2013-04-17

Computational Modeling and Inference of Alternative Splicing Regulation.

Ji Wen
University of Miami, baeywen@gmail.com

UNIVERSITY OF MIAMI

COMPUTATIONAL MODELING AND INFERENCE OF ALTERNATIVE SPLICING REGULATION

By

Ji Wen

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

May 2013

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

COMPUTATIONAL MODELING AND INFERENCE OF ALTERNATIVE SPLICING REGULATION

Ji Wen

Approved:

____
Xiaodong Cai, Ph.D.
Associate Professor of Electrical and Computer Engineering

____
Kamal Premaratne, Ph.D.
Professor of Electrical and Computer Engineering

____
Dimitris Papamichail, Ph.D.
Assistant Professor of Computer Science

____
M. Brian Blake, Ph.D.
Dean of the Graduate School

____
Mei-Ling Shyu, Ph.D.
Associate Professor of Electrical and Computer Engineering

____
Nigel John, Ph.D.
Lecturer of Electrical and Computer Engineering

WEN, JI (Ph.D., Electrical and Computer Engineering)

Computational Modeling and Inference of Alternative Splicing Regulation. (May 2013)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Xiaodong Cai.
No. of pages in text. (146)

Alternative splicing of precursor mRNA (pre-mRNA) provides an important means of regulating gene expression and generating transcriptomic and proteomic diversity in most eukaryotes. A number of special proteins, named splicing factors, can regulate the alternative splicing process by binding to certain short subsequences on pre-mRNA, named splicing regulatory elements (SREs). Therefore, identification of these SREs and prediction of their combinatorial effects are very important to the understanding of the mechanisms that regulate splicing.

In this dissertation, we develop two methods for identifying SREs and their interactions. In the first method, we use the traditional enrichment-based approach, which identifies SREs by comparing frequencies of all hexamers in two discriminative data sets generated from mouse RNA-Seq data. The SREs are identified as hexamers that are enriched in the positive data set but under-represented in the negative data set. We also analyze the position preference of the identified SREs and compare their frequencies in constitutive exons and alternatively spliced exons.

In the second method, we first derive a mathematical model for splicing regulation based on the principles of thermodynamics. We include the effects of both SREs and interactions between two SREs in the model. We then apply the model to identify

SREs and SRE interactions with linear regression. Since the linear regression model contains a very large number of variables, the traditional inference method does not perform well. To overcome this problem, we develop a novel framework for inferring the high-dimensional linear model.

Finally, we systematically study the alternative regions, arising from alternative splicing, alternative first exon or alternative last exon events in 105 breast cancer patients using RNA-Seq data. The identified aberrant alternative regions show very interesting associations with cancer development and provide important candidates for cancer diagnosis and cancer therapies.

To my beloved parents

for their endless love, understanding, support and encouragement

Acknowledgements

The research documented in this dissertation was carried out at the Department of Electrical and Computer Engineering, University of Miami, during 2008-2012. I am grateful to all who assisted and encouraged me in my graduate program studies.

I express my deepest appreciation to my advisor, Dr. Xiaodong Cai, who introduced me into the field of alternative splicing regulation. His support, guidance, encouragement and inspiration have been invaluable to my doctoral studies. I feel very fortunate and honored to have had the opportunity to work with him.

I also extend my sincerest thanks to my committee members, Dr. Kamal Premaratne, Dr. Mei-Ling Shyu, Dr. Nigel John and Dr. Dimitris Papamichail for attending my oral defense and for their insightful comments which improved my thesis.

I extend my thanks to the rest of the faculty and staff members of the Department of Electrical and Computer Engineering at the University of Miami. They helped me enormously with my education and generously provided various resources to further my research.

In addition, I thank my friends and colleagues, especially, Dr. Yu Zhang, Dr. Zhouyi Xu, Dr. Chao Chen, Anhui Huang, Tao Meng, Dianting Liu, Qiusha Zhu, Rosamund Coutts, Michelle Perez and all my other friends for their assistance and collaboration during my Ph.D studies.

Ji Wen

University of Miami

May 2013

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Biological Background

1.1.1 DNA, Gene and RNA

The functionalities of living cells in any organism are determined by many different types of molecules, particularly proteins. All the genetic information needed to synthesize protein molecules, which can be passed from generation to generation, is encoded in deoxyribonucleic acid (DNA) molecules [3]. There are four types of nucleotides in DNA including adenine (A), cytosine (C), guanine (G) and thymine (T) nucleotides. A DNA molecule consists of two chains of nucleotides forming base pairs (bp) between A and T, or C and G. All the DNA molecules in a cell form the genome of the cell or the organism. The process of producing proteins or ribonucleic acid (RNA) molecules from the information in a gene which is a segment of the genome is called gene expression. Gene expression in eukaryotes typically consists of three steps: transcription, splicing and translation as illustrated in Figure 1.1. First, the information in the gene to be expressed is transcribed to produce precursors of messenger RNA (pre-mRNA) molecules, which are then spliced to give rise to mature messenger RNA (mRNA) molecules. Finally, mRNAs are translated to yield proteins.

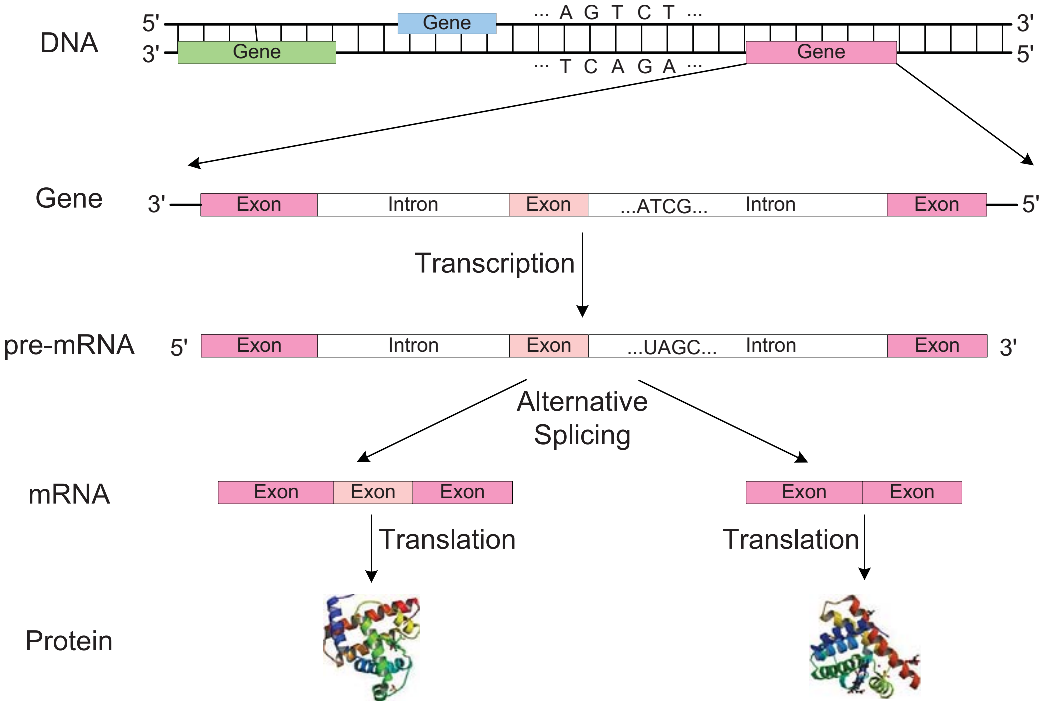


Figure 1.1: A sketch of gene expression that is proceeded in three steps: transcription, splicing and translation. Three genes are contained in this DNA segment. The last gene is roomed in to show the details of its three exons and two introns. This gene can produce two different mRNAs due to alternative splicing.

The set of all mRNA molecules that are produced in one or a population of the same type of cells is called transcriptome.

1.1.2 Alternative Splicing

In higher eukaryotes, protein coding genes are first transcribed to produce pre-mRNAs that need to be spliced out to produce mRNAs before being translated into proteins. The nucleotide sequence within a pre-mRNA that is removed by the splicing process forms introns and the remaining portion of the mRNA forms exons. The upstream margin of an intron is called 5' splice site (5' SS) or donor site, and the downstream margin of an intron is called 3' splice site (3' SS) or acceptor site.

In the human or mouse genome, there are about 30,000 protein-coding genes

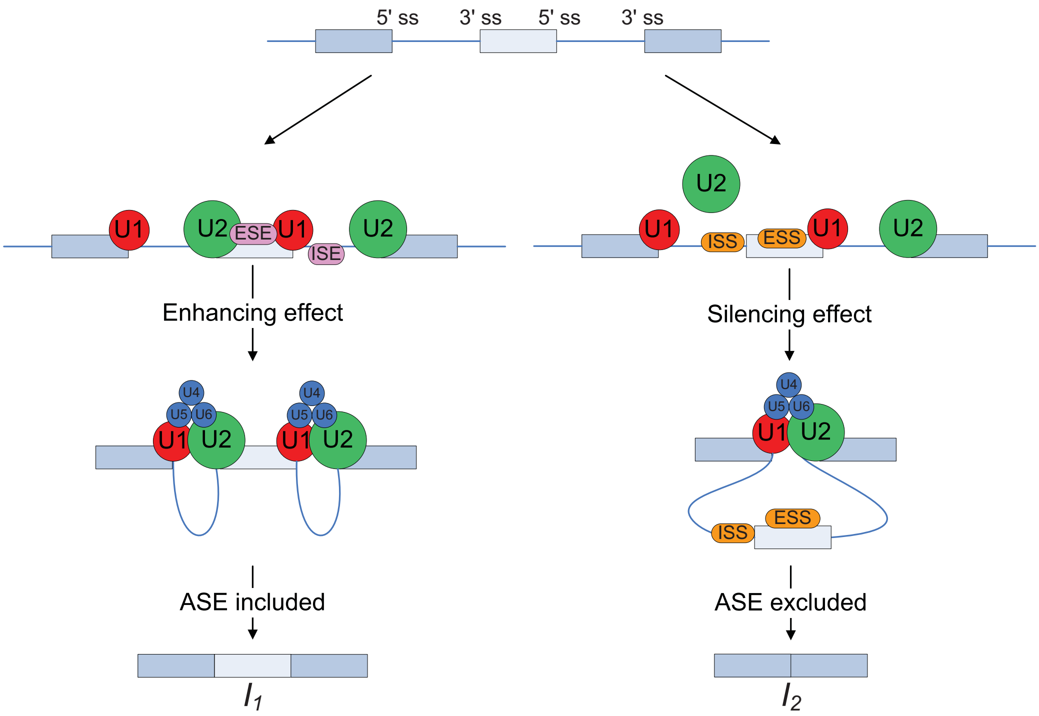


Figure 1.2: The alternative splicing process. The spliceosome assembly and alternative splicing regulation for an ASE is shown. The middle exon in the pre-mRNA is the alternative spliced exon. The other two exons are constitutive exons. 5’ (3’) ss stands for 5’ (3’) splice site.

[4] and about twice of different proteins encoded by these genes [5]. The process responsible for this discrepancy is alternative splicing (AS) that generates more than one proteins by selectively including different combinations of exons into the mRNA. Different mRNAs generated from one pre-mRNA are called isoforms. For example, in Figure 1.1, there are two isoforms for the third gene, which could be used to encode two similar but distinct proteins. The exon that is selectively included in an mRNA is called alternative spliced exon (ASE) or cassette exon, the other two exons that are included in all isoforms are called constitutive exons. ASE is the most common type of alternative splicing events. Other splicing events include alternative 5’ splice site usage, alternative 3’ splice site usage, intron retention, mutually exclusive exons, alternative first exon and alternative last exon.

The basal machinery of splicing is known as spliceosome, a large multicomponent ribonucleoprotein complex having U1, U2, U4, U5 and U6 small nuclear ribonucleoproteins (snRNPs) as its main building blocks [6]. Splicing begins with a multi-step process of spliceosome assembly around the splice sites (Figure 1.2). The accurate recognition of splice sites by the subunits of the spliceosome and the consequent assembly of the spliceosome are influenced by two factors.

The first factor is the splice site strength. The nucleotide sequences around the splice sites are far from random so that the spliceosome can recognize and bind to them. However, there is still small difference among each splice site which can affect the spliceosome binding probability or binding strength. This strength generally correlates with a "consensus value" which evaluates the conservation level of the splice site sequence. For example, Figure 1.3 illustrates the 5' splice site consensus that are constructed from 49,778 human 5' splice site [7]. The hight of A, T, C or G at each position represents the probability of the nucleotide presence at that position. The definition of "strong" and "weak" splice site usually refer to the consensus value. Thus the sequence CAG/GTAAGT has a high consensus value (strong splice site) and sequence GTG/GTGGGG has a lower consensus value (weak splice site). Generally speaking, alternative splice sites are slightly weaker than constitutive splice sites [8].

The second factor is a class of RNA binding proteins named splicing factors (SFs) that can bind to the *cis*-acting splicing regulatory elements (SREs) on the pre-mRNA. SFs can regulate splicing by facilitating or inhibiting the subunits of spliceosome to recognize the splice sites [9–11]. They can also regulate splicing through other mechanisms such as regulation of the transition from exon definition to intron definition [12, 13]. Moreover, multiple SFs and the spliceosome can interact cooperatively or antagonistically to affect the splicing process. According to the effects of SFs and

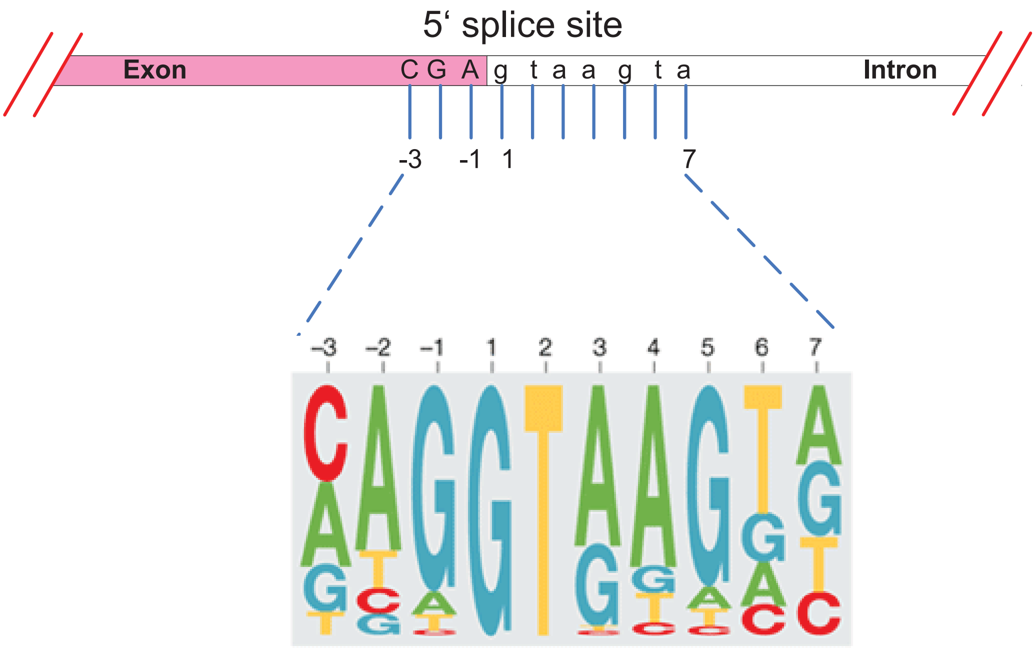


Figure 1.3: 5’ splice site profile constructed from 49,778 human 5’ss.

their binding positions, SREs can be classified as exonic splicing enhancers (ESEs) or silencers (ESSs) if they promote or inhibit the inclusion of the exon where they reside, and as intronic splicing enhancers (ISEs) or silencers (ISSs) if they enhance or inhibit the inclusion of the exon and reside in adjacent introns [9].

1.1.3 Next-generation Sequencing and RNA-Seq

AS can produce different isoforms from one gene, and different splicing levels determines different proportions of isoforms. Thus, to study splicing, we need a reliable method to measure the relative expression level of each isoform of a gene. RNA-Seq, also called whole transcriptome shotgun sequencing, is a revolutionary approach recently developed for this purpose [14].

The characterization of gene or isoform expression levels has long been of interest to many researchers. Before the advent of RNA-Seq, microarrays were the first choice of experiment for high-throughput transcriptome analysis [15]. However, microarrays

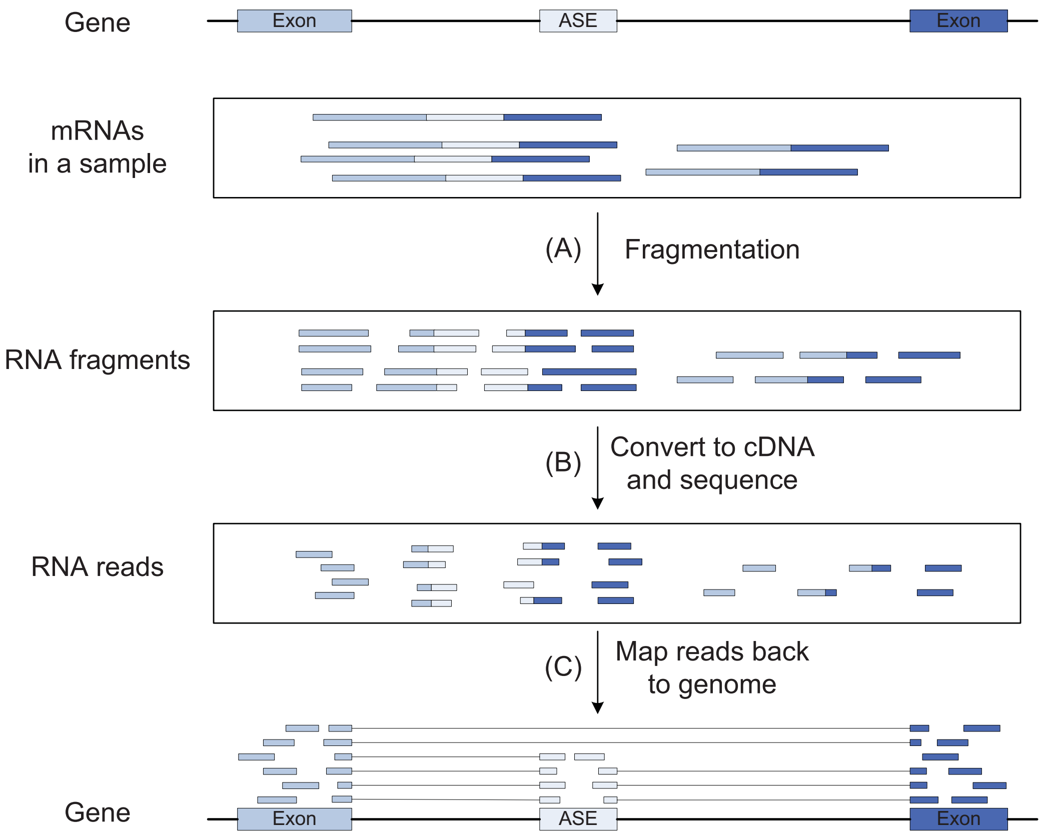


Figure 1.4: Overview of a typical RNA-Seq experiment and estimation of isoform expression levels. The schematic gene has two constitutive exons and one ASE. (A) All the mRNAs in a sample or cell is initially fragmented into small segments. (B) By random priming, the fragments are converted into cDNA fragments suitable for sequencing. A high-throughput sequencing technology is then applied to sequence these fragments to get millions of short reads. (C) Sequenced reads are mapped back to the genome to estimate the expression levels of the gene and its isoforms.

have several limitations compared to RNA-Seq. First, design of microarrays relies on existing knowledge about genome sequence. Second, microarray experiments have limited dynamic range of detection owing to background noise and saturation of signals [14]. Due to these limitations, more and more biologists choose RNA-Seq as their major tools instead of microarrays for their studies since the first generation of RNA-Seq studies published in 2008 [16, 17]. A simple illustration of RNA-Seq is shown in Figure 1.4.

RNA-Seq provides a powerful way to measure gene expression levels. In an RNA-

Seq experiment, the sequencing machine can output millions of short nucleotide sequences, named sequencing reads. The total number of reads produced from an experiment depends on the sequence depth, which is the average number of reads containing a given nucleotide in the transcriptome. Since each read is a specific combination of nucleotides, we can infer the location of each read in the genome using a sequence mapping algorithm. After mapping these short reads back to the genome, we can count the number of reads that are mapped to each gene to estimate the relative numbers of mRNA molecules, *i.e.*, the gene expression level of the gene. The simplest estimation algorithm is based on a reasonable uniform sampling assumptions, *i.e.*, the number of reads that can be mapped to a gene is approximately proportional to the total length of all the mRNA molecules of that gene and the sequencing depth. Therefore, the expression level of a gene can be measured in the unit of reads per kilobase of the transcript per million mapped reads (RPKM) [16].

RNA-Seq can also be used to estimate the expression levels of different isoforms. The simplest way is to count the reads that can be only mapped to a specific isoform, for example, the reads mapped to the different splice junctions and the reads mapped to the ASE in Figure 1.4. A better algorithm can also incorporate information from the reads that mapped to constitutive exons to increase estimation accuracy [18].

1.2 Motivation and Objectives

AS is a crucial step in the expression of most eukaryotic genes. It provides an important means of regulating gene expression and generating transcriptomic and proteomic diversity. Recent studies have found that ∼95% of human genes undergo AS [17, 19]. The importance of AS is highlighted recently by the findings that AS related mutations can cause many human diseases including cancer [20,21]. Therefore, Understanding of how splicing is regulated has draw much attention in the past decades, especially after the advent of RNA-Seq technology.

Several elements in pre-mRNA are important in splicing regulation. In addition to the core splicing signals at the 5' splice site, the 3' splice site and the branch point (A short nucleotide sequence required for splice site recognition), other splicing regulatory elements (SREs) including ESEs, ESSs, ISEs, and ISSs, are pivotal to ensure that splicing events occur accurately and efficiently [9, 10]. Identification of these SREs is of fundamental importance. Several experimental approaches such as systematic evolution of ligands by exponential enrichment (SELEX) [22], UV crosslinking and immunoprecipitation (CLIP) [23] and splicing reporter system [24], have been employed to identify SREs. However, experimental approaches can only identify SREs in a relatively small scale, and is labor-intensive. On the contrary, computational approaches provide a large-scale and efficient means to identify putative SREs that can be validated experimentally. Thus, the first *objective* of this research is to develop a reliable computational approach to identify SREs that are responsible for splicing regulation.

