junction discovery. Nucleic Acids Res. *38*, e178.

Wang, L., Feng, Z., et al. (2010b). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics *26*, 136-138.

Wang, Z., and Burge, Cb. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA *14*, 802–813.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57-63.

Ward, A.J., and Cooper, T.A. (2010). The pathobiology of splicing. J. Pathol. *220*, 152–163.

Warf, M.B., and Berglund, J.A. (2010). Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem. Sci. *35*, 169–178.

Westra, H.J., Peters, M.J., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease association. Nat. Genetics *45*, 1238-1243.

Wilson, R.A., and Keil, F.C. (1999). The MIT encyclopaedia of the cognitive sciences. MIT Press.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873-881.

Wu, T.D., and Watanabe, C.K. (May 2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics *21*, 1859–1875.

Zarnack, K., König, J., Tajnik, M., Martincorena, I., et al. (2013). Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements. Cell *152*, 453–466.