In addition to single SREs, cooperative interactions between multiple SREs is also important to splicing regulation. Interaction between different molecules is very common in biological systems. Many experimental results have suggested cooperative bindings or antagonistic effects between multiple SFs [25]. These interactions can cause their binding to SREs with some specific features that can be captured by computational approaches. For this reason, the second *objective* of this research is to derive a model for splicing regulation and develop a computational approach to identify the interactions between SREs.

Our last *objective* of this research is to identify aberrant splicing events that contribute to tumorigenesis. A large number of somatic mutations accumulate during the process of tumorigenesis. Many of them can cause aberrant gene expression level or generate aberrant proteins in tumor cells. However, more and more evidences show that aberrant splicing events are also prevalent in tumor cells [21]. The availability of RNA-Seq data for hundreds of individuals generated by The Cancer Genome Atlas (TCGA) [26] provides opportunities to investigate aberrant splicing events in tumors. Moreover, distinguishing the potential driver splicing events (contributing to tumor progression) to passenger splicing event (being effectively neutral) is also an important problem in tumorigenesis.

1.3 Contributions

The major contributions of this dissertation are listed in the following:

1. Computational Identification of Tissue-Specific Alternative Splicing Elements in Mouse Genes From RNA-Seq.

- Developed a novel strategy that applied a powerful discriminative approach to identify tissue-specific SREs using mouse RNA-Seq data.
- Analyzed the position distribution of SREs and found that a dozen of SREs were biased to a specific region.

2. A Thermodynamics-Based Model for Identifying Splicing Regulatory Elements and Their Interactions.

- Derived for the first time a thermodynamics-based model for AS regulation and applied it to identify SREs and their interactions in regulating AS.
- Developed a systematic and effective framework to identify SREs and SRE interactions from a large number of candidates, by incorporating the state-of-the-art techniques, and applied the framework to human RNA-Seq data.
- Identified a number of SREs and SRE pairs that are consistent with previous experimental and computational results.

3. Aberrant Isoform Expression in Cancer.

- Identified aberrant splicing, aberrant alternative first exon and aberrant last exon in cancer by comparing the RNA-Seq data in tumor cells with those in matched normal cells.
- Identified important genes with aberrant alternative regions that are associated with cancer development.

1.4 Dissertation Outline

The rest of this dissertation is organized as follows. In Chapter 2, we first give an overview of several existing SRE detection approaches. We then develop a discriminative approach to identify putative SREs in tissue-specific alternative splicing in three mouse tissues. After studying the position preference of the identified SREs in intron and exons, we compare our findings with previous computational results and experimental evidence.

In Chapter 3, we first propose a thermodynamics-based model for AS regulation, and show that our linear model is directly derived from the principles of thermodynamics, whereas inference of previous nonlinear thermodynamic model for gene transcription needs linear or nonlinear approximations. Since we need to consider a large number of candidate SREs and their interaction pairs, model inference is challenging. To overcome this challenge, we then develop a model inference framework that can effectively identify SREs and SRE pairs without overfitting the model to give a large false positive rate. Finally, we compare the performance of our model with previous transcription model and discuss several examples that are well-supported by previous computational and experimental result.

In Chapter 4, we first introduce current efforts and discoveries in cancer biology, then a new concept of alternative region is introduced and a new framework are proposed based on publicly available RNA-Seq data to identify the alternative regions that are aberrantly regulated in tumor cells compared to matched normal cells. The significantly aberrant genes and their functional enrichment are discussed to highlight the discovery of this framework.

In Chapter 5, we summarize our findings in the current work, and discuss possible future work.

CHAPTER 2

Identification of Tissue-Specific Alternative Splicing Elements in Mouse Genes

2.1 Motivation

Computational approaches provide an effective means of identifying putative SREs that can be validated experimentally. According to the regulatory pattern of splicing factors, we can classify the SREs into constitutive SREs and tissue-specific SREs, which are bound by constitutive and tissue-specific splicing factors, respectively.

A number of constitutive SREs have been identified from constitutively spliced exons using computational methods and some of them have been demonstrated in experiments to function as predicted. Fairbrother *et. al.* [27] searched for ESEs as hexamers that are more abundant in exons with weak splice sites than in exons with strong splice sites based on the assumption that ESEs should be over-present in exons with weak splice sites. They also assumed that the hexamers are more abundant in exons compared to flanking intronic regions. The identified ESEs by Fairbrother *et. al.* are known as “RESCUE-ESEs”. Zhang *et. al.* [28] used a similar approach. The octamers were detected as ESEs (ESSs) if they were present more (less) frequently in the noncoding exons compared to both the pseudo exons and the 5’ untranslated regions (UTRs) of intronless genes. Several other methods were also developed by comparing the frequencies of short nucleotide sequences in different data sets [8, 29].

Alternative splicing plays an important role in generating tissue specificity. Recent high-throughput studies based on microarray have shown that 42% cassette exons examined are differently expressed in at least one of 48 human tissues [30]. This percentage even reaches 72% in a recent RNA-Seq study [17]. Tissue-specific alternative splicing is thought to be largely regulated by tissue-specific splicing factors and tissue-specific expression of constitutive splicing factors [12, 31]. Therefore, it is important to identify SREs that are targets of these splicing factors.

Several studies for identification of SREs related to alternative splicing have been performed. Brudno *et al.* [32] identified brain-specific intronic SRE from a relatively small data set that includes 25 brain-specific cassette exons. More recently, Castle *et al.* [30] measured the expression level of a large number of exons and exon-exon junctions in 48 human tissues using microarray, and then determined up- and down-regulated cassette exons in each tissue. From these cassette exons, they identified 143 tissue-specific motifs. Wang *et al.* [33] determined the ratio of expression level of cassette exons in different pairs of human tissues from exon arrays and used a linear regression model to identify tissue-specific SREs.

The key technique used in most computational methods for identifying SREs is to find short nucleotide sequences (typically hexamers or octamers) that are over-represented in a positive data set relative to a background data set. For example, constitutive RESCUE-ESEs [27] are hexamers that are over-represented in constitutive exons with weak splice sites comparing to introns and constitutive exons with strong splice sites. In another example [30, 32], tissue-specific SREs were identified by contrasting the frequencies of hexamers in a positive data set including cassette exons and their flanking intronic region to the frequencies of hexamers in a background set including sequences neighboring to the cassette exons. However, if a more reliable negative data set, where SREs over-represented in the positive data are most unlikely present, is used, such a discriminative approach will significantly improve the power of detecting SREs as already demonstrated in identifying transcriptional factor binding sites [34–36].

In this chapter, we used mouse RNA-Seq data [16] to determine a positive and a negative data set for each type of SREs in a specific tissue. For example, the positive data set for ESEs contains cassette exons that are *included* in the dominant isoforms of genes, while the negative data set consists of the cassette exons that are *excluded* in the dominant isoforms of genes. We then employed a discriminative approach to identify putative SREs. Since the expression level of each mRNA isoform can be calculated from the RNA-Seq data more accurately than from exon microarray data used in previous work [14, 15], our method can reliably determine the positive and the negative data sets, which enables our discriminative approach to identify SREs more reliably.

2.2 Materials and Methods

2.2.1 Overview of Our Method

Our goal is to identify short motifs that are over-represented in the region flanking alternative splicing sites in a specific tissue. Alternative splicing usually occurs at weak splicing sites with highly conservative flanking sequences [37]. So these over-represented motifs most likely function as enhancers or silencers to assist the spliceosome to make a splicing decision.

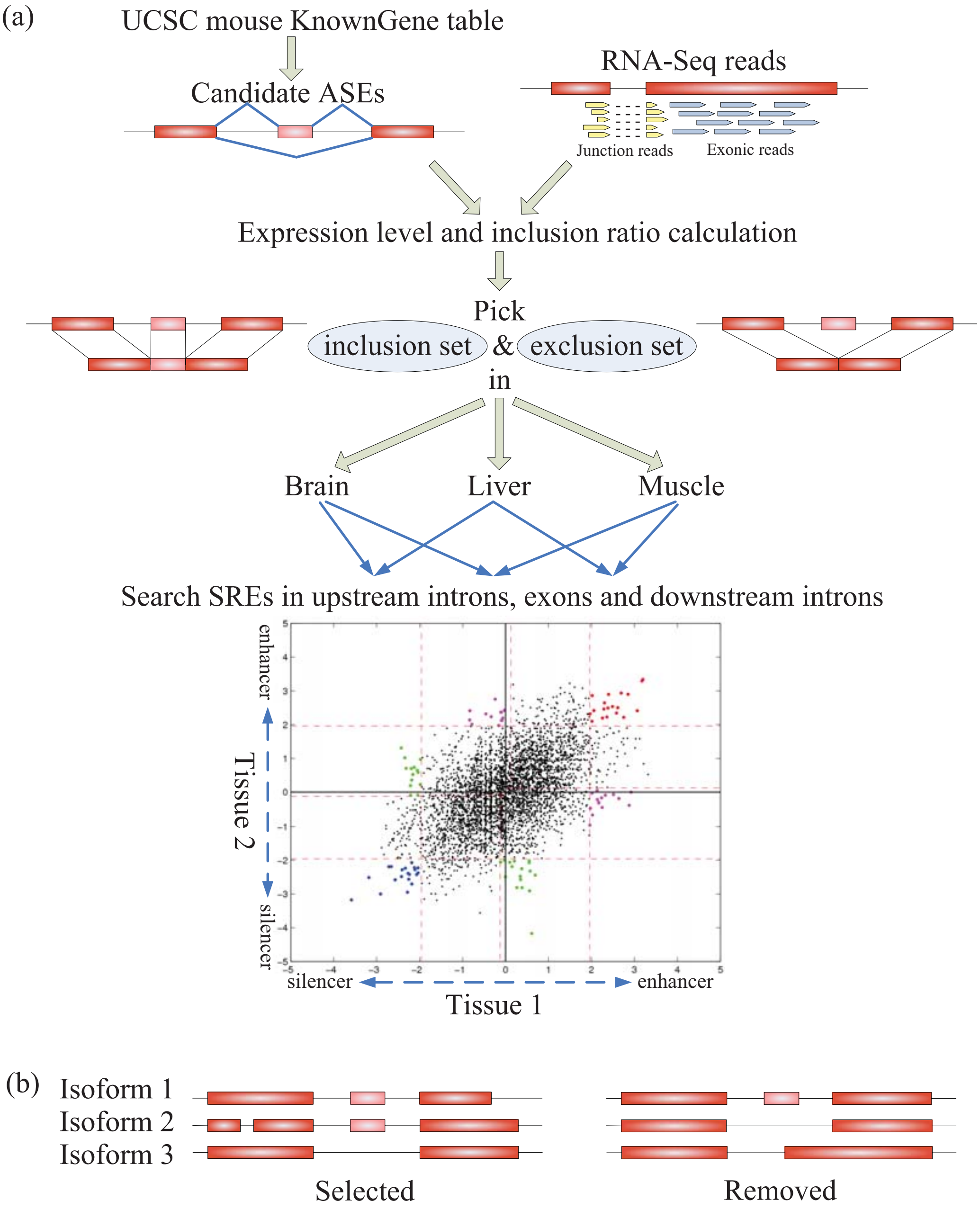


Figure 2.1: (a) Schematic flow chart for the identification of tissue-specific SREs. (b) Example of genes with more than two isoforms which were selected or excluded in our analysis. The left one was selected for further analysis since the ASE was either included or skipped in each isoform. The right one was not selected in our analysis because isoform 3 does not strictly skip the ASE.

Our approach could be divided into several steps depicted in Figure 2.1(a). We first identified cassette exons which are also referred to as alternatively spliced exons (ASEs) from the UCSC mouse KnownGene table. For a specific tissue, we then divided the set of ASEs into an inclusion set and an exclusion set as follows. Using the RNA-Seq data, we calculated the expression level of each isoform of genes that contain ASE(s). If a majority ($\geq 90\%$) of the isoforms of a gene include an ASE, then the ASE is in the inclusion set; on the other hand, if only a minority ($\leq 10\%$) of the isoforms of a gene include an ASE, then the ASE belongs to the exclusion set. The inclusion and exclusion sets also include 400 intronic nucleotides upstream and 400 intronic nucleotides downstream of the selected ASEs. For each tissue, we compared the frequency of each hexamer in the inclusion set with the frequency of the same hexamer in the exclusion set to determine if the hexamer is over-represented. The hexamers that are over-represented in one tissue but not over-represented in the other tissue are identified as putative tissue-specific SREs, while the hexamers over-represented in both tissues are identified as SREs common in both tissues. Finally, annotations in three tissues were integrated and similar putative SREs were clustered to form a motif. The selection of parameters used to construct the inclusion and exclusion set are determined by empirical research by comparing the result with experimental validated database.

2.2.2 Data Sets

Mouse RNA-Seq data of Mortazavi *et al.* [16] for 3 tissues (brain, liver and skeletal muscle) were selected in our study. The mouse genome and the KnownGene table were also downloaded from the University of California Santa Cruz genome database (UCSC) Mouse July 2007 (mm9). The mouse RNA-Seq data set contains 140 millions reads of 25 nucleotides (nt). Mortazavi *et al.* have mapped these reads against the expanded mouse genome which consists of the standard UCSC mm9 genome and the 42 nt splice-crossing sequence for each exon junction documented in the UCSC KnownGene table. Reads that could be mapped to multiple loci of the genome were excluded, and 30∼40 million uniquely mapped reads for each tissue from two replicates were used in our analysis.

We also selected 922 distinct hexamers from the database SpliceAid [38] as experimentally validated SREs for comparison purpose. SpliceAid is the latest database that collects experimentally assessed target RNA sequences bound by splicing proteins in humans. However, some sequences in SpliceAid are relatively long, and thus part of such long sequences may not be core splicing motifs. In fact, if we take hexamers from all the sequences in SpliceAid, we can get a total of 2321 distinct hexamers which may contain many false SREs. To get more reliable SREs, we only took all sequences assessed by SELEX from the SpliceAid. Since the length of the randomized sequences used in SELEX was usually larger than the length of a protein binding site, multiple alignment of the selected sequences was performed to locate the imbedded consensus sequences [22]. These consensus sequences were manually checked and extracted, which gave 922 distinct hexamers.

2.2.3 ASE Selection

We selected ASEs and some intronic nucleotides flanking the ASEs to identify SREs for the following reasons. First, AS predominantly generates ASE events in both human and mouse [16, 17]. Second, other AS events may not generate sequence data compatible to those generated by ASE events. For example, alternative 5' or 3' splice site usage lacks an alternative 3' or 5' splice site [39]. ASEs were selected from the KnownGene table with a strict criterion. An ASE was selected if at least one isoform include the ASE and at least one other isoform do not include any part of the ASE and only these two types of isoforms exist. For example, the right AS event in Figure 2.1(b) was not selected in our analysis because although an ASE was included in isoform 1 and skipped in isoform 2, this exon and its flanking regions might also contain SREs governing alternative 3' splice site since isoform 3 included this ASE partially. Different genes with overlapped open reading frame were also excluded for the simplicity and accuracy of calculating gene expression levels.

2.2.4 Calculation of Expression Level and Inclusion Ratio

Expression level for each transcript isoform of a gene was calculated with the algorithm of Jiang *et al.* [18]. This algorithm modeled the count of RNA-Seq reads falling into a region of each gene as a Poisson variable with a mean proportional to the length of the region. For an exon of length l, Jiang *et al.* used the effective exon length $l - r$ in the mean of the Poisson random variable, where r is the read length, since $l - r$ is the number of possible loci of the exon that a read could be mapped to. However, since we only kept uniquely mapped reads and excluded the ambiguous reads that could be mapped to multiple places of the genome, we used an effective exon length $l - r - m$, where m is the number of 25-nt subsequences of the exon mapped by multiple-mapped reads. To find out these multi-mappable regions, we re-mapped all possible 25-nt subsequences of candidate ASEs and splice junctions against the same expanded genome described above using Bowtie (version 0.9.9.3), an ultrafast and memory-efficient program for the alignment of short DNA sequences to a large genome [40].

After the expression level of each isoform of genes with ASEs were calculated, the inclusion ratio of an ASE in a specific tissue was calculated as the ratio of the expression level of the isoform with the ASE to the total expression level of all isoforms of the gene.

2.2.5 SRE Searching

For each tissue, all ASEs with an inclusion ratio $\geq$0.9 were put together as the exonic inclusion set, and 400 intronic nucleotides upstream or downstream of the ASEs were put together as the intronic inclusion set. All ASEs with inclusion ratio ≤ 0.1 were selected as the exonic exclusion set, and 400 intronic nucleotides upstream or downstream of the ASEs were selected as the intronic exclusion set. The 15 nt long splicing acceptor site consensus Y_{10}NCAG/G and the 9 nt long donor site consensus MAG/GURAGU [41] were not included in corresponding exonic and intronic sequences.

To identify ESEs and ESSs, we calculated the frequencies of each of 4096 possible hexanucleotides, f_{TI} and f_{TS}, in the exonic inclusion set and exclusion set of tissue T. The z-score [27, 28] of the hexamer in tissue T was then given by

$$Z_T = \frac{f_{TI} - f_{TS}}{\sqrt{(\frac{1}{N_{TI}} + \frac{1}{N_{TS}})p\ (1-p)}}$$

where N_{TI} and N_{TS} are the total number of hexamers in the inclusion and exclusion sets, respectively, and $p = (N_{TI}f_{TI} + N_{TS}f_{TS})/(N_{TI} + N_{TS})$. Tissue-specific ESEs were identified as over-represented hexamers in the exonic inclusion set of tissue T_1 but not over-represented in the exonic inclusion set of tissue T_2. To test the statistical significance of over-representation under the null hypothesis of $f_{T_1I} - f_{T_1S} = 0$, we considered hexamers with $Z_{T_1} > 2.1701$ (p-value< 0.03, two-tail test) as being over-represented. The 0.03 cutoff value for the p-value was selected based on the distribution of p-values as will be described in Discussion. To test the statistical significance of non-over-representation, we assumed the null hypothesis of over-representation as $Z_{T_2} = 2.1701$ and considered hexamers with $Z_{T_2} < -1.8808 + 2.1701 = 0.2893$ (p-value< 0.03, one-tail test) as non-over-represented hexamers.

For each pair of tissues (3 pairs in total), we compared the z-score of each hexamer as shown in Figure 2.2. Hexamers with $Z_{T_1} \geq 2.1701$ in tissue T_1 but $Z_{T_2} \leq 0.2893$ in tissue T_2 were considered as tissue T_1-specific ESEs (p-value$< 0.03^2$). Hexamers with both Z_{T_1} and $Z_{T_2} \geq 2.1701$ (p-value$< 0.03^2$) were identified as ESEs common to both tissues. The interval (0.2893, 2.1701) of a z-score corresponds to the unsure region where we do not have statistical evidence to decide whether a hexamer is over-represented or not.

Similarly, tissue-specific ESSs were identified as hexamers over-represented in the exonic exclusion set of tissue T_1, but not over-represented in the exonic exclusion set of tissue T_2. Therefore, hexamers with $Z_{T_1} \leq -2.1701$ in tissue T_1 but $Z_{T_2} \geq -0.2893$ in tissue T_2 were considered as tissue T_1-specific ESSs (p-value$< 0.03^2$). Hexamers with both Z_{T_1} and $Z_{T_2} \leq -2.1701$ (p-value$< 0.03^2$) were identified as ESSs common to both tissues.

This searching process was repeated for upstream intronic sequences of 400 nt long and downstream intronic sequences of 400 nt long. The z-score of each hexamer was computed for each pair of tissues as depicted in Figure 2.2. Tissue-specific and common ISEs and ISSs in upstream and downstream introns were then identified based on the z-scores in the same way as ESEs and ESSs were identified.

2.2.6 Integration and Clustering

After implementing the above steps, we got 6 classes of SREs which include ESE, ESS, us' ISE, us' ISS, ds' ISE and ds' ISS, where us' and ds' stand for upstream and downstream, respectively. We integrated all of them into one table (Appendix table A.1) to make their relationship more clear with the following annotation rule. Every SRE is associated with three characters to indicate its role in brain, liver and muscle. The first character can be "B", "−" or "?" to indicate that the SRE is present, absent or unsure in brain. Similarly, the second character can be "L", "−" or "?" and the third character can be "M", "−" or "?". Note that tissue specificity is a relative concept. For example, an ESE can be present in brain but not in other tissues. We annotated this type of ESE as $\mathrm{ESE}_{\mathrm{B}--}$. An ESE can also be present in both brain and liver but not in muscle. We represented this type of ESE as $\mathrm{ESE}_{\mathrm{BL}-}$. If an SRE was present in all three tissues, we referred to it as a common SRE.

We also clustered similar SREs using the hierarchical clustering algorithm [27] for each of 6 classes of SREs to determine splicing motifs. The Hamming distance was used in the clustering algorithm as the dissimilarity metric between any two SREs. We say that two SREs have incompatible annotations if any character associated with the SREs is a letter (B, L or M) in one SRE but a "−" in the other SRE. For example, two SREs, annotated as "BL-" and "BLM", respectively, are incompatible for their annotation in muscle, but two SREs annotated as "BL-" or "BL?" are compatible. Since we do not want to put those incompatible SREs into the same cluster, we add a sufficiently large value (> 6) to the dissimilarity distance between any two incompatible SREs. Therefore, each cluster only contains compatible SREs including SREs annotated with "?". The dissimilarity cutoff for each cluster was chosen to be 2.0, which was relatively small to make the clustering result more reliable.

2.2.7 Position Bias Test

The chi-square goodness of fit test was adopted to determine if an SRE is uniformly distributed in a selected region or is biased toward certain specific locations. The selected region includes introns or exons in which the SRE was predicted. For example, for SREs annotated as $\text{ESS}_{-\text{LM}}$, the exonic exclusion sets of liver and muscle were used to test position bias. Since the exons have different lengths, we only chose exons of $\geq$110 nucleotides, and took 55 nucleotides from each end of the exon. For introns, we took first 395 nucleotides upstream or downstream of the exon. An SRE which is a hexamer can be mapped to 390 positions of an intronic sequence or 100 positions of an exonic sequence. We therefore divided each exonic or intronic sequence into 10 or 39 intervals, each with 10 nucleotides. A significance level of 0.01 was used to reject the null hypothesis that an SRE uniformly appears in all intervals. Since for a uniform distribution, the χ^2 test is robust when the average number of the SREs falling into an interval is $\geq$2 for a significance level as small as 0.01 [42], only SREs with $\geq$78 counts in intronic sequences or $\geq$20 counts in exonic sequences were chosen in the analysis.

2.2.8 Comparison with Constitutive Data

We collected two sets of sequences from the KnownGene table and put them together as the data set for constitutive exons. We took 49,649 internal exons of genes with only one isoform as the first data set. The second set consists of 34,403 exons locating in alternatively spliced genes but included by all isoforms. In addition to exons, intronic sequences of 400 nucleotides upstream or downstream of the constitutive exons were also collected as the intronic constitutive data set. Note that these data sets from constitutive exons and their flanking intronic regions are similar to the inclusion set of ASEs, since all exons in the data sets are constitutively included in the mature mRNA. We compared the frequency of an SRE in the constitutive data with the frequency in its corresponding positive data. The frequency of an SRE in the negative data was also compared with that in the positive data. For example, when we examined $\mathrm{ESE}_{\mathrm{BLM}}$, the constitutive data were constitutive exons; the positive data were exonic inclusion set of brain, liver and muscle; and the negative data were the exclusion sets of these three tissues. If we examined ds' $\mathrm{ISS}_{-\mathrm{L}-}$, the constitutive data were constitutive downstream intronic sequences; the positive data were downstream intronic exclusion set of liver; and the negative data were downstream intronic inclusion set of liver. Frequency comparison was only performed for clusters annotated without "?" (369 SREs in total).

2.3 Results

2.3.1 Putative Enhancers and Silencers

As shown in Table 2.1, we got 300∼400 ASEs in the inclusion and exclusion sets of three tissues. The average length of ASEs is 123 nucleotides, and the average length of upstream and downstream introns is 5900 and 6116 nucleotides, respectively. This is consistent with the observation that ASEs are generally short and flanked by long introns [37].

Table 2.1: Number of ASEs used in SRE searching.

	brain	liver	muscle
inclusion set	399	369	411
exclusion set	372	454	408

The z-scores for hexamers in liver and muscle data sets and the regions defining

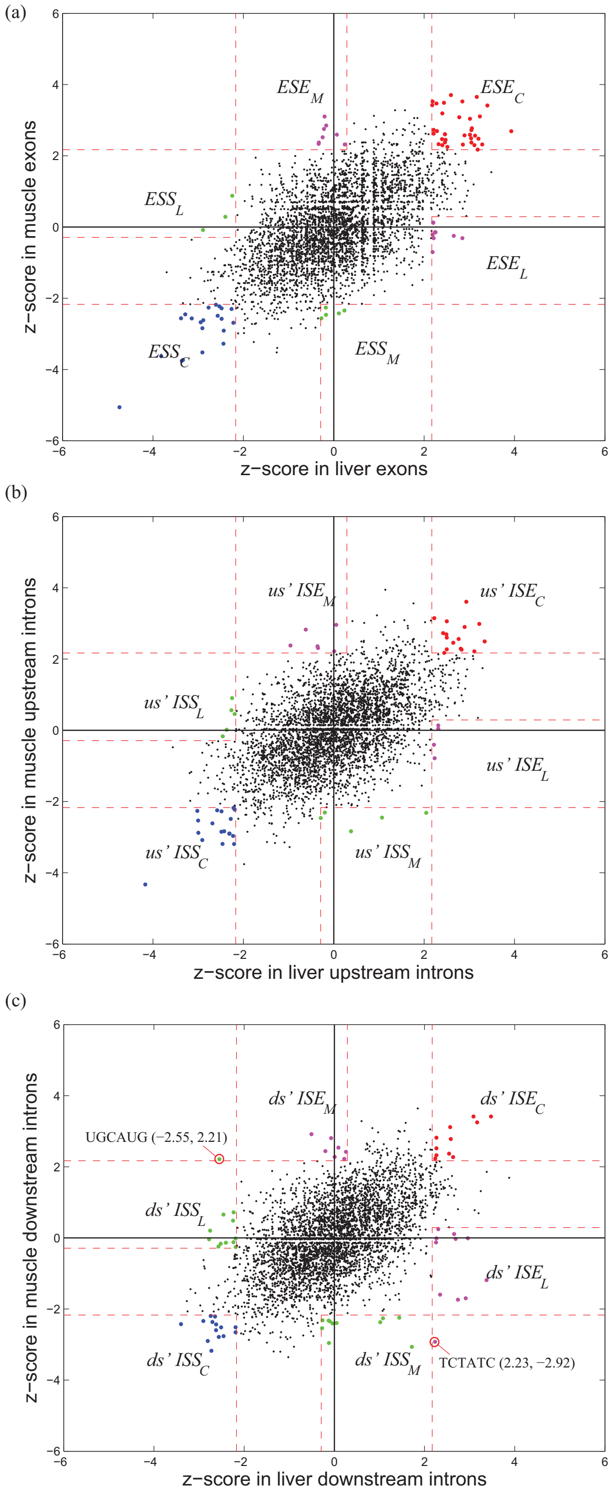


Figure 2.2: z-scores for all hexamers in liver and muscle. (a) z-scores in exons. $\mathrm{ESE_C}$, $\mathrm{ESE_L}$ and $\mathrm{ESE_M}$ stand for common ESE, liver-specific ESE and muscle-specific ESE, respectively. (b) z-scores in 400 nt intronic sequences upstream of the exons. (c) z-scores in 400 nt intronic sequences downstream of the exons.

each type of the SREs are plotted in Figure 2.2. The z-scores for hexamers in other two pairs of tissues (brain versus liver and brain versus muscle) are included in Appendix figure A.1 and A.2. It is seen from these figures that most hexamers are not over-represented in any tissue and thus are not an SRE, as expected.

After integrating all the SREs identified in these figures, we predicted 456 putative enhancers and silencers which are listed in Appendix table A.1. The statistics of these 456 SREs are summarized in Table 2.2. The second row annotated with "BLM" contains 45 SREs common to all three tissues. The next three rows consist of SREs annotated with "BL?", "B?M" and "?LM", which are common to two tissues but may or may not be an SRE in the third tissue. The next 12 rows contain a total of 221 tissue specific SREs. Note that only 18, 8 and 15 SREs are unique to brain, liver and muscle, respectively.

Table 2.2: Number of common and tissue-specific SREs.

Anno.	ESE	ESS	us' ISE	us' ISS	ds' ISE	ds' ISS	Total
B L M	15	11	6	3	4	6	45
B L ?	17	15	7	11	13	13	76
B ? M	12	14	9	9	9	11	64
? L M	21	10	10	13	6	7	67
B – ?	2	2	5	6	7	4	26
B ? –	6	1	2	8	6	4	27
– L ?	3	2	10	8	4	4	31
? L –	10	3	2	4	8	8	35
– ? M	6	4	6	5	2	4	27
? – M	6	2	4	2	5	7	26
B L –	0	0	1	0	1	0	2
B – M	0	1	0	1	1	0	3
– L M	0	0	0	2	0	1	3
B – –	8	1	4	2	3	0	18
– L –	0	0	1	1	2	4	8
– – M	2	2	2	2	2	5	15
TSSE	43	18	37	41	41	41	221

*The last row contains the total number of tissue-specific splicing elements for each type of SREs which equals to the sum of rows with annotation '–'.

To systematically examine the quality of our SREs and choose reliable candidate SREs for further analysis, we ranked all the SREs without "?" in their annotations by their final p-values (product of p-values in three tissues). The top 15 SREs and the relevant experimental evidence reported in the literature are listed in Table 2.3. Some of the 15 SREs may actually come from the same motif; for example, CCUGCC, CUGCCU and GCCUGC may come from the CCUG repeat. We examined some well studied SREs by comparing studies in the literature and their annotations in our result. Some of our annotations match previous studies very well. For example, UCUCUC and CUCUCU are both identified as us' ISS$_{-\mathrm{LM}}$ in our analysis. Previous experimental study has identified the conserved CUCUCU sequence within intron regions as splicing silencer in nonneuronal cells, since it is responsible for repressing splicing of neuron-specific N1 exon of mouse *c-src* transcript in nonneuronal cells [43], possibly by interacting with PTB proteins [44]. We will discuss several interesting SREs in the following sections.

2.3.2 Comparison with SREs identified in previous methods

We compared our results with constitutive exonic splicing enhancers RESCUE-ESE in mouse [56, 57]. Among 508 mouse RESCUE-ESEs, only 43 (8%) are included in the SREs we identified, 20 of which are also ESEs in our analysis. Note that our SREs include 108 ESEs and less than 20% (20/108) are also RESCUE-ESE. This shows that most of our SREs are different from RESCUE-ESEs possibly due to their tissue-specificity.

In addition, we compared our results with tissue-specific SREs in human identified recently by other two groups [17, 30], as shown in Figure 2.3. Using human RNA-Seq data, Wang *et al.* [17] identified 362 SREs in 15 tissues and cell lines, of which 51

Table 2.3: 15 SREs with most significant *p*-values.

SREs	Annotation	*p*-value	Ref.	Related experimental results
CCUGCC	ESS$_{\text{BLM}}$	2.43e-18	[45]	CCUG repeats specifically interact with MBNL1.
UCUAUC	ds' ISS$_{--\text{M}}$	7.53e-13	[46] [47]	Downstream ISE UCUAUC, bound by protein HRP-2, regulates alternative splicing of exon 16 in *unc-52* gene of *C. elegans*.
CUCUCU	us' ISS$_{-\text{LM}}$	1.23e-12	[43] [48] [49]	Within polypyrimidine tract, interact with PTB, responsible for the skipping of N1 exon of mouse. *c-src*.
CUGCCU	ESS$_{\text{BLM}}$	1.51e-10		Same as CCUGCC.
CUAUCU	ds' ISS$_{--\text{M}}$	1.64e-10		May be from the same motif as UCUAUC.
GCGCGC	ds' ISS$_{--\text{M}}$	2.20e-10		
GCCUGC	ESS$_{\text{BLM}}$	2.78e-10		Same as CCUGCC.
AAAUAA	ESS$_{\text{BLM}}$	3.16e-10		
UGCAUG	ds' ISE$_{--\text{M}}$	3.95e-10	[50] [51] [52]	When bound by tissue-specific factor, Fox-1 Protein family, it acts as splicing enhancer.
UGCAUG	ds' ISS$_{-\text{L}-}$	1.15e-09		
ACACAC	us' ISS$_{--\text{M}}$	1.58e-09	[53]	Intronic CA repeat could function as enhancers or silencers, depending on its proximity to the 5' ss.
UGGAGC	ESE$_{\text{BLM}}$	2.79e-09		
UUCUUC	ds' ISS$_{\text{BLM}}$	2.94e-09	[54]	It is the second pyrimidine-rich(PY) elements in the three PY elements downstream of CFTR exon 9.
AUCUAU	ds' ISE$_{\text{BL}-}$	7.24e-09		
GCAGCA	us' ISE$_{\text{BLM}}$	7.40e-09	[55]	splicing factor CUGBP1 interacts with GCA repeats located within the MEF2A mRNA.

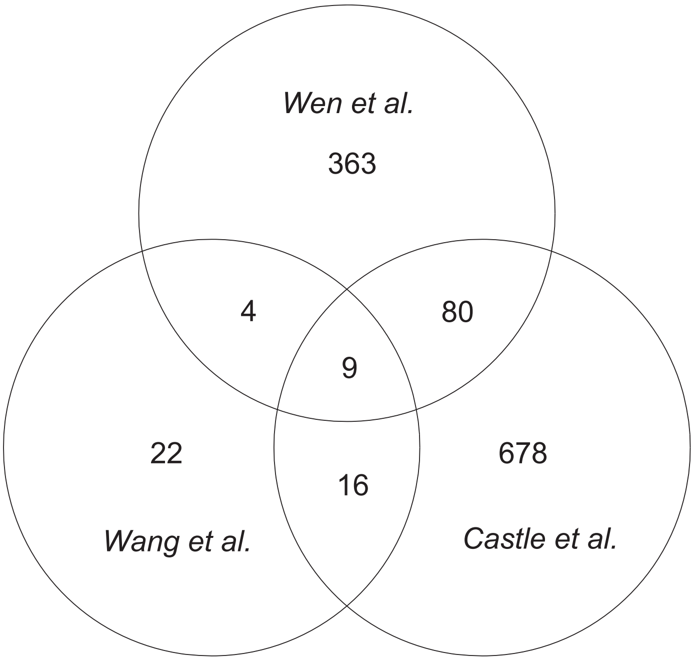


Figure 2.3: Venn diagram for the number of SREs identified in three studies.

distinct elements were identified as SREs specific to brain, liver and muscle. Among these 51 SREs, 13 (25.5%) are also included in the SREs we identified. To compare our SREs with the results of Castle *et al.* [30], we extracted all the hexamers significantly over-represented (p-value$< 10^{-3}$) in up-regulated or down-regulated cassette exons in samples related to brain, liver and muscle. This gave 783 distinct hexamers in total, of which 89 (11.4%) are also included in the SREs we identified. As shown in Figure 2.3, the number of SREs identified by any two studies is relatively small. Overall, 20% (93/456) of our SREs are in the SREs of Wang *et al.* and/or Castle *et al.*

We also compared our results with two tissue-specific motifs in mouse identified by Sugnet *et al.* [58] from their microarray data. The first CU-rich motif with consensus sequence UGYUUUC was identified by Sugnet *et al.* in upstream of brain-included exons. The most similar SREs in our results are UGAUUU (us' $\mathrm{ISE}_{?\mathrm{LM}}$) and UGAUUG (us' $\mathrm{ISE}_{\mathrm{BL}?}$). The second motif with consensus sequence UACUAAC was identified by Sugnet *et al.* in downstream intron of muscle-included exons. We can find two hexamers in our ds' ISE consistent with this motif, which are CCAAAC (ds' ISE$_{\mathrm{B?M}}$) and CGCUAA (ds' ISE$_{\mathrm{B?M}}$).

2.3.3 Comparison with experimental validated SREs

We further compared our results with 992 hexamers selected from the SpliceAid databases [38] (See Methods for the selection of 992 hexamers). About 26% (118/456) of our SREs are among these 992 hexamers (see Appendix table A.1 for details). On the other hand, 20% (10/51) of the SREs identified by Wang *et al.* [17] and 22% (173/783) of the SREs of Castle *et al.* [30] are also included in the 992 hexamers. This shows that our study gives slightly higher portion of experimentally validated SREs than other two studies.

2.3.4 Position Bias of SREs

Since splicing factors function primarily in the vicinity of a splice site [59], it is possible that positions of some SREs are biased towards certain locations, while nonfunctional sequences may tend to locate more randomly. To test if SREs have position bias, we adopted χ^2 goodness of fit test as described in Methods. Under the selection criterion described in Methods, 156 SREs were selected for position bias test (Appendix table A.2). About 46% (71/156) of SREs show significant position bias at a significance level of 0.01. We ranked these SREs according to their p-values, and visually examined the position distribution of top 30 SREs with most significant p-values. We found that 12 SREs' positions show significant position bias, eight of which were depicted in Figure 2.4. It is interesting to see that not all the SREs are biased towards a splice site. The common us' ISS AAGAUU and the common us' ISE UUGUAC occupy the position near 200 nt upstream of the acceptor site as their

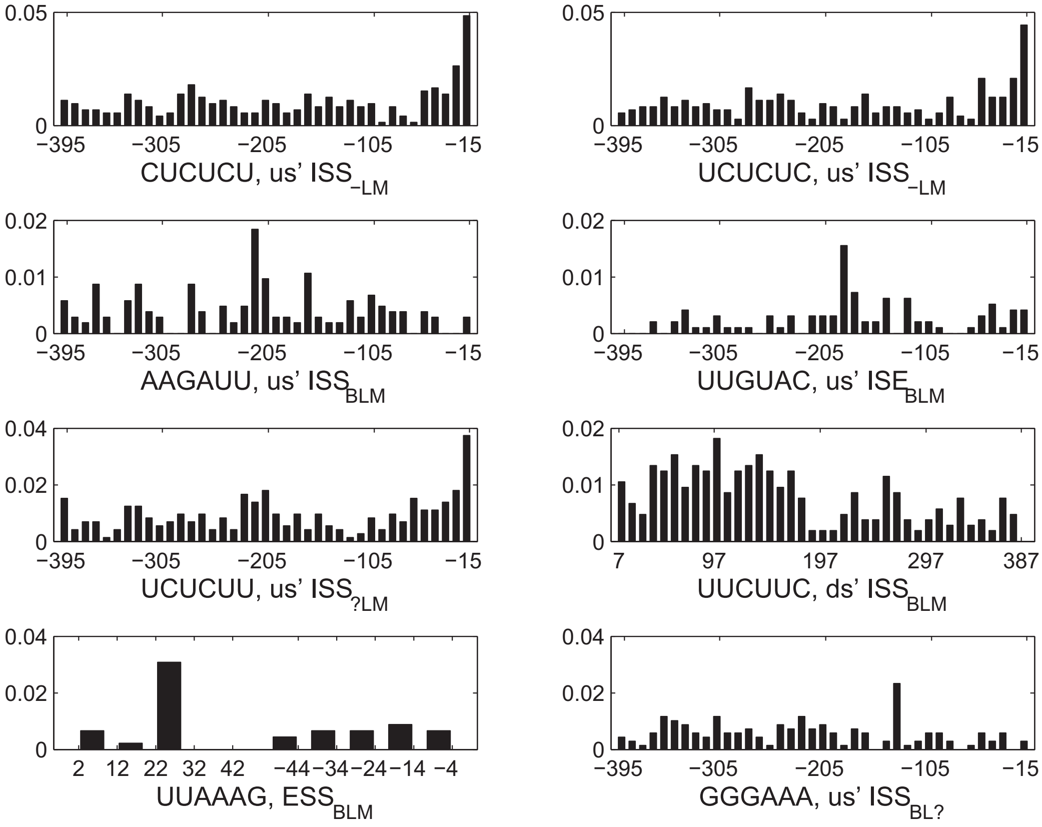


Figure 2.4: Position distribution of top 8 SREs with smallest p-values in the position bias test. Each bar represents the average number of SREs falling into a region of 10 nt divided by the number of intron or exon sequences used in analysis.

preferred location. The common ds’ ISS UUCUUC is abundant almost evenly in the region <170 nt downstream of the ASE but is less abundant in the region further away from the splice site. The common ESS UUAAAG prefers the interval between 22 and 31 nt downstream of the 5’ end of the ASE. We also checked position distributions of other SREs with p-values ≤ 0.01, but did not find any general pattern for position bias.

Two tissue-specific SREs, CUCUCU (p-value=1.33e-16) and UCUCUC (p-value =2.59e-14), which were identified as upstream ISS$_{-\mathrm{LM}}$, showed most significant p-values. They were clustered with other two SREs UCUCUU (us’ ISS$_{?\mathrm{LM}}$, p-value=1.58e-10) and CUCUUU (us’ ISS$_{?\mathrm{LM}}$, p-value=3.50E-06) in the clustering result described

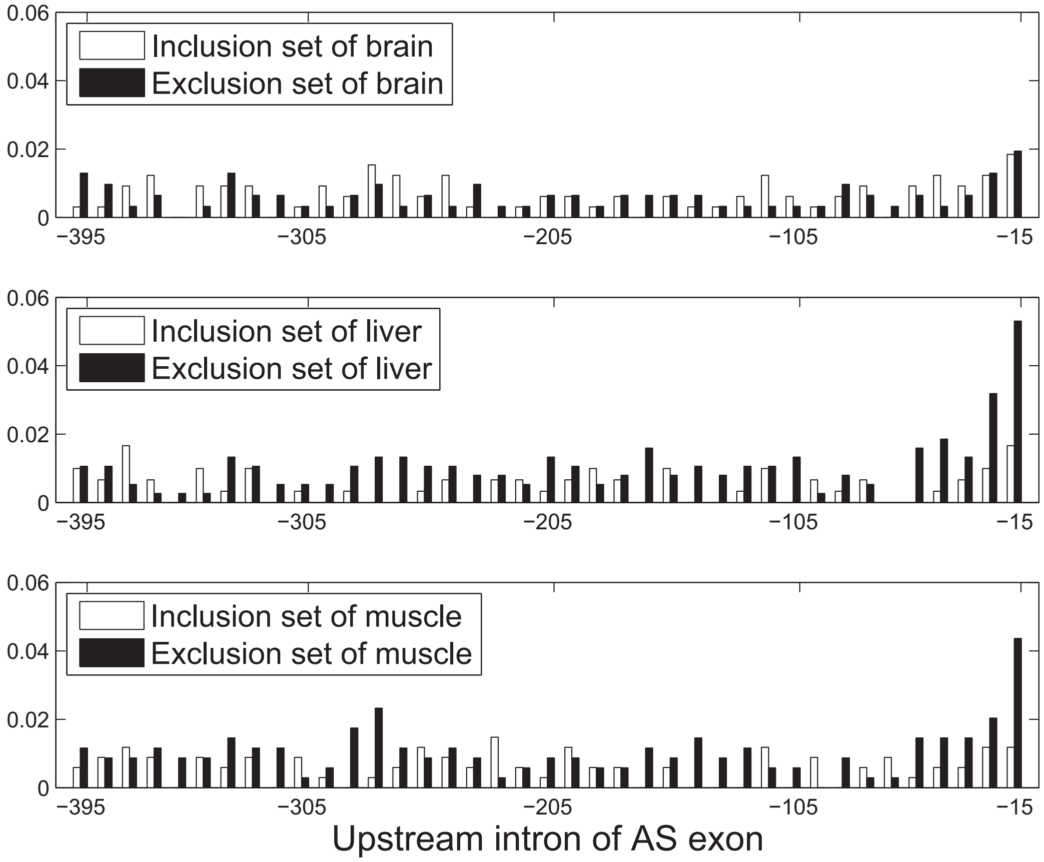


Figure 2.5: position distribution comparison for the us' $\text{ISS}_{-\text{LM}}$ CUCUCU in upstream introns of the exclusion set and the inclusion set of brain, liver and muscle. Each bar represents the average number of SREs falling into a region of 10 nt normalized by the number of intronic sequences used in analysis.

later. The position distributions of CUCUCU, UCUCUC and UCUCUU were shown in Figure 2.4. We also compared the distribution of CUCUCU in three tissues' exclusion sets with that in inclusion sets as shown in Figure 2.5. The SRE UCUCUC has a very similar position distribution that is not shown here. We can see from Figure 2.5 that in the exclusion set of liver and muscle, CUCUCU is not only abundant, but also shows a significant position bias towards the acceptor site while in the inclusion set, it is less abundant and almost evenly distributed. Since the region between 15 and 30 nucleotides upstream of a 3' splice site coincides with the location of the polypyrimidine tract, it is highly possible that CUCUCU and UCUCUC are part of the polypyrimidine tract. Note that Castle *et al.* [30] found that UCUCU is enriched in the region from 35 to 110 nt upstream of tissue-regulated ASEs in human tissues.

This result may imply the main position where this SRE takes effect, since it is consistent with the finding that polypyrimidine-tract binding protein (PTB; also known as hnRNP I) silence splicing by binding to the polypyrimidine tract and blocks the binding of U2AF [48, 49]. Interestingly, this SRE was annotated as us' $\mathrm{ISS}_{-\mathrm{LM}}$ which implies that it is specific to liver and muscle but not a us' ISS in brain. We also checked its position distribution in the brain data set, but no position bias was found as shown in Figure 2.5. Our annotation is consistent with the experimental evidence showing that skipping of neuron-specific N1 exon of mouse *c-src* in nonneuronal cells requires conserved CUCUCU elements within polypyrimidine tract and downstream intron [43].

Comparing with the results of Castle *et al.* [30] and Wang *et al.* [17], we got some identical and some different findings for the SRE CUCUCU. Both our result and the result of Castle *et al.* [30] indicate that CUCUCU is an upstream ISS in liver, but Wang *et al.* [17] did not identify it as an SRE in liver. In muscle, we identify CUCUCU as an upstream ISS, but the data of Castle *et al.* [30]indicate that it is an upstream ISE, and Wang *et al.* [17] did not identify it as an SRE. In brain, the data of Castle *et al.* [30] show that CUCUCU is over-represented in the upstream of up-regulated ASEs (which is equivalent to our upstream intronic inclusion set) but not in the upstream of down-regulated ASEs (which is equivalent to our upstream intronic exclusion set). Based on this observation, we may identify CUCUCU as an upstream ISE in brain. However, data of Castle *et al.* [30] also show that the expression of PTBP1, whose target motif is CUCUCU, is down-regulated in brain. Hence, if CUCUCU is an ISE, there must be another unknown SF that binds to it.

Another possibility is that PTBP1 is the only SF that can bind to CUCUCU, and CUCUCU is always a silencer; however, its silencing function is lost in brain due to the low level of PTBP1. The RNA-Seq data of both Mortazavi *et al.* [16] that are used in our study and Wang *et al.* [17] indicate that CUCUCU is over-represented in both upstream intronic inclusion and exclusion sets of brain. This is in conflict with the microarray data of Castle *et al.* [30]. Nevertheless, combining the RNA-Seq data of Wang *et al.* [17] and Mortazavi *et al.* [16] and the expression level of PTBP1 reported by Castle *et al.* [30], we can eliminate the possibility that CUCUCU is an ISE, but predict it to be an upstream ISS with lost silencing function in brain. Note that if we only use the information of CUCUCU without using the information about the expression level of PTBP1, the data of Castle *et al.* [30] will predict CUCUCU to be an ISE which is likely wrong; the non-discriminative method will predict CUCUCU to be both upstream ISS and ISE which conflict with each other; on the other hand, our discriminative method does not identify CUCUCU to be an SRE in brain. Our result in this complicated case seems most reasonable, because the most reasonable prediction is that CUCUCU is an ISS generally but not function in brain, as we discussed earlier. These studies indicate that CUCUCU may play an important and complicated role in tissue-specific splicing particularly in brain and worth further experimental investigation.

2.3.5 Clustering Results

Some of the 456 SREs are very similar to each other. These similar SREs may come from the same motif that is bound by the same splicing factor. Our clustering process resulted in 247 clusters as shown in Table 2.4 and Appendix table A.3. Relatively large number of clusters is due to the fact that we used a relatively small cutoff value

Table 2.4: Number of common and tissue-specific splicing motifs.

Anno.	ESE	ESS	us'ISE	us'ISS	ds'ISE	ds'ISS	Total
B L M	15(44)	13(36)	9(18)	7(14)	9(25)	10(23)	63(160)
B L ?	2(3)	3(3)	3(3)	2(3)	4(4)	3(3)	17(19)
B ? M	0(0)	3(4)	1(1)	2(2)	0(0)	1(1)	7(8)
? L M	3(4)	2(3)	0(0)	1(2)	1(1)	1(1)	8(11)
B – ?	0(0)	1(1)	2(2)	3(4)	3(4)	2(4)	11(15)
B ? –	2(2)	0(0)	0(0)	4(5)	3(3)	2(2)	11(12)
– L ?	2(2)	1(1)	4(4)	2(2)	1(1)	2(2)	12(12)
? L –	4(5)	1(2)	1(1)	0(0)	1(1)	3(4)	10(13)
– ? M	0(0)	1(1)	1(1)	1(1)	1(1)	1(2)	5(6)
? – M	1(1)	1(1)	1(3)	0(0)	2(3)	1(2)	6(10)
B L –	5(11)	1(3)	2(5)	3(7)	4(10)	2(6)	17(42)
B – M	4(8)	2(4)	3(7)	4(9)	2(3)	3(8)	18(39)
– L M	4(10)	1(2)	5(13)	8(20)	0(0)	2(4)	20(49)
B – –	7(12)	1(1)	4(5)	2(4)	3(7)	0(0)	17(29)
– L –	0(0)	0(0)	1(2)	1(2)	3(7)	3(8)	8(19)
– – M	3(6)	3(6)	3(5)	1(2)	3(4)	4(8)	17(31)
TSSM	32(57)	13(22)	27(48)	29(56)	26(44)	25(50)	152(277)

*The last row contains the total number of each type of tissue-specific splicing motifs (TSSM). The number of hexamers in each type of motifs is shown in parenthesis.

(2.0) for the dissimilarity distance between any two SREs in the same cluster.

After the clustering process, we re-annotated each cluster to eliminate some "?" annotation. For example, one cluster consists of sequences with annotation "BL?" and "BL-", then we re-annotate all the elements in the cluster as "BL-". The high ratio between the number of tissue-specific SREs and common SREs (221/45) observed in Table 2.2 was decreased to 152/63 in Table 2.4, but the ratio is still significantly large, implying that tissue-specific motifs may play a very important role in splicing regulation. The average number of SREs per cluster is 160/63 = 2.54 for common SREs and 277/152 = 1.82 for tissue-specific SREs, which implies that tissue-specific motifs may be more conservative than common motifs.

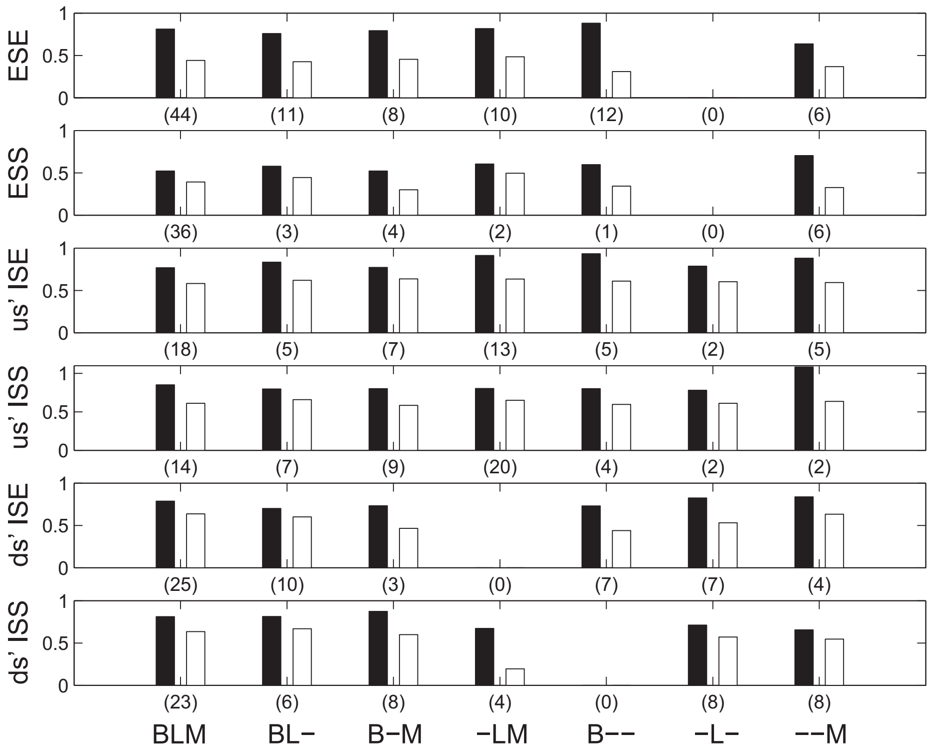


Figure 2.6: Comparison of frequencies of different SREs in different data sets. The first bar in each group stands for the ratio of the frequency in constitutive data to the frequency in the positive data. The second bar stands for the ratio of frequency in the negative data to the frequency in the positive data. Number of SREs used in each comparison is shown in parenthesis.

2.3.6 Frequencies of Identified SREs in Constitutive Exons

The 456 SREs were identified based on their frequencies in the inclusion and exclusion sets of the ASEs. We also wished to know the frequencies of these SREs in the constitutively spliced exons and their flanking intronic regions to gain more insight of the role of these SREs. Using the constitutive data described in Materials and Methods, we compared the frequency of SREs in different data sets, as depicted in Figure 2.6. For the clarity of comparison, frequencies in the constitutive data and the negative data have been normalized by the frequencies in the positive data.

First, let us look at the frequencies of the enhancers. We would expect the constitutive exon data set to have abundant enhancers to assist splicing. However, it is seen from Figure 2.6 that enhancers we identified have lower frequencies in the constitutive exon data set than in the inclusion set of ASEs. This may be due to the following two reasons. First, most of the tissue-specific enhancers may be different from the enhancers present in the constitutively spliced exons and flanking introns. This may also explain why most of enhancers we identified are not RESCUE-ESEs. Second, tissue-specific enhancers are more abundant in ASEs than in constitutively spliced exons. We also compared the frequencies of constitutive RESCUE-ESEs in constitutive exons and our inclusion sets of brain, liver and muscle, but no frequency difference was found.

The frequencies of silencers are expected to be lower in the constitutive data set than in the exclusion set of the ASEs. Indeed, this is observed in Figure 2.6. Comparing the relative frequencies of ESS with frequencies of other silencers and enhancers, we see that the relative frequencies of ESS are generally the lowest. This may imply that ESSs play a stronger role in AS than ISSs and ISEs.

Another observation from Figure 2.6 is that the frequencies of all SREs in the constitutive data set are higher than the frequency in the corresponding negative data set (exclusion set for enhancers and inclusion set for silencers) of ASEs. Therefore, if we use the constitutive data as the negative control data as did in [27, 29, 32, 60] to identify SREs, we would lose some detection power. To verify this, we calculated the z-score and the corresponding p-value of each of the 369 SREs by replacing the negative data set used in the previous analysis with the constitutive data set. We found that 179 SREs have a p-value ≥ 0.03, implying that 48.5% (179/369) of these SREs would not be identified if we have used constitutive data set as the negative data set. Among these 179 SREs, 34% (60/179) SREs could be found in the 992 hexamers selected from SpliceAid; whereas among the remaining 190 SREs, only 22% (42/190) can be found in the 992 hexamers. This indicates that more percentage of true positive SREs can be lost if the non-discriminative approach is employed.

2.3.7 Special SREs That Can Be Both Enhancer and Silencer

Among 456 SREs we identified, two SREs are special because they were identified as an enhancer in one tissue but a silencer in another tissue. These two SREs are UGCAUG and UCUAUC, whose z-score are shown in Figure 2.2(c).

UGCAUG was annotated as a downstream ISE$_{--\mathrm{M}}$ and downstream ISS$_{-\mathrm{L}-}$. Our annotation ISE$_{--\mathrm{M}}$ (muscle-specific ISE) of UGCAUG is consistent with the computational result [32] and experimental observation [50, 51], as well as with the results of Wang *et al.* [17] and Castle *et al.* [30]. Our annotation ds' ISS$_{-\mathrm{L}-}$ is also consistent with the result of Castle *et al.* [30], but Wang *et al.* [17] did not predict UGCAUG to be an SRE in liver. To the best of our knowledge, this putative role of downstream ISS in liver has not been reported in any experimental results, although it was experimentally verified to be an upstream ISS [61]. Further experimental investigations worth being carried out to see if it is a liver-specific ISS as Castle *et al.* and we predicted. If this is true, new splicing factors binding to this hexamer may be identified, given the fact that Fox-1 is not expressed in liver [61].

We did not identify UGCAUG to be an upstream ISE in brain as previous computational work did [32, 50]. The data of Wang *et al.* [17] also indicate that UGCAUG is an ISE in brain, but enrichment in brain is not so significant as in heart and muscle. The data of Castle *et al.* [30] are more complicated, because UGCAUG is over-represented in the downstream intronic region of both up- and down-regulated ASEs in several brain cells including medulla oblongata, thalamus and in fetal brain, but is over-represented in the downstream intronic region of only up-regulated ASEs in other brain cells including cerebellum and hippocampus. Hence, if we use the non-discriminative method, we would predict UGCAUG to be both ISE and ISS in the same downstream region and in the same type of cell, which is obviously a conflictive result. To find out why our method using the data of Mortazavi *et al.* [16] did not predict UGCAUG to be an ISE in brain, we rechecked the data and found that UGCAUG is moderately abundant in both the inclusion and exclusion sets of brain. The z-score of our discriminate approach is lower than the critical value at the 0.03 significance level. Generally, for those sequences that are abundant in both inclusion and exclusion sets, our discriminative approach will not predict them to be an SRE, but the non-discriminative will give a conflictive prediction: such sequences are both an enhancer and a silencer. Given the different results in different studies and the fact that UGCAUG is a binding target of Fox-1 protein family specifically expressed in brain [61, 62], more carefully designed experiment is needed to investigate the role of UGCAUG in brain, especially in different brain cell types.

2.4 Discussion

Reads from RNA-Seq give information about how exons are connected which can be explored in the investigation of AS. RNA-Seq also provides more accurate measurement of expression levels of transcripts and their isoforms across a very broad dynamic range than other methods such as microarray [14]. Capitalizing on these two advantages of RNA-Seq, we identified ASEs from the mouse RNA-Seq data set [16] and calculated the expression levels of isoforms of the genes containing the selected ASEs. This enabled us to determine reliable positive and negative data sets for SREs and then to employ a powerful discriminative approach to identify enhancers and silencers regulating alternative splicing. We chose the RNA-Seq data for three mouse tissues [16] rather than more comprehensive RNA-Seq data for 15 human tissues and cell lines [17] due to the following two reasons. First, unlike the human RNA-Seq data [17], the mouse RNA-Seq data [16] have not been explored to predict any SREs. Second, as demonstrated in [16], the RNA-Seq reads generated from the protocol using RNA fragmentation provide more uniform coverage along the transcripts than those generated from the protocol using cDNA fragmentation [17], and thus, the mouse RNA-Seq data can be used to calculate the expression level of each isoform of each gene more accurately.

As shown in [34–36], a discriminative approach using reliable positive and negative data can significantly increase the power of detecting motifs that are over-represented in the positive data set relative to the negative data, without increasing the false positive rate. However, most computational methods for identifying SREs do not employ the discriminative approach. These include the ones used to identify RESCUE-ESEs from constitutively spliced exons [27] and tissue-specific SREs from microarray data [30] as we discussed in Introduction. Similar to the method used to identify RESCUE-ESEs, intronic sequences flanking constitutively spliced exons were used as background data to identify brain-specific SREs [32]. The putative ESEs and ESSs (PESEs/PESSs) were identified by comparing the frequencies of octamers in constitutively spliced non-protein-coding exons with those in a negative control set including the pseudo exons and 5' untranslated regions of intronless gene [28]. Although this negative set may be more reliable than the one used in identifying RESCUE-ESEs, it may not be as reliable as the negative data in our method due to the following arguments. Pseudo exons are good negative sequences for identifying ESEs because they are never spliced. However, although the ASEs in our exclusion set are also not spliced in a tissue or under certain condition, they are spliced in other tissue(s) or under other conditions. This is a stronger indication that these ASEs in our exclusion set may lack the ESEs that assist the splicing of ASEs in the positive data. Similar arguments hold for other enhancers or silencers. In the identification of ESS from pseudo exons [29], constitutively spliced exons and their flanking intronic regions were used as the negative data set, which is again not as reliable as the ASEs and their flanking intronic regions in our inclusion set because these ASEs can also be skipped under different conditions.

Another advantage of our discriminative approach is that it can identify both common and tissue-specific SREs. This is an important feature because both tissue-specific splicing factors and tissue-specific expression of constitutive splicing factors may play a role in regulating alternative splicing. If we use constitutively spliced exons as the negative data as used in [27,29,32,60], we would not only lose detection power as shown in the Results, but also miss those common SREs present in constitutively splice exons. As a side note, similar to the method used to identify PESE/PESS [28], our method do not have problem of sequence bias such as codon or CpG bias, since our positive and negative data sets have similar sequence composition. If a sequence is abundant in both inclusion and exclusion sets, our discriminative approach generally will not predict it as an SRE, but the non-discriminative approach will likely predict it to be both an enhancer and a silencer, which obviously is a conflictive and confusing result. On the other hand, if an SRE is abundant in both the data set from which we try to identify the SRE and the background data set, non-discriminative approach can not identify such an SRE, but our discriminative approach using negative data set is very likely able to identify it.

In order to reduce the false positive rate without losing detection power, we used

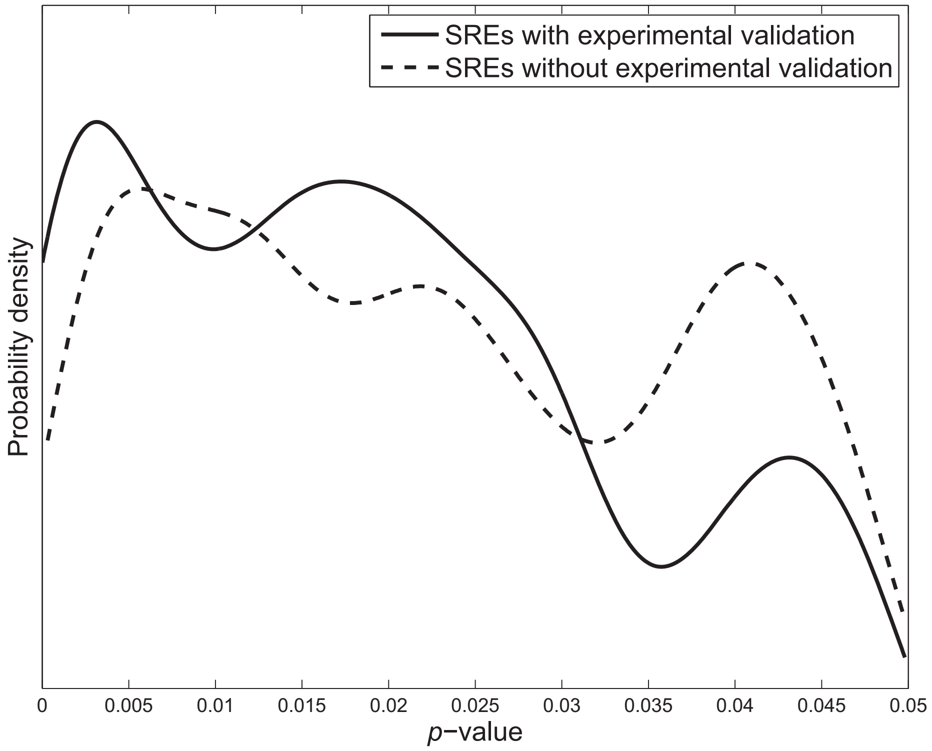


Figure 2.7: Probability density of p-values of the SREs with or without experimental validation. SREs are computationally identified at a significance level of 0.05.

a validating process to determine the cutoff p-value which was chosen to be 0.03. Specifically, we first used a cutoff p-value equal to 0.05. This gave 799 SREs, 200 of which could be found at least one match in the 992 hexamers selected from SpliceAid [38] containing experimentally identified SREs. We plotted the distribution of the p-values of these 200 SREs and of the remaining 599 SREs, as shown in Figure 2.7. It is seen that at a p-value$<$ 0.03, the probability of experimentally validated SREs is generally higher than the probability of SREs without experimental validation, and that this trend is reversed at p-value$>$ 0.03. Therefore, we selected 0.03 to be the cutoff p-value.

About 26% (118/456) of 456 SREs we identified can be found in database with experimentally validated SREs. This percentage is slightly higher than that for the SREs identified by Wang *et al.* [17] and Castle *et al.* [30] from human tissues. About

48% (221/456) of our SREs are tissue-specific, which shows that tissue-specific SREs play an important role in regulating alternative splicing as observed early. Although only 10% (45/456) SREs are common to all three tissues in this study, it does not imply that common SREs are less important, because 45% (207/456) SREs were common to two tissues but unsure to the other tissue. If more data are available, we may identity these SREs as common or tissue-specific SREs. Only 18% (20/108) of our ESEs are included in RESCUE-ESE identified from constitutively splice exons, and only 14% (15/108) of our ESEs are annotated as common to three tissues. This shows that much more tissue-specific ESEs are involved in regulating tissue-specific splicing than constitutive ESEs.

It worths some discussions on three SREs: CUCUCU (us' $\mathrm{ISS}_{-\mathrm{LM}}$), UGCAUG (ds' $\mathrm{ISE}_{--\mathrm{M}}$ and ds' $\mathrm{ISS}_{-\mathrm{L}-}$) and UCUAUC (ds' $\mathrm{ISS}_{--\mathrm{M}}$ and ds' $\mathrm{ISE}_{?\mathrm{L}-}$). The first two have been repeatedly identified as an SRE in both experimental and computational approaches [30, 43, 48–52], but our study reveal some new information. Specifically, our position analysis showed that CUCUCU appears at 15 nt to 30 nt upstream of the ASE skipped in liver and muscle but not brain with much higher frequency than any other locations. Since these locations are in the polypyrimidine tract, CUCUCU most likely functions in the polypyrimidine tract as a tissue-specific silencer. While previous results showed that an SRE can be an enhancer or silencer depending on its location. For example, UGCAUG can be a ds' ISE or a us' ISS. Our analysis showed that UGCAUG and UCUAUC can function as an enhancer in one tissue but a silencer in another tissue from the same intronic region downstream of the ASE, which calls further investigation about the mechanism that these two SREs function.

2.5 Concluding Remarks

Tissue-specific alternative splicing is a key mechanism for generating tissue-specific proteomic diversity in eukaryotes. Splicing regulatory elements (SREs) in pre-mature messenger RNA play a very important role in regulating alternative splicing. In this chapter, we use mouse RNA-Seq data to determine a positive data set where SREs are over-represented and a reliable negative data set where the same SREs are most likely under-represented for a specific tissue and then employ a powerful discriminative approach to identify SREs. We identified 456 putative splicing enhancers or silencers, of which 221 were predicted to be tissue-specific. Most of our tissue-specific SREs are likely different from constitutive SREs, since only 18% of our exonic splicing enhancers (ESEs) are contained in constitutive RESCUE-ESEs. A relatively small portion (20%) of our SREs is included in tissue-specific SREs in human identified in two recent studies. In the analysis of position distribution of SREs, we found that a dozen of SREs were biased to a specific region. These findings provide insight into the mechanism of tissue-specific alternative splicing and give a set of valuable putative SREs for further experimental investigations.

CHAPTER 3

A Thermodynamics Model for Identifying Splicing Regulatory Elements and Their Interactions

3.1 Motivation

Current computational methods for the detection of SREs can be largely categorized into three approaches. The first enrichment-based approach is to identify SREs as short nucleotide sequences (typically hexamers or octamers) that are statistically enriched in a carefully selected set of introns and exons against a background or negative dataset. A large part of the current methods developed in [8,27,28,30] and our method presented in Chapter 2 belong to this category. The second conservation-based approach utilizes comparative genomic methods to identify evolutionarily conserved motifs in introns and exons, which can also be combined with the enrichment-based approach to identify SREs [46,63,64]. The third regression-based approach exploits both sequence information and expression levels of different isoforms in a unified framework [33,65]. Comparing with the other two approaches, the regression-based approach offers flexibility of identifying combinatorial regulatory effects of multiple SREs. However, the current regression methods were not developed systematically from a theoretical base, which may limit their performance.

Multiple SFs could act cooperatively to promote or repress splicing by regulating exon or intron definition [25]. These interactions will cause their binding SREs have specific features that can be captured by computational approaches. Several recent computational works have studied cooperative SRE pairs in AS regulation. Ke *et al.* [66] searched for frequently co-occurred SRE pairs from two ends of exons that mediate exon definition. Friedman *et al.* [67] identified cooperative SRE pairs from two ends of human and mouse introns possibly mediating intron definition. Suyama *et al.* [68] analyzed conserved pentamers that often co-occur in the same region of upstream or downstream introns, which may arise from cooperative binding of different SFs or actually from a single long motif. These different types of SRE pairs from different regions reveal that cooperative interaction between SREs may be a common mechanism in AS regulation. However, these SRE pair detection methods did not incorporate expression data into the analysis, and like the enrichment-based approach to the detection of single SRE, they could not exploit sequence and expression data in a systematic way, which may limit their detection power.

In this chapter, we employ the principles of thermodynamics to overcome the shortcomings of enrichment-based approach and other existing methods by introducing a rigorous linear regression model for AS regulation. The regression-based model has been used to identify transcription factor binding site (TFBS) for a long time [69–72]. However, this model is an approximation of a nonlinear thermodynamics-based model as shown in [71, 73–75], and applying it directly to TFBS interaction detection [76] is skeptical. Specifically, we first derive a novel thermodynamics-based regression model for AS regulation of alternatively spliced exons (ASEs) which can capture both main effects of individual SREs and combinatorial effects of multiple SREs. We then develop a systematic framework to infer the regression model, which in turn identifies both single SREs and different types of cooperative SRE pairs. The key feature of our model inference framework is that we employ the shrinkage technique [77] to identify a small number of SREs and SRE pairs from a huge number of all possible SREs and their pairs. Our numerical results show that our model can explain a significant portion of the variance in the data comparable to the best result for transcription achieved by a non-linear thermodynamic model [73]. Using an RNA-Seq data set [17], we identify 619 SREs and 196 SRE pairs, some of which are verified with previous experimental results.

The remaining part of this chapter is organized as follows: In Section 3.2 we derive the thermodynamic model for splicing regulation. In Section 3.3, we discuss the methods used to determine the regulatory effects. In Section 3.4, we develop the framework for model inference, and apply the model and inference framework to a human RNA-Seq data set. In Section 3.5, the performance of our model is evaluated and the identified SREs and SRE pairs are presented. In Section 3.6, we discuss the merits of our thermodynamics-based model by comparing it with previous models and other SRE pair detection algorithms. Finally, conclusions are drawn in Section 3.7.

3.2 The Thermodynamic Model for AS Regulation

3.2.1 The Model of Spliceosome Assembly

Consider a gene containing an ASE that can generate two isoforms I_1 and I_2, either with or without the ASE. Since splicing is coupled with transcription and the product emerging from this coupled process is either I_1 or I_2, we can consider the splicing of each pre-mRNA independently. Splicing begins with a multi-step process of spliceosome assembly around the splice sites and the branch point. We model the assembly of spliceosome S to the splice sites of ASE on each individual pre-mRNA as a single chemical reaction:

$$\begin{array}{ccc} S + RNA & \rightleftharpoons & RNA{\cdot}S, \\ \downarrow & & \downarrow \\ I_2 & & I_1 \end{array} \tag{3.1}$$

where RNA denotes the state in which S is not fully assembled, and $RNA{\cdot}S$ represents the state in which S is fully assembled around the ASE. This simplification is similar to the one used in the derivation of a thermodynamics-based model [73–75] for transcription where assembly of the RNA polymerase (RNAP) complex is simplified to one reaction. We assume that I_1 is produced from the $RNA{\cdot}S$ state and I_2 is produced from the RNA state. The probability of having a spliceosome assembled around the ASE is equal to the probability of the $RNA{\cdot}S$ state in reaction (3.1), which can be expressed as $P_s = \frac{[RNA{\cdot}S]}{[RNA]+[RNA{\cdot}S]}$, where [RNA·S], [RNA] and [S] stand for the concentrations of RNA·S, RNA and S, respectively. Since the equilibrium constant of reaction (3.1) is given by $k_s = \frac{[RNA{\cdot}S]}{[RNA][S]}$, we can write P_s as follows:

$$P_s = \frac{[RNA{\cdot}S]}{[RNA]+[RNA{\cdot}S]} = \frac{k_s[S]}{1+k_s[S]} = \frac{q_s}{1+q_s}, \tag{3.2}$$

where $q_s = k_s[S]$. Similar to the gene expression model [71, 76], the dynamic changes of the concentration of I_1 denoted as E_{I_1} can be written as:

$$\frac{dE_{I_1}}{dt} = k_g P_s - k_d E_{I_1}, \tag{3.3}$$

where k_g and k_d are constants related to synthesis and degradation rates, respectively. In the steady state where $dE_{I_1}/dt = 0$, we have $E_{I_1} = \frac{k_g}{k_d} P_s = \alpha P_s$, where

$\alpha = k_g/k_d$. Likewise, concentration of the second isoform $E_{I_2} = \gamma(1 - P_s)$, where γ is another constant. Thus, the ratio of E_{I_1} and E_{I_2} can be written as:

$$\frac{E_{I_1}}{E_{I_2}} = \frac{\alpha}{\gamma}\frac{P_s}{1 - P_s} = c \cdot q_s, \tag{3.4}$$

where constant $c = \alpha/\gamma$. Therefore, the probability of producing I_1 is equal to the probability that the pre-mRNA is bound by the spliceosome.

3.2.2 The Regulatory Model of One SF

Now we consider an SF that can bind to an SRE around the ASE and influence the assembly of the spliceosome. The pre-mRNA can have four possible states: 1) bound by both S and SF ($RNA{\cdot}S{\cdot}SF$), 2) bound by S only ($RNA{\cdot}S$), 3) bound by SF only ($RNA{\cdot}SF$), and 4) bound by neither of them (RNA). Then the probability P_s of having a spliceosome assembled around the ASE is equal to the probability of states 1) and 2). Following [73–75], we can write P_s as $P_s = (z_1+z_2)/(z_1+z_2+z_3+z_4)$, where z_i is the Boltzmann weight for state i. Let w be the cooperative factor reflecting the interaction between SF and S, $q_{sf} = k_{sf}[SF]$ where $k_{sf} = \frac{[RNA{\cdot}SF]}{[RNA][SF]}$, then it is not difficult to find that $z_1 = wq_sq_{sf}$, $z_2 = q_s$, $z_3 = q_{sf}$ and $z_4 = 1$, which yields:

$$P_s \quad = \quad \frac{q_s + wq_sq_{sf}}{1 + q_{sf} + q_s + wq_sq_{sf}}. \tag{3.5}$$

If $w = 1$, SF and S bind to the transcript independently and P_s in (3.5) is simplified to that in (3.2). If $w > 1$, the binding of SF to SRE increases the probability of spliceosome assembly, which implies that the SF is an enhancer. If $w < 1$, binding of the SF has a negative effect on spliceosome assembly and the SF is a repressor. The ratio of the expression levels of I_1 and I_2 can be written as:

$$\frac{E_{I_1}}{E_{I_2}} = \frac{\alpha}{\gamma}\frac{P_s}{1-P_s} = c \cdot q_s \frac{1+wq_{sf}}{1+q_{sf}}. \tag{3.6}$$

3.2.3 The Interacting Model of Two SFs

Multiple SFs and the spliceosome can also interact cooperatively or antagonistically to affect the splicing process. If two SFs can cooperatively bind to their SREs around the ASE and interact with the spliceosome, it is not difficult to derive the following ratio:

$$\frac{E_{I_1}}{E_{I_2}} = c \cdot q_s \frac{1+w_1q_{sf_1}+w_2q_{sf_2}+w_{12}w_1w_2q_{sf_1}q_{sf_2}}{1+q_{sf_1}+q_{sf_2}+w_{12}q_{sf_1}q_{sf_2}}, \tag{3.7}$$

where $q_{sf_1} = k_{sf_1}[SF_1]$ with $k_{sf_1} = \frac{[RNA \cdot SF_1]}{[RNA][SF_1]}$, q_{sf_2} and k_{sf_2} are defined similarly for SF_2, $w_i, i = 1, 2$, is the cooperativity factor between SF_i and S, and w_{12} is the cooperativity factor between two SFs. If $w_{12} = 1$, there is no cooperative interaction between two SFs, and they enhance or repress spliceosome assembly independently. In this case, (3.7) can be simplified as

$$\frac{E_{I_1}}{E_{I_2}} = c \cdot q_s \frac{1+w_1q_{sf_1}}{1+q_{sf_1}}\frac{1+w_2q_{sf_2}}{1+q_{sf_2}}. \tag{3.8}$$

If $w_{12} \neq 1$, we can express (3.7) as:

$$\frac{E_{I_1}}{E_{I_2}} = c \cdot q_s \frac{1+w_1q_{sf_1}}{1+q_{sf_1}}\frac{1+w_2q_{sf_2}}{1+q_{sf_2}}\ \phi, \tag{3.9}$$

where $\phi = (\frac{1+w_1q_{sf_1}+w_2q_{sf_2}+w_{12}w_1w_2q_{sf_1}q_{sf_2}}{1+q_{sf_1}+q_{sf_2}+w_{12}q_{sf_1}q_{sf_2}})/(\frac{1+w_1q_{sf_1}}{1+q_{sf_1}}\frac{1+w_2q_{sf_2}}{1+q_{sf_2}})$. If we define $b_1 = \log(\frac{1+w_1q_{sf_1}}{1+q_{sf_1}})$, $b_2 = \log(\frac{1+w_2q_{sf_2}}{1+q_{sf_2}})$ and $b_{12} = \log(\phi)$, we can write (3.9) as:

$$\log(\frac{E_{I_1}}{E_{I_2}}) = \log(c \cdot q_s) + b_1 + b_2 + b_{12}. \tag{3.10}$$

The first term reflects the basal level of splicing determined by the spliceosome alone, the second and third terms are the effects of interactions between the spliceosome and each individual SF, while the last term is the effect of the interaction between two SFs. This can be seen from the fact that $b_i = 0$ if $w_i = 1$ and $b_{12} = 0$ if $w_{12} = 1$. In other words, if the ith SRE affects splicing, then $b_i \neq 0$; otherwise $b_i = 0$; Similarly, if two SREs interact with each other and affect splicing jointly, then $b_{12} \neq 0$; otherwise $b_{12} = 0$.

3.2.4 Removal of Exon-Specific Effects

Note that $b_i, i = 1, 2$, is determined by k_{sf_i}, $[SF_i]$ and w_i. Since an SF can bind to the same set of SREs around different ASEs, we assume that k_{sf_i} and w_i are the same for different ASEs. Therefore, if we consider a set of ASEs in the same tissue or under the same condition, where $[SF_i]$ is fixed, b_i is identical for these ASEs. Similarly, we assume that b_{12} is a constant for different ASEs in the same tissue. On the other hand, since different exons may have strong or weak splice sites and different genes may have different degradation rates, the first term $\log(c{\cdot}q_s)$ in (3.10) may be different for different exons even in the same tissue or under the same condition. Since our goal is to infer b_1, b_2 and b_{12} from data of multiple ASEs in the same tissue, we need to remove the exon-specific effects from the model.

If expression levels of isoforms in two tissues t_1 and t_2 are available, the first term in (3.10) is identical in these two tissues, and can be removed by forming the following model:

$$\log(\frac{E_{I_1}^{t_1}}{E_{I_2}^{t_1}}) - \log(\frac{E_{I_1}^{t_2}}{E_{I_2}^{t_2}}) = (b_1^{t_1} - b_1^{t_2}) + (b_2^{t_1} - b_2^{t_2}) + (b_{12}^{t_1} - b_{12}^{t_2}), \tag{3.11}$$

The data of tissue t_2 can be regarded as a reference. Subtraction of the reference data from the data of tissue t_1 removes the exon-specific effects. When data of multiple tissues are available, we can arbitrarily choose a tissue as the reference. However, since the expression level of each isoform is estimated from the noisy measurements, a better reference can be obtained by averaging the data of multiple tissues, which is similar to the strategy used in [65]. Specifically, suppose we have a set of data $E_{I_1}^t, E_{I_2}^t$ for tissue $t = 1, ..., T$, we can remove the first term in (3.10) by forming the following model:

$$\begin{aligned} &\log(\frac{E_{I_1}^{t_1}}{E_{I_2}^{t_1}}) - \frac{1}{T-1}\sum_{\substack{t=1\\t\neq t_1}}^{T}\log(\frac{E_{I_1}^{t}}{E_{I_2}^{t}}) \\ &= (b_1^{t_1} - \frac{1}{T-1}\sum_{\substack{t=1\\t\neq t_1}}^{T} b_1^t) + (b_2^{t_1} - \frac{1}{T-1}\sum_{\substack{t=1\\t\neq t_1}}^{T} b_2^t) + (b_{12}^{t_1} - \frac{1}{T-1}\sum_{\substack{t=1\\t\neq t_1}}^{T} b_{12}^t). \end{aligned} \tag{3.12}$$

If we define $y = \log(\frac{E_{I_1}^{t_1}}{E_{I_2}^{t_1}}) - \frac{1}{T-1}\sum_{\substack{t=1\\t\neq t_1}}^{T}\log(\frac{E_{I_1}^{t}}{E_{I_2}^{t}})$, $\beta_1 = b_1^{t_1} - \frac{1}{T-1}\sum_{\substack{t=1\\t\neq t_1}}^{T} b_1^t$, $\beta_2 = b_2^{t_1} - \frac{1}{T-1}\sum_{\substack{t=1\\t\neq t_1}}^{T} b_2^t$, and $\beta_{12} = b_{12}^{t_1} - \frac{1}{T-1}\sum_{\substack{t=1\\t\neq t_1}}^{T} b_{12}^t$, then (3.12) can be simplified as:

$$y = \beta_1 + \beta_2 + \beta_{12}. \tag{3.13}$$

3.2.5 The Final Model Used for Inference

So far, we have assumed that y can be measured without any error. If the measurement error is taken into account, equation (3.13) becomes:

$$y = \beta_1 + \beta_2 + \beta_{12} + \epsilon, \tag{3.14}$$

where ϵ is the measurement error modeled as a Gaussian random variable with zero mean. Model (3.14) is derived under the assumption that two SREs are present to regulate splicing. Since we do not know which SRE or SRE pair contributes to splicing, we can include all potential SREs and their pairwise interaction in the model by adding a parameter to the model for each potential SRE and SRE pairs. Moreover, when we use this model to identify SREs and their interactions, we need to apply it to a set of ASEs. However, different ASEs may have different SREs. To overcome this problem, we include all possible SREs (typically hexamers) and their pairwise interactions in the model, but multiply β_i by a binary variable $x_i \in \{0, 1\}$ that indicates if the corresponding SRE is present in the ith ASE, and similarly multiply β_{ij} by $x_i x_j$. This gives rise to the following model:

$$y = \beta_0 + \sum_{i \in \mathcal{M}} x_i \beta_i + \sum_{(i,j) \in \mathcal{I}} x_i x_j \beta_{ij} + \epsilon, \tag{3.15}$$

where $y = \log(f^t) - \frac{1}{N-1} \sum_{t'=1, t' \neq t}^{N} \log(f^{t'})$, $f^t = E^t_{I_1} / E^t_{I_2}$ in tissue t, and N is the total number of tissues; $\mathcal{M}$ and $\mathcal{I}$ are the set of potential SREs and the set of potential SRE pairs, respectively; x_i is a binary variable to indicate presence ($x_i = 1$) or absence ($x_i = 0$) of the ith SRE; β_i reflects the contribution of the ith SRE to the splicing of the ASE, and β_{ij} indicates the cooperative contribution of the ith SRE and the jth SRE; ϵ is the measurement error, which is modeled as a Gaussian random variable with zero mean. Thus, given the splicing profile y of a set of ASEs and a set of candidate SREs, we can identify SREs or SRE pairs that have regulatory effect by finding β_i or β_{ij} that are not equal to zero with certain statistical significance.

We can also include interactions involving more than two SREs, but this will dramatically increase the number of unknowns that is already very large, which will make model inference extremely difficult if not impossible. For this reason, we only include pairwise interactions in our model.

3.3 Inference of Regulatory Effects

Our model inference framework described later in the next section will infer β_i and β_{ij} in the linear regression model (3.15). If β_i or β_{ij} is not equal to zero with certain statistical significance, then we determine that the ith element or the pair involving the ith and jth elements has regulatory effect. However, it is unclear whether it is an enhancer or silencer, since it is the sign of $w_i - 1$ or $w_{ij} - 1$, not the sign of β_i or β_{ij} that determines an enhancing or inhibitory effect. We will next show that we can infer the regulatory effect of the ith SRE from β_i and $[SF_i]$ that is the concentration of the SF that can bind to the SRE. The enhancing or inhibitory effect of an SRE pair however is difficult to infer.

Let us first consider the situation where only one reference sample is used as in (3.11). In this case, $\beta_i = b_i^{t_1} - b_i^{t_2} = f([SF_i^{t_1}]; w_i) - f([SF_i^{t_2}]; w_i)$, where $f([SF_i^t]; w_i) = \log(\frac{1+w_i k_{sf_i}[SF_i^t]}{1+k_{sf_i}[SF_i^t]})$. Then it can be easily shown that the sign of β_i is the same as the sign of $(w_i - 1)([SF_i^{t_1}] - [SF_i^{t_2}])$. Therefore, if $([SF_i^{t_1}] - [SF_i^{t_2}])\beta_i > 0$, then $w_i > 1$ and SRE i is an enhancer; otherwise, $w_i < 1$ and SRE i is a silencer.

For model (3.13) or (3.15) where multiple tissues are used as the reference, the following proposition can be used to infer the regulatory effect of an SRE:

Proposition 1 *Given the condition*

$$[SF_i^{t_1}] > \frac{1}{T-1} \sum_{t \neq t_1} [SF_i^t], \tag{3.16}$$

we have $w_i > 1$ if $\beta_i > 0$, or $w_i < 1$ if $\beta_i < 0$. Given the condition

$$[SF_i^{t_1}] < \left[\frac{1}{T-1} \sum_{t \neq t_1} \frac{1}{[SF_i^t]} \right]^{-1}, \tag{3.17}$$

we have $w_i > 1$ if $\beta_i < 0$, or $w_i < 1$ if $\beta_i > 0$.

Proof: For simplicity, we will omit subscript i throughout the proof. Define $b = f(q^t) = \log(\frac{1+wq^t}{1+q^t})$, where $q^t = k_{sf}[SF^t]$ as defined earlier. Then $\frac{df(q^t)}{dq^t} = \frac{(w-1)}{(1+wq^t)(1+q^t)}$. If $w > 1$, $f(q^t)$ is a monotonically increasing function; otherwise, $f(q^t)$ is a monotonically decreasing function (Figure 3.1).

Define q_w such that $f(q_w) = \frac{1}{T-1}\sum\limits_{t\neq t_1} f(q^t)$, which gives:

$$q_w = \frac{1 - \prod\limits_{t\neq t_1}(\frac{1+wq^t}{1+q^t})^{\frac{1}{T-1}}}{\prod\limits_{t\neq t_1}(\frac{1+wq^t}{1+q^t})^{\frac{1}{T-1}} - w}, \tag{3.18}$$

where $0 < w < 1$ or $w > 1$. From (3.18), we obtain the following result:

$$\begin{aligned} q_0 = \lim_{w\to 0^+} q_w &= \prod_{t\neq t_1}(1+q^t)^{\frac{1}{T-1}} - 1 < \frac{1}{T-1}\sum_{t\neq t_1} q^t, \\ q_\infty = \lim_{w\to\infty} q_w &= \frac{1}{\prod\limits_{t\neq t_1}(\frac{1+q^t}{q^t})^{\frac{1}{T-1}} - 1} > \frac{1}{\frac{1}{T-1}\sum\limits_{t\neq t_1}(\frac{1+q^t}{q^t}) - 1} = \frac{T-1}{\sum\limits_{t\neq t_1}\frac{1}{q^t}}, \end{aligned}$$

Define $q_a = q_w$ for $w > 1$ and $q_b = q_w$ for $0 < w < 1$. In Lemma 1 given later at the end of this section, we prove that q_w is a monotonically decreasing function of w. Therefore, we have the following inequalities:

$$\frac{T-1}{\sum\limits_{t\neq t_1}\frac{1}{q^t}} < q_a < q_b < \frac{1}{T-1}\sum_{t\neq t_1} q^t. \tag{3.19}$$

From (3.12) and (3.13), we have $\beta = f(q^{t_1}) - \frac{1}{T-1}\sum\limits_{t\neq t_1} f(q^t) = f(q^{t_1}) - f(q_w)$. When $w > 1$, due to the fact that $f(q^t)$ is an increasing function, we have $\beta > 0$ if $q^{t_1} > q^a$ or $\beta < 0$ if $q^{t_1} < q^a$. Similarly, when $w < 1$, we have $\beta < 0$ if $q^{t_1} > q^b$ or $\beta > 0$ if $q^{t_1} < q^b$. This is illustrated in Figure 3.1. Now suppose that $[SF^{t_1}] > \frac{1}{T-1}\sum\limits_{t\neq t_1}[SF^t]$, since $q^t = k_{sf}[SF^t]$, we have $q^{t_1} > \frac{1}{T-1}\sum\limits_{t\neq t_1} q^t$. Using (3.19), we have $q^{t_1} > q_b > q_a$. Therefore, we can infer that $w > 1$ if $\beta > 0$, or $w < 1$ if $\beta < 0$ as illustrated in Figure 3.1. Similarly, if $[SF^{t_1}] < [\frac{1}{T-1}\sum\limits_{t\neq t_1}\frac{1}{[SF^t]}]^{-1}$, we have $q^{t_1} < \frac{T-1}{\sum\limits_{t\neq t_1}\frac{1}{q^t}}$, and thus,

$q^{t_1} < q_a < q_b$. We can infer that $w > 1$ if $\beta < 0$, or $w < 1$ if $\beta > 0$, again as illustrated in Figure 3.1. Note that q_a and q_b are determined by unknown parameter w and they can be anywhere in between q_∞ and q_0. Therefore, if $[\frac{1}{T-1}\sum\limits_{t\neq t_1}\frac{1}{[SF^t]}]^{-1} < [SF^{t_1}] < \frac{1}{T-1}\sum\limits_{t\neq t_1}[SF^t]$, we can not determine whether $w > 1$ or $w < 1$.

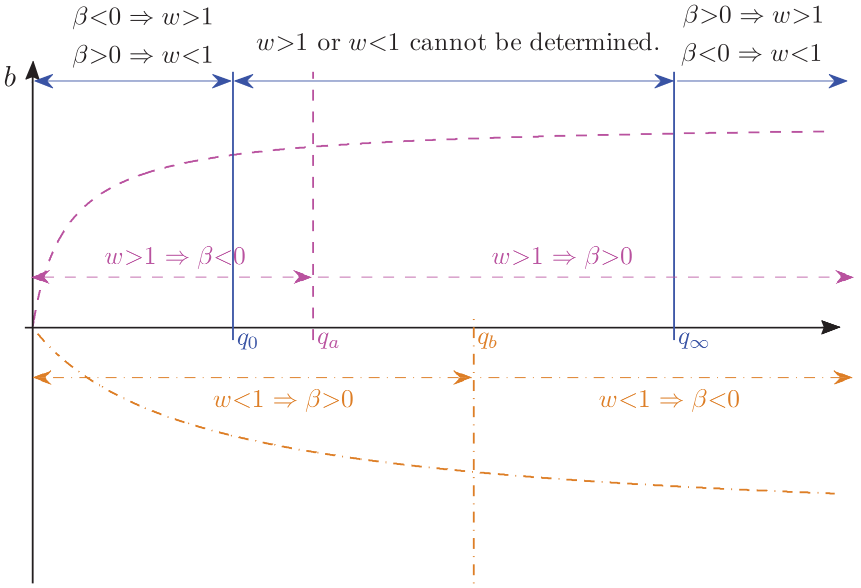


Figure 3.1: Illustration of Proposition 1. The dashed curve is $b = f(q) = \log(\frac{1+wq}{1+q})$ for $w > 1$ and the dashed lines with arrows define the region where $\beta < 0$ or $\beta > 0$. The dash-dot curve is $b = f(q) = \log(\frac{1+wq}{1+q})$ for $w < 1$ and the dash-dot lines with arrows define the region where $\beta > 0$ or $\beta < 0$. The solid lines define the decision region of w described in Proposition 1.

Lemma 1

$$h(x) = \frac{1 - \prod\limits_{i=1}^{N}(\frac{1+xq_i}{1+q_i})^{\frac{1}{N}}}{\prod\limits_{t=1}^{N}(\frac{1+xq_i}{1+q_i})^{\frac{1}{N}} - x}. \tag{3.20}$$

is a monotonically decreasing function for $x \in (0,1)\bigcup(1,\infty)$*, when* q_i*,* $i = 1, ..., N$ *are positive.*

Proof: Define $g(x) = \prod(\frac{1+xq_i}{1+q_i})^{\frac{1}{N}}$. Then $\frac{dh(x)}{dx} = \frac{g'(x)(x-1)-g(x)+1}{[g(x)-x]^2}$, where $g'(x) = \frac{dg(x)}{dx}$. Let us define the numerator of $h(x)$ as $J(x)$, since the denominator is positive, we next prove that $J(x) \leq 0$, which implies that $\frac{dh(x)}{dx} \leq 0$ and therefore $h(x)$ is a decreasing function.

$$\begin{aligned}
J(x) &= g'(x)(x-1) - g(x) + 1 \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{1+xq_i}g(x)(x-1) - g(x) + 1 \\
&= \left[\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{1+xq_i}(x-1) - 1 + \frac{1}{g(x)}\right]g(x) \\
&= \left[\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{1+xq_i}(x-1) - 1 + \prod_{i=1}^{N}(\frac{1+q_i}{1+xq_i})^{\frac{1}{N}}\right]g(x) \\
&\leq \left[\frac{1}{N}\sum_{i=1}^{N}\frac{1+xq_i}{1+xq_i} - 1\right]g(x) \\
&= 0,
\end{aligned}$$

where the first inequality is due to the facts that $g(x) > 0$ and that $\prod\limits_{i=1}^{N}(\frac{1+q_i}{1+xq_i})^{\frac{1}{N}} \leq \frac{1}{N}\sum\limits_{i=1}^{N}(\frac{1+q_i}{1+xq_i})$, since $x > 0$ and $q_i > 0, i = 1, ..., N$. Thus, $h(x)$ is monotonically decreasing in $(0,1)$ and $(1,\infty)$. Moreover, since we have:

$$\lim_{x\to 1^-} h(x) = \lim_{x\to 1^+} h(x) = \frac{\sum\limits_{i=1}^{N}\frac{q_i}{1+q_i}}{N - \sum\limits_{i=1}^{N}\frac{q_i}{1+q_i}},$$

$h(x)$ is a monotonically decreasing function for $x \in (0,1)\bigcup(1,\infty)$, and $\frac{dh(x)}{dx} = 0$ if and only if $q_1 = q_2 = ... = q_N$.

3.4 Framework for Model Inference

We applied the thermodynamic model (3.15) derived in Section 3.2,

$$y = \beta_0 + \sum_{i\in\mathcal{M}} x_i\beta_i + \sum_{(i,j)\in\mathcal{I}} x_i x_j \beta_{ij} + \epsilon,$$

to the identification of SREs and SRE pairs involved in alternative splicing, where y, $\mathcal{M}$, $\mathcal{I}$, x_i, β_i and β_{ij} are described in Section 3.2.5.

We first determined a set of ASEs from the UCSC KnownGene table [78]. We extracted all the hexamers in five regions around ASEs as candidate SREs (Figure 3.2A), and obtained y for this set of ASEs. Then, model inference was carried out using four components described in Figure 3.2B. Since we considered all 6-mers and their interactions, the regression model (3.15) contained 5×4^6 variables for the main effects and $> 10^8$ variables for the interactions. The huge number of variables not only requires huge computation for model inference, but also may yield a large number of false SREs. To overcome these problems, we developed the four-component framework in Figure 3.2B to select reliable SREs without overfitting the model. We will describe each of these steps in detail in this section.

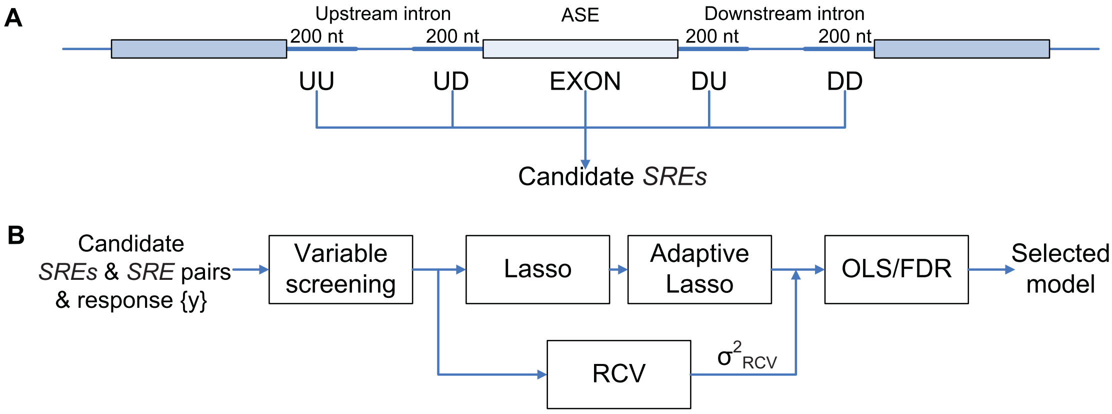


Figure 3.2: Five regions around ASEs used to extract SREs and the model inference framework. (A) All hexamers in the five regions around ASEs are considered as candidate SREs. UU (upstream/upstream) stands for the 5' end of the upstream intron. UD (upstream/downstream) denotes the 3' end of the upstream intron. DU and DD are defined in a similar way. EXON stands for the ASE region. (B) The inference framework for detecting active SREs and SRE pairs.

3.4.1 ASE selection

The KnownGene table of human February 2009 assembly (hg19) was downloaded from the University of California Santa Cruz (UCSC) genome database [78]. We chose UCSC Known Genes as the reference gene annotation, since they contain a comprehensive gene set that is constructed mostly from experimental data in Genbank and Uniprot [79]. For each gene, the KnownGene table gives all known isoforms of its mRNA transcripts. An exon was selected as an ASE in our dataset if the following five criteria were satisfied: 1) at least one isoform includes the exon, 2) at least one isoform does not include the exon, 3) the upstream 5' splice site is the same in all isoforms, and similarly the downstream 3' splice site is the same in all isoforms, as illustrated in Figure 3.3A, 4) the upstream 3' splice site is the same in all isoforms with the ASE, and similarly the downstream 5' splice site is the same in all isoforms with the ASE, as illustrated in Figure 3.3A, and 5) both the upstream and downstream introns are of $\geq$ 400 nts. With these strict criteria, the five regions of the ASE shown in Figure 3.2A are defined without ambiguity. Note that isoforms in Figure 3.3B do not satisfy criterion 3), and thus ASEs with isoforms in Figure 3.3A and 3.3B are not included in our data set. Similarly, isoforms in Figure 3.3C do not satisfy criterion 4), and ASEs with isoform in Figure 3.3A and 3.3C are not included in our data set. To ensure a reliable estimate of the expression ratio, we only kept ASEs with gene expression level greater than 3 RPKM (reads per kilobases per million mapped reads). This gave a set of ASEs for each tissue. The number of ASEs for each tissue is given in Table 3.1 and more detailed description of the ASEs including their genomic coordinates are given in Table S2 of [2]. Most ASEs were used for model inference in almost all tissues. Some ASEs were not used in a specific tissue because they did not pass the minimum expression requirement in that tissue. Note that although almost the same set of ASEs were used, the splicing response variable y of the same ASE was different in different tissues, and thus the data used for model inference were in fact different for different tissues.

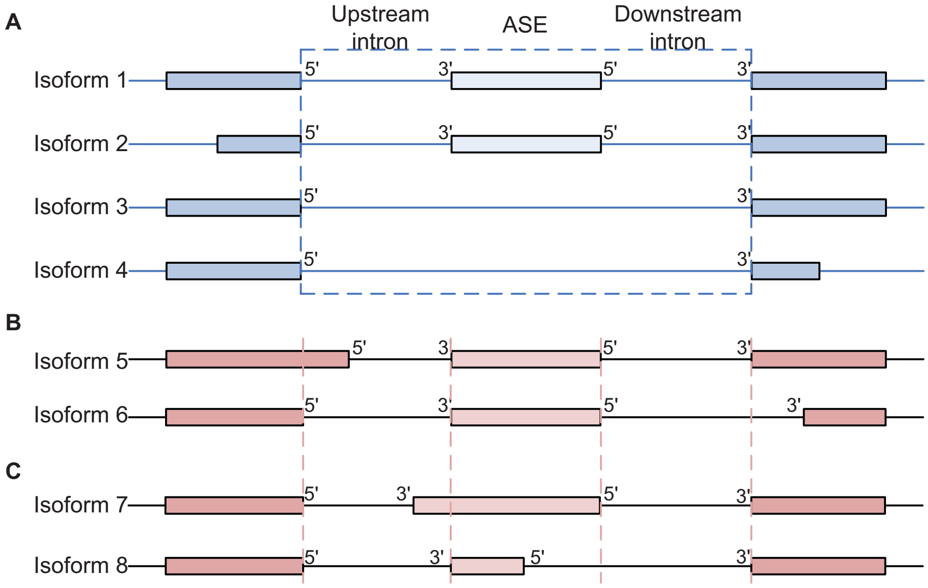


Figure 3.3: Illustration of ASE selection criteria 3 and 4. (A) ASEs that satisfy both criteria 3 and 4. (B) ASEs that do not satisfy criterion 3. ASEs with isoforms in (A) and (B) are not included in our ASE data set. (C) ASEs that do not satisfy criterion 4. ASEs with isoforms in (A) and (C) are not included in our ASE data set.

3.4.2 RNA-Seq Data

The data set in [17] includes RNA-Seq reads from 9 tissues: adipose, whole brain, breast, colon, heart, liver, lymph node, skeletal muscle and testes, as well as several cerebellar cortex samples and cell lines. We only used RNA-Seq data of 9 tissues, which contains over 200 million reads of 32 nts, to detect SREs and cooperative SRE pairs.

3.4.3 Estimation of Expression Level and Inclusion Ratio

We started by mapping the RNA-Seq reads against an expanded human genome (hg19) downloaded from the UCSC genome database, allowing up to two mismatches, using Bowtie (version 0.12.7) [40]. The expanded human genome consists of the UCSC hg19 whole genome reference sequence and the 56 nt long splice-crossing sequences for each exon junction documented in the UCSC KnownGene table. Reads that could be mapped to multiple loci of the genome were excluded, and 140 million uniquely mapped reads were kept for the following analysis.

We next calculated the expression level of each isoform including or excluding a selected ASE in 9 tissues using the algorithm of Jiang *et al.* [18]. Since we only kept uniquely mapped reads, for an exon of length l, we used an effective exon length $l-r-m$, where r is the read length and m is the number of multi-mappable positions of the exon. To find out m, we re-mapped all possible 32-nt subsequences of candidate ASEs and splice junctions against the same expanded genome described above using Bowtie [40]. Moreover, to minimize the effect of non-uniformity of read distribution [14], we only used three exons, including the ASE itself, the adjacent upstream and downstream exons to estimate the expression level of each isoform.

After the expression level of each isoform of the selected gene was calculated, the inclusion ratio (IR) of an ASE in a specific tissue was calculated as the ratio of the expression level of the isoforms with the ASE to the total expression level of all isoforms of the gene, *i.e.*, $IR = \frac{E_{I_1}}{E_{I_1}+E_{I_2}}$, where E_{I_1} is the total expression level of isoforms including the ASE and E_{I_2} is the total expression level of isoforms excluding the ASE.

3.4.4 RNA Sequence Elements

For each tissue and each ASE, we extracted all hexamers in five regions around the ASE, including the 200 nts intronic region adjacent to the upstream 5' splice site (UU in Figure 3.2A), the 200 nts intronic region adjacent to the upstream 3' splice site (UD in Figure 3.2A), the ASE region (EXON in Figure 3.2A), the 200 nts intronic region adjacent to the downstream 5' splice site (DU in Figure 3.2A) and the 200 nts intronic region adjacent to the downstream 3' splice site (DD in Figure 3.2A). The EXON region is the ASE itself if the ASE is less than 200 nts; otherwise, it is the combination of the first and last 100 nts of the ASE. Since the 5' and 3' splice sites have the consensus sequences MAG/GURAGU and $\mathrm{Y_{10}NCAG/G}$ [8,41], respectively, we excluded the sequences in the window from -3 to 6 around the 5' splice site and in the window from -14 to 1 around the 3' splice site in our analysis.

3.4.5 Variable Screening

In the first variable screening component in Figure 3.2B, we used a strategy similar to the sure independence screening method [80] to reduce the dimensionality of the feature space, thereby improving variable selection in terms of both speed and accuracy. Specifically, for the ith hexamer, $i \in (1, ..., 4^6)$, we used the following simple linear regression to test the correlation between its presence in one of the five regions of ASEs with the response variable:

$$y_e = \beta_0 + x_{ei}\beta_i + \epsilon_e, \tag{3.21}$$

where $x_{ei}, e = 1, ..., n$ is a binary variable to indicate if the ith hexamer is present ($x_{ei} = 1$) or absent ($x_{ei} = 0$) in one of the five regions of the eth ASE, and y_e

is the splicing response of the eth ASE as defined earlier, and $\epsilon_e, e = 1, ..., n$ are independent and identically distributed normal random variables. In some samples, we have inclusion ratio $IR_e = 1$ or $IR_e = 0$, usually due to the low read abundance of the minor isoform. For these samples, we set $\log(\frac{E_{e,I_1}}{E_{e,I_2}}) = 10$ for $IR_e = 1$ or set $\log(\frac{E_{e,I_1}}{E_{e,I_2}}) = -10$ for $IR_e = 0$, which is equivalent to $IR_e \approx 0.9999$ or $IR_e \approx 1e^{-5}$. Hexamers having a significant correlation with a p-value < 0.05 were kept in set $\mathcal{M}$ for further analysis with the Lasso and the adaptive Lasso.

In the next step, for each pair of the retained hexamers, their interaction was tested using the model:

$$y_e = \beta_0 + x_{ei}\beta_i + x_{ej}\beta_j + x_{ei}x_{ej}\beta_{ij} + \epsilon_e. \tag{3.22}$$

Interaction terms with a p-value < 0.05 were also kept in set $\mathcal{I}$ for further analysis. To reduce the possible false positive effects, we also required that the co-occurrence frequency of the two hexamers in an interaction pair was significant (p-value < 0.05 from a hypergeometric test based on the null hypothesis that the presence of the first hexamer is independent of the presence of the second hexamer) in the five regions of the selected ASEs defined earlier, and that any hexamer or hexamer-pair must be present in at least 1% of the ASEs.

3.4.6 The Lasso and The Adaptive Lasso

In the second component, we adopted the Lasso [81] and the adaptive Lasso [82] to perform penalized multiple regression to select variables. The Lasso is known to shrink many variables with no or small correlation with the response variable to zero [81], and thus yields a sparse model that only contains a small number of variables. The shrinkage techniques has been widely applied to various problems to solve the biological complexity [83–85]. Using both the Lasso and the adaptive Lasso was to ensure more reliable variable selection as suggested in [86].

We define $\mathbf{y} = (y_1, y_2, ..., y_e, ..., y_n)^T$ and $\mathbf{x}_i = (x_{1i}, x_{2i}, ..., x_{ei}, ..., x_{ni})^T$, where y_e and x_{ei} are defined earlier. We also define $\mathbf{x}_i.{*}\mathbf{x}_j$ as element-wise multiplication of two vectors. The Lasso procedure was performed by solving the following problem [81]:

$$\{\hat{\beta}_0, \hat{\beta}_i, \hat{\beta}_{ij}\} = \underset{\beta_0,\beta_i,\beta_{ij}}{\operatorname{argmax}} \left\{ \left\| \mathbf{y} - \beta_0 - \sum_{i\in\mathcal{M}} \mathbf{x}_i\beta_i - \sum_{(i,j)\in\mathcal{I}} \mathbf{x}_i.{*}\mathbf{x}_j\beta_{ij} \right\|^2 + \lambda\Big(\sum_{i\in\mathcal{M}} |\beta_i| + \sum_{(i,j)\in\mathcal{I}} |\beta_{ij}| \Big) \right\}. \tag{3.23}$$

The optimal value of parameter λ was obtained using 100-fold cross-validation based on the mean squared prediction error. Then we chose $\hat{w}_i = 1/|\hat{\beta}_i|$ and $\hat{w}_{ij} = 1/|\hat{\beta}_{ij}|$ and solved the following adaptive Lasso problem [82]:

$$\{\hat{\beta}'_0, \hat{\beta}'_i, \hat{\beta}'_{ij}\} = \underset{\beta'_0,\beta'_i,\beta'_{ij}}{\operatorname{argmax}} \left\{ \left\| \mathbf{y} - \beta'_0 - \sum_{i\in\mathcal{M}} \mathbf{x}_i\beta'_i - \sum_{(i,j)\in\mathcal{I}} \mathbf{x}_i.{*}\mathbf{x}_j\beta'_{ij} \right\|^2 + \lambda\Big(\sum_{i\in\mathcal{M}} \hat{w}_j|\beta'_i| + \sum_{(i,j)\in\mathcal{I}} \hat{w}_{ij}|\beta'_{ij}| \Big) \right\}. \tag{3.24}$$

The optimal value of λ was also obtained using 100-fold cross-validation. We solved these problems using the coordinate descent algorithm of Friedman et al. [87] implemented in the 'glmnet' package.

3.4.7 Refitted Cross-validation

Although the Lasso and the adaptive Lasso only retained a small number of variables in the model, we wanted to ensure that the overfitting problem did not occur. To this end, we added the third component named refitted cross-validation (RCV) [88] to the inference procedure.

RCV is a technique to estimate residual variance in linear regression models of ultrahigh dimension [88]. In our case, the n samples were randomly split into two even datasets. We applied the Lasso to the first dataset to select a set $\mathcal{V}$ of variables from the variables in $\mathcal{M}$ and $\mathcal{I}$ resulted from the variable screening procedure. We then again used the Lasso to refit the model with the variable set $\mathcal{V}$ to the second dataset. The refitting process selected a set $\mathcal{V}'$ of variables from $\mathcal{V}$. Finally, the variance of the residual error $\hat{\sigma}_1^2$ is estimated from the second dataset with variables in $\mathcal{V}'$ using the OLS method. We reversed the role of the two datasets, and obtained another estimate of the variance of the residual error, $\hat{\sigma}_2^2$. The final estimate is then defined as $\hat{\sigma}^2 = (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2$. We repeat this process 100 times by randomly splitting the dataset, and the average variance σ_{RCV}^2 of the 100 estimates was the final estimate of the residual variance.

We then used σ_{RCV}^2 to select the final optimal model. The adaptive Lasso procedure in the second component produced a sequence of models for different values of λ. For each λ, we extracted the variables of the model and estimated the residual variance σ^2 with these selected variables using the OLS method. We then compared the minimum value of the residual variance, σ_{min}^2, with σ_{RCV}^2. If $\sigma_{min}^2 > \sigma_{RCV}^2$, we selected the model (variables) that gave σ_{min}^2; otherwise, we identified the value of λ that yielded a residual variance equal to σ_{RCV}^2 and selected the model (variables) with this λ.

3.4.8 Correction for Multiple Testing

The adaptive Lasso together with RCV selected selected a set of variable in the final model, but it did not give p-value for each variable. In the last component, we used OLS method to refit the model and calculated the p-value for each variable. Based on these p-values, we chose the variables at an FDR ≤ 0.01 [89]. The final model contained these variables and the percentage of the variance explained (R^2) by this final model was calculated as:

$$R^2 = 1 - \sum_{e=1}^{n} (y_e - \hat{y}_e)^2 / \sum_{e=1}^{n} (y_e - \bar{y})^2, \tag{3.25}$$

where $\hat{y}_e$ is the predicted value of y_e from the final model, and $\bar{y}$ is the sample mean of y_e. Note that R^2 was also used in [69, 73, 76] as the figure of merit for performance evaluation, which will be used for comparison in the following Section.

3.5 Results

3.5.1 Performance of the model and the regression framework

As described in Section 3.2 and 3.4, we selected a set of ASEs for each tissue from the KnownGene table and calculated the inclusion ratio $E_{I_1}/(E_{I_1} + E_{I_2})$ from the RNA-Seq data [17] for each ASE, which was then used to calculate the response y in model (3.15). We applied our thermodynamic model and inference framework to this data set to detect SREs and SRE pairs. The number of ASEs used in model inference, the number of SREs and SRE pairs in the final model and the percentage variance explained by the final model are given in Table 3.1. In each tissue, our thermodynamic model explained 49.1%-66.5% of the variance in the data (see R^2 in

Table 3.1), which was comparable to that achieved by the best model for transcription reported in [73]. More specifically, the linear model for gene transcription in [69] explained 9.6% of the variance on average, while the regression spline model for gene expression that incorporated interaction terms explained 13.9% to 32.9% of the variance [76]. Most recent work by Gertz *et al.* [73] fitted a nonlinear thermodynamic model to the expression data of synthetic genes. Their models explained 44-59% of the variance in gene expression. Thus, considering the non-synthetic genes we used, our model has captured a large fraction of the variance in the splicing response.

Overall, 619 different SREs and 196 SRE pairs were detected from different tissues (Table 3.1, Table S1 in [2] and Appendix table B.1). These SREs and SRE pairs consist of 854 different hexamers. Many SREs are very similar to each other, which may arise from an SRE longer than 6 nucleotides (nts) or from a degenerated motif. Note that although they were detected in different tissues, the SRE and SRE pairs *per se* do not give rise to tissue-specific isoforms, because they are always there in pre-mRNA sequences in different tissues. It is the tissue-specific expression of SFs (*e.g.*, Figure 3.4) that regulates tissue-specific splicing through SREs.

Table 3.1: Summary of the final selected models in each tissue

Tissue	No. of ASEs	No. of SREs	No. of SRE pairs	R^2 (%)
Adipose	1345	60	30	52.9
Brain	1411	65	19	49.1
Breast	1399	82	24	58.3
Colon	1236	64	13	49.4
Heart	1302	84	21	60.5
Liver	995	76	19	66.5
Lymph node	1405	77	24	55.7
Skeletal muscle	1174	85	18	63.2
Testes	1601	79	28	51.0

3.5.2 Comparison with experimental validated SREs

We compared our 854 SREs with the results of Castle *et al.* [30], who performed a systematic screening of all 4 to 7-mers for cis-regulatory motifs enriched near ASEs using microarray. We only kept 5 to 7-mers with a p-value smaller than 10^{-3} (Bonferroni-corrected p-values as described in [30]) for the comparison. In total, 137 non-redundant 5 to 7-mers were left (91 5-mers, 28 6-mers and 18 7-mers). We considered it as a match if a 7-mer contains one of our SREs, a 5-mer is part of our SREs, or a 6-mer exactly matches one of our SREs. This yielded 103 k-mers, $k = 5, 6, 7$, that could find at least one match in our SREs. In order to evaluate the significance of the overlap, we generated a list of 6-mers from the 137 k-mers of Castle *et al.* with the following procedure. For each 5-mer, 8 different 6-mers containing the 5-mer were obtained by padding a nucleotide to the beginning or the end of the 5-mer. For each 7-mer, 2 different 6-mers were obtained by extracting the first or the last 6 nts. In total, 639 different 6-mers were extracted from Castle's k-mers, $k = 5, 6, 7$. A significant number of 6-mer (180) were found in both our 854 SREs and the 639 6-mers obtained from 137 k-mers of Castle *et al.* (p-value=9.37e-7 from Fisher's exact test).

We also compared our 854 SREs with the binding sites of 25 SFs experimentally identified with SELEX [51, 90–108] or RNAcompete [109]. Each of [51, 90–108] attempted to determine the binding sites of 1 to 3 SFs using SELEX, and [51, 90–108] reported SELEX results for a total of 25 SFs. For each of 25 SFs, we obtained a set of RNA sequences selected with SELEX from one of [51, 90–108]. If there are more than one SELEX results for an SF, we used the most recent SELEX result. We then extracted the consensus sequences embedded in this set of RNA sequences as the binding sites of the SF. If a consensus binding site is 4 or 5 nt long, one more nucleotide was also extracted from each side of the original selected sequence to obtain hexamers. If a consensus binding site is longer than 6 nts, all the hexamers included in the consensus were extracted. For RNAcompete, the 7-mers listed in Figure 2 of [109] were used as consensus binding sites, and two hexamers were taken from each 7-mer. In total, 709 different hexamers were obtained from the consensus binding sites. A significant number (175) of hexamers were found in both our 854 SREs and the 709 hexamers obtained from SELEX or RNAcompete (p-value=0.004 from Fisher's exact test). In Table S1 of [2], we gave the SF that was identified in [51, 90–109] to bind to one of the 175 hexamers. Although the overlap between our predicted SREs and the binding sites of SFs determined with SELEX and RNAcompete is statistically significant, a relative large number (679) of our predicted SREs are not included in these experimental results, which implies that our result contains some novel SREs.

3.5.3 Experimental Evidence of SREs

Several well-defined SREs involved in tissue-specific AS have been detected in our work. We will take SREs bound by Fox-1 protein, polypyrimidine tract binding protein (PTB), quaking protein (QKI), muscleblind-like protein (MBNL) as examples (listed in Table 3.2) to illustrate how to understand the results and compare them with the available experimental evidences.

Table 3.2: Selected examples of the detected SREs

SF	SRE(region)	p-value	Tissue	Effect
Fox-1	UGCAUG(DU)	7.02E-13	muscle	enhancer
Fox-1	UGCAUG(UU)-GCAUGU(UU)	1.34E-3	heart	unknown
PTB	CUCUCU(UD)	9.76E-6	lymph node	silencer
PTB	UUCUCU(UD)	9.33E-4	adipose	silencer
QKI	ACUAAC(UD)	3.36E-8	muscle	silencer
MBNL	UGCUGC(UU)	1.22E-6	lymph node	unknown
MBNL	UGCUGC(EXON)	2.61E-3	muscle	unknown

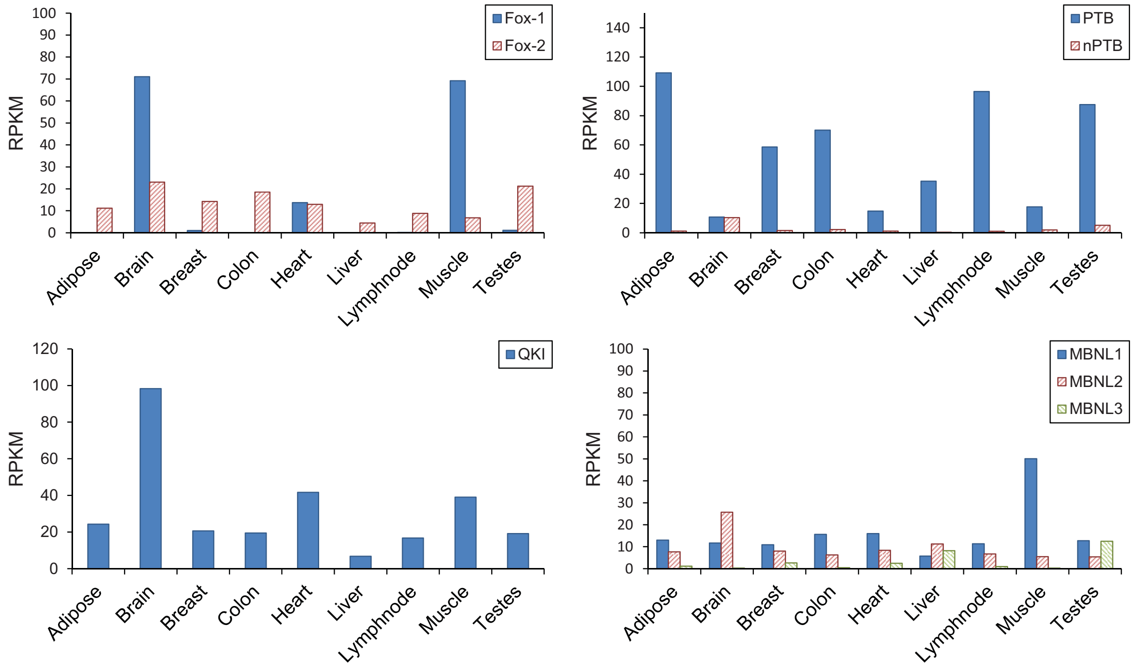


Figure 3.4: Expression level of several splicing factors in 9 tissues. The expression levels are calculated from the RNA-Seq data as reads per kilobases per million mapped reads (RPKM).

Fox-1

Fox protein recognizes [U]GCAUG as its SRE and it has been shown to be one of the most conserved regulators of tissue-specific AS in metazoans. Fox-1 is exclusively expressed in brain, heart, and skeletal muscle as reported in [61], which is consistent with the RNA-Seq data we used as shown in Figure 3.4. Its paralog Fox-2 has relatively low expression level in all tissues. In our results, we detected two SREs containing UGCAUG as summarized in Table 3.2. These 2 SREs were detected in heart or muscle, which is consistent with the tissues where Fox-1 is expressed. Note that the second SRE UGCAUG(UU)-GCAUGU(UU) detected in upstream region in heart is actually an interaction term in the regression model. Since our model includes interaction between SREs in the same region or from different regions, it is possible that an SRE longer than 6 nts is detected as an interaction term. This interaction term actually arises from a 7-mer SRE UGCAUGU(UU) [110] (Among the 28 UGCAUG(UU)-GCAUGU(UU) pairs used for inference in heart, 27 are derived from UGCAUGU(UU)). Using the inference method for regulatory effects described in Materials and Methods, we found that the SRE that we identified from muscle are enhancers in the downstream region (DU) (Table 3.2), which is consistent with the computational analysis [17] and the experimental evidence [61, 111].

PTB

Another well-defined SF is PTB which binds to UC-rich SREs and has high binding affinity to UCUU and/or UCUCU [49, 112]. Two SREs we identified are possible binding sites of PTB (Table 3.2). They are found in the upstream region in adipose and lymph node. Both SREs are detected from the tissues where PTB are over-expressed as shown in Figure 3.4. Using Proposition 1 in Materials and Methods, These two SREs were determined to be a silencer in the upstream of ASEs. This result coincides with position-dependent alternative splicing activity of the PTB as identified using microarrays [113].

QKI

One of QKI's binding site ACUAAY [103,114] was detected in the upstream intron of the ASE in muscle (Table 3.2). The SRE ACUAAC was previously reported as an over-represented motif in the downstream region of ASEs in muscle [65] and was also predicted as an upstream intronic SRE specific to the central nervous system [115]. Our result indicates that this upstream SRE detected in muscle is a silencer, which is consistent with the experimental result [116].

MBNL

The MBNL family protein can bind to the motif YGCU(U/G)Y [117]. Two sequences of this motif can be found in our SREs (Table 3.2). However, the function of this motif is complicated. First, both MBNL and CELF family proteins can bind to similar motifs [118, 119], although different protein isoforms may have different binding specificities [45]. Second, MBNL can either promote or repress splicing of specific ASEs on different pre-mRNAs by antagonizing the activity of CELF proteins [117]. Thus, although our result indicates that this SRE is related to AS regulation in muscle and lymph node, it is difficult to interpret its regulatory effect based on current result.

3.5.4 Experimental Evidence of SRE pairs

We also identified 196 different cooperative SRE pairs (Table S1 in [2] and Appendix table B.1). Thirty-nine percent (77 SRE pairs) of them are interactions of SREs in the same region. As discussed in the previous section, a part of this kind of interaction may come from a single SRE longer than 6 nts. Among the remaining interactions between different regions (119 SRE pairs), the most frequent interactions are SREs in region pairs UD-DU, UU-UD and UD-DD. Region pair UD-DU may reflect the effect in the exon definition stage of spliceosome assembly, and region pair UU-UD may reflect the effect in the intron definition stage. A summary of interaction between SRE pairs in different regions is given in Figure 3.5. Several detected SRE pairs are listed in Table 3.3.

Interestingly, Several previously identified SREs were detected in our interaction results. They either cooperate with their own or other different SREs in a different region or in the same region.

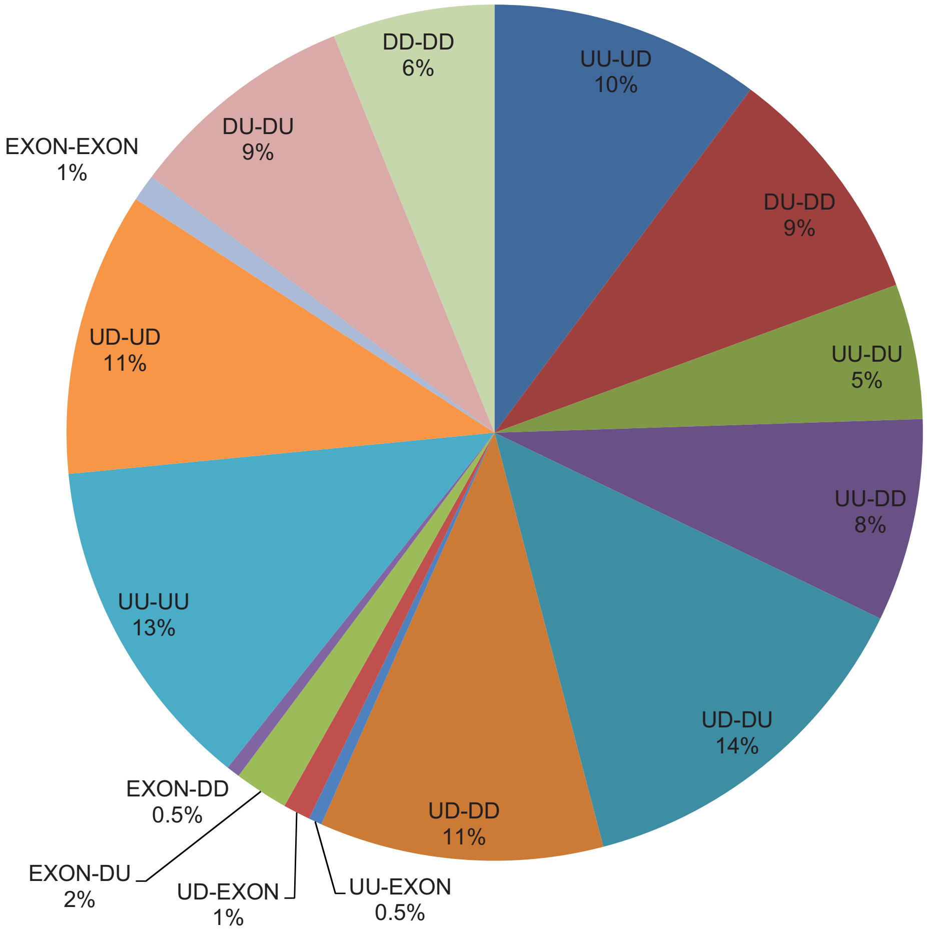


Figure 3.5: Percentage of different types of SRE pairs. Two hundred and forty-one SRE pairs are detected in different or same regions. This figure shows breakdown of the SRE pairs in different regions.

Interaction Between PTBs

Two SRE pairs bound by PTB [49, 112] were identified in our result. All the four SREs located at the upstream region of the 3' splice site which can also be a part of an extended polypyrimidine tract, although the 3' splice site consensus of 15 nts containing the classical polypyrimidine tract [8, 9] has been removed in our analysis. One SRE pair UUCUCU-UCCUCU was identified in the upstream intron in brain. The other interaction detected involves PTB's binding site UCUUCC in the UD region and CCUUCU in the DD region (Table 3.3). The downstream CCUUCU resembles

Table 3.3: Selected examples of the detected cooperative SRE pairs

SF_1	SRE_1(region)	SF_2	SRE_2(region)	p-value	Tissue
PTB	UUCUCU(UD)	PTB	UCCUCU(UD)	2.02E-4	brain
PTB	UCUUCC(UD)	PTB	CCUUCU(DD)	1.02E-6	brain
F/H	GGGGCA(DU)	F/H	UGGGGA(DD)	1.31E-3	liver
E/K	CCCCAG(UU)	E/K	CCGCCC(UU)	8.02E-8	lymph node
E/K	ACCCCU(DU)	E/K	CCCCUC(DU)	8.70E-4	adipose
T/H/S	AUUUAU(UU)	T/H/S	UAAAUG(UD)	2.33E-10	brain
T/H/S	AUUUAC(DU)	T/H/S	AAUAAA(DD)	3.04E-9	lymph node
T/H/S	AAAUUU(UD)	T/H/S	UUUUUU(DU)	3.71E-7	lymph node
T/H/S	AAUAUG(UU)	T/H/S	AAAAAU(UD)	3.75E-5	heart
T/H/S	AUUUAG(UD)	T/H/S	UUAUAU(DU)	1.02E-4	testes
T/H/S	UUUAAU(UD)	T/H/S	UUUUAU(DD)	1.45E-4	adipose
T/H/S	AAAUUC(UU)	T/H/S	AAAUUU(DU)	6.08E-4	lymph node
T/H/S	UUUUAA(UD)	T/H/S	CAUUAU(DU)	8.93E-4	adipose
T/H/S	UUUAAU(UU)	T/H/S	AAAAUA(UD)	1.09E-3	heart
T/H/S	GUUUUA(UD)	T/H/S	UUUUAU(DD)	1.90E-3	adipose
T/H/S	AAUAUU(UD)	T/H/S	AUUUUG(DD)	3.06E-3	breast
T/H/S	UAUUUA(UD)	QKI	AACUAA(DU)	2.47E-6	liver
T/H/S	AUUUAA(UU)	PTB	CUUUUC(UD)	6.12E-4	heart

*F/H represent splicing factor hnRNP F or hnRNP H.
*E/K represent splicing factor hnRNP E or hnRNP K.
*T/H/S represent TIA-1/TIAR, Hu family or Sam68 proteins.

the PTB binding sites and might also be bound by PTB. In a cooperative model for the regulatory mechanism of PTB, it was proposed that a single PTB or a PTB dimer could loop out the branch point upstream of 3' splice site or the ASE by binding to several pyrimidine tracts upstream and downstream to repress splicing [120]. A single PTB-binding element had weak silencing activity, while multiple PTB-binding sites at both upstream and downstream of the ASE or of a branch point could have strong inhibitory effect [121]. This model and the experimental results in [120, 121] are consistent with the interaction pairs we identified.

Interaction Between hnRNP F/H

hnRNP F/H can bind to GGGA and G-rich SREs [12, 122]. The proposed mechanism underlying the splicing regulation of hnRNP F/H involves an interaction between proteins bound at both ends of an intron that loops out the intron and brings distantly separated exons into closer proximity [123]. Consistent with this model, one pair of SREs GGGGCA(DU)/UGGGGA(DD) (Table 3.3) putatively bound by the hnRNP F/H was detected in two ends of the downstream intron.

Interaction in hnRNP E/K

Both hnRNP E and hnRNP K proteins contain three copies of the KH domain arranged in a very similar manner, and they are the major poly-C binding proteins in mammalian cells [99, 124]. We extracted all the SRE pairs in our result that contain at least 4 Cs. Two such SRE pairs (Table 3.3) were found to resemble the binding motif of the hnRNP E/K [99]. One SRE pair is due to a longer motif ACCCCUC at downstream intron. The other SRE pair was detected as interaction in the same region. This result may reflect the fact that the three KH domains bind RNA synergistically while a single KH domain appears to have very low level of RNA binding activity [125].

Interaction Between AU-rich Elements

An interesting outcome of this work was the identification of many AU-rich elements in SREs and SRE pairs. The AU-rich elements have been identified previously by comparative analysis as a large class of conserved mammalian ISEs [64]. In plant splicing system, the AU-rich sequences in upstream and downstream introns appear to be involved in early intron recognition and stabilization of spliceosomal complex [126];

and multiple short AU-rich SREs could cooperatively modulate splice site usage [127]. However, the interaction between AU-rich SREs has not been reported in mammals. In our result, The SRE pairs involved in two AU-rich SREs from different regions are listed in Table 3.3. Among these SRE pairs, 7 pairs were detected in upstream and downstream introns, and 4 were detected at two ends of the same intron. These AU-rich SREs resemble the binding sites of TIA-1/TIAR, Hu protein and Sam68 [93,128]. Hu proteins can bind to both upstream and downstream SREs. It has been speculated that binding of Hu protein to multiple sites both upstream and downstream of an ASE could loop out the ASE and as a result block exon definition to repress splicing [128]. Moreover, Hu proteins can promote exon inclusion by binding to AU-rich sequences that are conserved at both upstream and downstream of ASE [129]. The intricate experimental results stress the importance of further mutation analysis to study the cooperative interactions of AU-rich binding proteins.

We also identified many interactions between AU-rich elements and binding sites of other known SFs. For example, We found one interactions between upstream AU-rich SRE and a downstream SRE resembling QKI's binding site. Another example is the interactions between upstream AU-rich SRE and a CU-rich SRE resembling polypyrimidine tract (Table 3.3). These results indicate that AU-rich SREs and cooperative pairs may play an important regulatory role in mammalian AS and worth further experimental investigations. In summary, various cooperative mechanisms could be detected in our result as an interaction, which implies the important role of interaction in splicing regulation.

3.6 Discussion

Our regression model for splicing regulation is derived from the same thermodynamic principle regarding protein and nucleic acids interaction as the one used to derive the model for transcription [73–75, 130]. However, our model uses the ratio of the expression levels of two isoforms or equivalently the ratio of binding and unbinding probabilities of the spliceosome to the mRNA as the response variable, whereas the model for transcription uses the gene expression level or equivalently the binding probability of the RNAP to the promoter as the response variable. For this reason, our model for splicing is linear with respect to unknown parameters to be inferred, whereas the model for transcription is nonlinear. While the nonlinear model for transcription with relatively small number of unknown parameters can be directly inferred [73], a linear approximation [69–72] and an nonlinear approximation using splines [65, 76] have been proposed to facilitate model inference. It was shown that the spline approximation [72, 76] can offer significantly better performance than the linear approximation in terms of the variance explained by the model. The linear regression model [33] and the spline regression model [65] for splicing regulation use the expression level of an isoform as the response variable. Therefore these two models are also approximate models. In contrast, our regression model is directly derived from the thermodynamic principle and the linearity of our model with respect to unknown parameters enables efficient model inference even when the number of unknown parameters is very large. This explains why the variance explained by our model is comparable to the best result achieved by the nonlinear thermodynamic model for transcription [73]. Our model may be improved to explain more variance, for example, by including interactions involved more than two SREs, by accounting for the number of occurrences of each SRE, and by including SREs of length not necessarily equal to six nucleotides, although this can increase the complexity of model inference. On the other hand, if we are interested in the identification of SREs interacting with the binding sites of a specific SF, we can design well-controlled experiments which measure the splicing profile of all the genes before and after knockdown of a specific SF, and then apply our model to identify SRE pairs related to the SF.

To the best of our knowledge, this is the first time that the regression-based approach is employed to systematically identify cooperative SRE pairs. Moreover, our regression framework can identify SRE pairs without being affected by GC content. Current methods for identifying cooperative SRE pairs in splicing regulation use hypergeometric test to find the co-occurrence of SRE pairs over-presented in different regions [66–68]. Since they only use sequence information, some sequence features that are not related to splicing regulation may increase the FDR. For example, without correction of GC content, the majority of the motif pairs detected in [66, 67] with high p-values share similar GC contents, being either GC-rich or AU-rich. For this reason, they corrected the GC content by grouping the ASEs with similar GC content [66,67]. However, If AU-rich or GC-rich SRE pairs indeed have regulatory effects, correction for GC content might introduce bias and under- or over-estimate the statistical significance of the SRE pairs. In fact, it has been shown that the AU-rich motif is conserved in mammalian introns [64] and could be bound by TIA-1/TIAR, Hu protein or Sam68 [93, 128]. In our work, since our model uses the isoform expression information in addition to the sequence information, it can automatically handle the GC content problem. If there is no correlation between the splicing response and the GC- or AU-rich SRE pairs, these pairs will not be identified since they are just sequence features irrelevant to splicing regulation. On the other hand, if they do have regulatory effect, our method will identify them as SRE pairs.

Since we want to detect all possible SREs and SRE pairs, our linear regression model contains a very large number of candidate SREs and their pairs as variables. The challenging problem in model inference is to select correct variables without overfitting the data. We employed four techniques to tackle this problem. First, variable screening is used to exclude SREs and SRE pairs that are present in less than 1% of the samples or have no or very small correlation with the response variable. Second, regularized inference methods, the Lasso and the adaptive Lasso, were employed in conjunction with cross-validation to select a small number of SREs and SRE pairs. Third, RCV is used to estimate the residual variance which was further used to prevent the possible overfitting problem. Finally, the FDR was calculated to retain only the most statistically significant SREs and SRE pairs in the final model. Overall, these steps combine the state-of-the-art techniques and form an effective framework to reduce the FDR and prevent model overfitting, without compromising the power of detection.

The SREs and SRE pairs we identified have a significant overlap with SREs identified with experiments. The regulatory effects of several well-defined SREs were correctly inferred from our model. For several different interaction proposed based on the experimental results, our model successfully identified them as SRE pairs in the same regions and provided more insight into their interactions. Note that we can identify SRE pairs at two ends of intron [67], at two ends of exon [66], and in the same region [68] in one framework, and capture the combinatorial regulatory effects of multiple SREs more faithfully. We also report AU-rich SRE pairs as a putative interaction pattern that is important and prevalent in human splicing regulation. In summary, our thermodynamic regression model provides a useful platform for discovering splicing regulators and unraveling splicing regulatory mechanisms.

3.7 Concluding Remarks

Although many splicing factors (SFs) and their binding sites have been identified in the past decades, their combinatorial regulatory effects are yet to be illuminated. In this chapter, we developed a linear regression model to integrate combinatorial signals of *cis*-acting splicing regulatory elements (SREs) and their cooperative effects based on the thermodynamics of binding and interactions among SFs, spliceosome and pre-mRNA. We also developed a systematic framework for model inference that can identify SREs and their interactions from a large number of potential SREs without overfitting the model. Applying the thermodynamic model to a human RNA-Seq data set consisting of 9 tissues, we demonstrated that the final selected model can explain 49.1%-66.5% variance of the data, which is comparable to the best result achieved by thermodynamic models for transcription [73]. In total, we identified 119 SRE pairs between different regions of cassette exons that may regulate exon or intron definition in splicing, and 77 SRE pairs from the same region that may arise from a long motif or two different SREs bound by different SFs. Some SRE pairs are consistent with the interaction models that have been proposed based on previous experimental results, and various cooperative mechanisms between SREs could be identified from our results. These results show that our thermodynamic model and inference method provide a means of quantitative modeling of splicing regulation and a useful tool for identifying SREs and their interactions.

CHAPTER 4

Aberrant Isoform Expression in Cancer

4.1 Introduction to Cancer Biology

Cancer is a disease of genome aberration [131]. Somatic mutations play an important role in transforming normal cells into cancerous cells. The major somatic mutations known in the cancer genome include nucleotide substitution mutations, small insertion/deletions (indels), copy number variations, chromosomal rearrangements, and nucleic acids of foreign origin. These somatic mutations can cause either tumor promotion or tumor suppression by changing the expression level of related proteins or functions of related molecules. Therefore, to understand the pathogenesis of cancer, it is very important to identify these somatic mutations from the sequenced cancer genomes, as well as to find aberrantly expressed genes from the measured gene expression levels. The advent of next-generation sequencing technologies provide a revolutionary tool to study cancer genomics. Their applications to cancer studies have accelerated our understanding of tumorigenesis [132, 133].

Over the past decades, many studies have been performed to learn the molecular architecture in human cancers. It has been shown that various layers of this architecture contribute to the pathogenesis of most human malignancies, including variation in gene expression [134], deregulation of miRNAs [135], alternative promoter usage [136], alternative cleavage and polyadenylation [137] and alternative splicing [21]. Alternative promoter usage produces alternative first exons (AFEs) and alternative cleavage and polyadenylation produces alternative last exons (ALEs) which are involved in different mechanisms from alternative splicing. However, due to restrictions in patients' samples and the experimental design, the effects and the importance of these biological process in cancer is yet to be discovered, especially for alternative splicing.

Several previous studies have tried to profile the alternative splicing and to find differential splicing between cancer cells and normal cells using microarray [138–140]. However, the design of microarrays needs a reference annotation of human genome. Cancer cells always include a large number of mutations. If the mutations occur around the splice site or create new splice sites, they will give rise to new splice isoforms, which can not be detected by microarray. Moreover, in experiments that used tumor samples without matched normal samples that are derived from the normal cells of the same patient, splicing polymorphisms can be identified as aberrant splicing events, even though they are irrelevant to cancer pathogenesis.

4.2 Aberrant Alternative Region

According to the exon shuffling theory [141], exons in genes can encode specific functional modules. Duplication, permutation and rearrangement of these exons would brought different functional modules together to produce new proteins. It has also been shown that there is a significant correlation between the borders of exons and protein domains on genomic scale and this exon-domain correlation is consistently stronger in more complex organisms [142]. In human genes, different isoforms usually arise from different combination of a group of building blocks–exons [143]. Thus, the aberrant isoform expression in tumor may only related to loss or gain of some functional modules or pathogenic domains. Inspired by this notion, we proposed a method to identify “alternative regions” aberrantly included or excluded in tumor. The concept of alternative region is a natural generalization of alternative spliced exon, in order to discover aberrant AFE and aberrant ALE in a consistent framework. Thus the alternative regions can arise from alternative promoter usage, alternative polyadenylation and all kinds of alternative splicing events. Accordingly, the problem of differential splicing discovery was generalized to the problem of aberrant alternative region (AAR) discovery.

The proposed framework first use RNA-Seq data to assemble patient-specific transcriptome in tumor and matched normal samples independently. Then alternative region of each gene are identified based on all isoforms in cancer and normal samples of an individual. Statistical test for differential inclusion of these alternative region will generate a list of AAR and prevalence of each AAR in patients are evaluated. The AARs may harbor a pathogenic domain, therefore, if many breast cancer patients share a same AAR, the AAR and corresponding gene are highly probable to be associated with cancer pathogenicity.

The proposed method has several advantages compared to previous efforts. First, statistical test in AAR context is more powerful than inclusion ratio test in isoform context. Several isoforms may share an identical alternative region and thus may share similar functional modules in their encoded proteins. The inclusion ratio difference of the alternative region will be larger than the inclusion ratio difference of each single isoform and thus are easier to be identified. Second, our method can identify differential splicing event as well as AFE and ALE event in one framework. Third, since patient-specific transcriptome are used, new splice junctions and isoforms, either patient-specific or ubiquitous, can be detected in addition to the ones documented in reference annotation. Fourth, since matched normal samples are used for comparison, splicing polymorphism will not be identified as aberrant splicing which reduces the false positive rate. Fifth, compared to microarray data, RNA-Seq data overcome many restrictions originated from experiment design, and thus increase the possibility of new discoveries. The only challenge of this framework is the much more extensive requirement of high performance computing for large size of RNA-Seq data.

4.3 Materials

Several large-scale cancer genome characterization efforts have been initiated by utilizing the next-generation sequencing technologies [144]. For example, The Cancer Genome Atlas (TCGA) project has applied microarray and next-generation sequencing technologies to profile somatic mutations, gene and microRNA expressions, DNA methylation and single-nucleotide polymorphism (SNP) in a large number of stringently qualified tumor samples [26, 145]. Most data generated in TCGA are made publicly available although the raw sequence and clinical data are subjected to controlled-access restriction for the protection of patient privacy.

In this chapter, since we focus on the aberrant isoform expression, especially alterative splicing in cancer, the RNA-Seq data are used for analyzing the gene expression in isoform level. In total, we have used 210 clinical breast cancer RNA-Seq data for tumor and matched normal samples in 105 breast cancer patients downloaded from TCGA data portal (Appendix table C.1). The matched normal sample is derived from the same organ site as the tumor from the same patient. The file size of the raw RNA-Seq data for each sample ranges from 3.2 GB to 12.9 GB, with a median file size of 6.3 GB.

4.4 Analysis Workflow

The proposed method for identification of aberrant isoform expression can be separated into several steps, as illustrated in Figure 4.1. Specifically, the first step is mapping all the RNA-Seq reads to human reference genome. The second step is to assemble the patient-specific transcriptome based on both the reads that can be mapped to the genome continuously or to splice junctions. The third step is traditionally differential gene expression analysis. The forth step is the novel analysis for differential inclusion of alternative regions.

4.4.1 RNA-Seq Mapping

Since we need to reassemble the patient-specific transcriptome, the reads that can be mapped to the human genome discontinuously, which arise from mRNA exon-exon junctions (also named junction read), are of particular interest. To this end, the paired 75-bp RNA-Seq reads were mapped to human reference genome assembly (NCBI release GRCh37, UCSC release hg19) using tophat v2.0.4 [40,146], a fast splice junction mapper for RNA-Seq reads. The RNA-Seq reads for tumor and normal samples for each individual were mapped to the reference genome separately.

4.4.2 Patient-Specific Transcriptome Assembly

Cancer cells undergo extensive genomic changes and thus the transcriptome may be different in different patients. Novel isoforms can exist in a patient-specific manner.

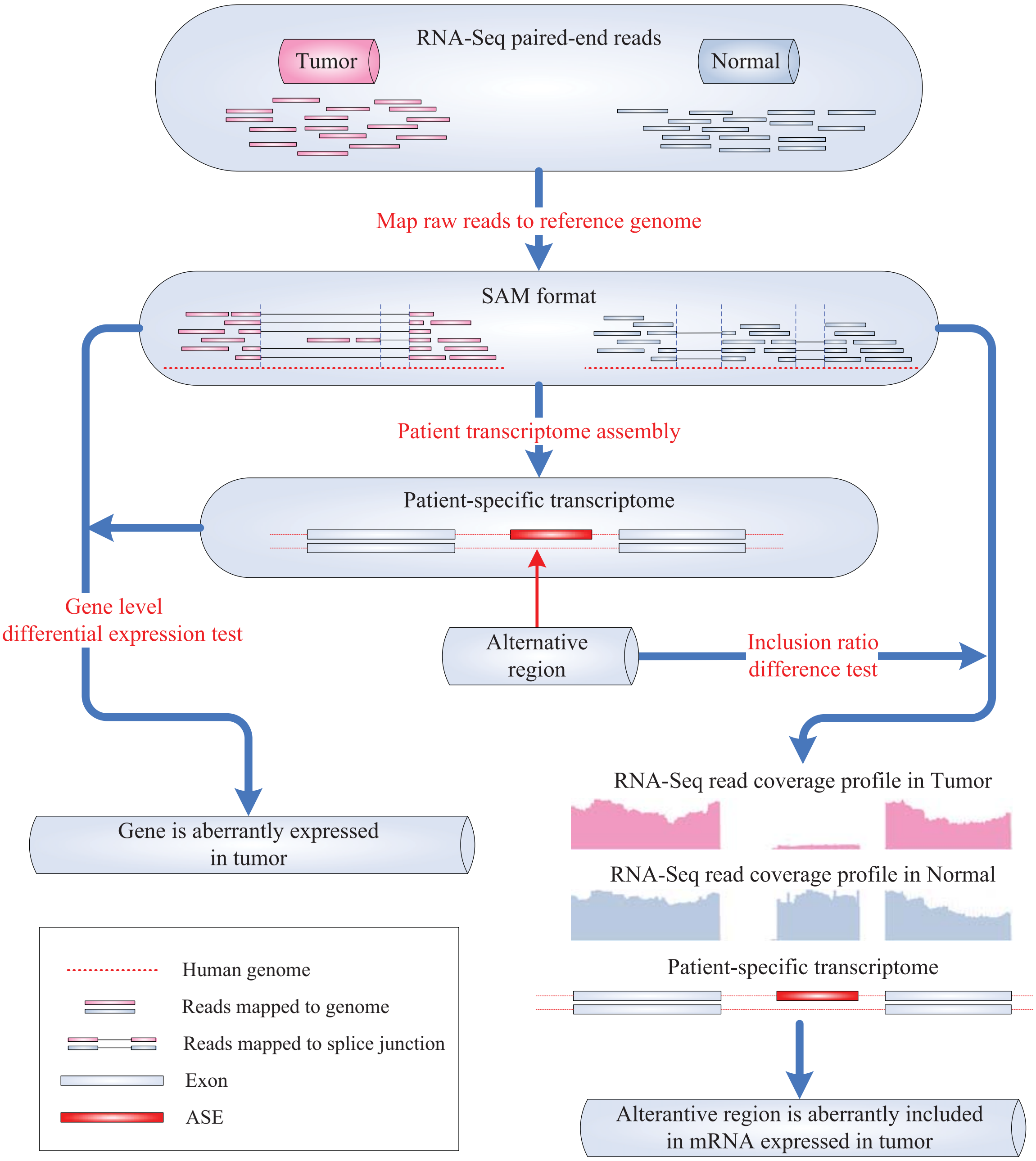

RNA-Seq paired-end reads
Tumor
Normal
Map raw reads to reference genome
SAM format
Patient transcriptome assembly
Patient-specific transcriptome
Gene level
differential expression test
Alternative
region
Inclusion ratio
difference test
Gene is aberrantly expressed
in tumor
RNA-Seq read coverage profile in Tumor
RNA-Seq read coverage profile in Normal
Patient-specific transcriptome
Alterantive region is aberrantly included
in mRNA expressed in tumor
Human genome
Reads mapped to genome
Reads mapped to splice junction
Exon
ASE

Figure 4.1: Workflow for identification of aberrantly expressed alternative region in human breast cancer.

We first assembled the transcriptome for each tumor and normal sample separately using cufflinks v2.0.2 [147] with RABT option (reference annotation based transcript assembly) [148]. UCSC human Known Genes annotations [78] were used as the reference annotation. To test differential gene expression and differential alternative region inclusion ratio for each pair of tumor and normal samples, the tumor and normal transcriptome were then merged for each patients following the strategy recommended by cufflinks [149].

4.4.3 Differential Gene Expression in Cancer

Properly normalized RNA-Seq fragment counts can be used as a measure of relative abundances or expression levels of assembled genes or isoforms. We used cuffdiff in the cufflinks package [147] to estimate gene and isoform expression levels in unit of FPKM (Fragments Per Kilobase of exon per Million fragments mapped) for each pair of tumor and normal samples. Since different RNA-Seq experiments can have different sequencing depth, the expression levels of 210 samples were normalized by geometric normalization method [150], which has been shown to perform much better than the other normalization methods [151]. The log2 transformed gene expression ratio between 105 tumor and matched normal samples were then tested using SAM v4.00 [152] to identify differentially expressed genes.

4.4.4 Alternative Region Identification

The alternative region was defined in this work as the mRNA sequence that are not included in all isoforms of the gene. This is in contrast to constitutive region which is defined as the region shared by all the isoforms. For example, in Figure 4.2, the hypothetical gene can express 4 different isoforms by including different blocks into the mRNA. Thus, we defined 5 alternative regions (AR) for this gene. The 1st and 2nd AR arise from alternative promoter usage and the 3rd-5th ARs arise from alternative splicing. The 3rd AR is due to alternative splicing of an cassette exon. The 4th AR is due to alternative 5' splice site usage. The 5th AR is due to an intron retention. For each AR, we defined the isoforms including the AR as set $I_{incl.}$ while the isoforms excluding the AR as set $I_{excl.}$. The inclusion ratio of the AR was defined as $\sum_{i \in I_{incl.}} E_i / \sum_{i \in (I_{incl.} \cup I_{excl.})} E_i$, where E_i is the expression level of isoform i. Take AR3 in Figure 4.2 for example, the inclusion ratio the AR3 is equal to the sum of the expression of isoform 1 and 3 divided by sum of all four isoforms' expression.

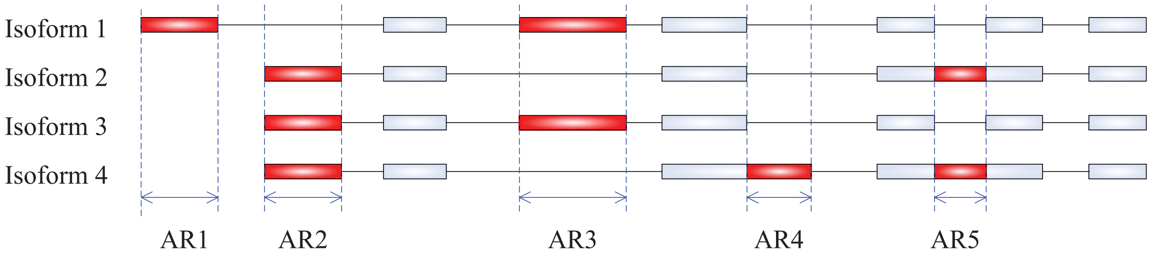


Figure 4.2: Illustration of alternative regions. Four different isoforms were plotted along the genomic coordinates. Blocks stand for exons, black lines indicate introns. Five alternative regions were defined for this hypothetical gene.

4.4.5 Differential Inclusion of Alternative Regions

The alternative regions may have important consequences in cancer development and thus may serve as markers in cancer patients. Thus, it is very important to identify such regions. Currently, there are two different kinds of approaches for this purpose. The first one is the Fisher exact test of isoform-specific read counts [153–155]. The second one is the Bayesian approach that models read counts as a sampling process from a mixture of distinct isoforms [156, 157]. The Fisher exact test utilized the count number of reads mapped to alternative regions and of reads mapped to splice junctions in two conditions to construct a contingency table, and test the differential AR usage [155], and thus, only local information is used in the differential test. Moreover, since the splice junctions, which are the only distinctive regions for isoforms without the AR, are usually very short compared to exon regions, the statistical test may not be stable. On the contrary, Bayesian approach can utilized all the reads mapped to the entire gene region to estimate the inclusion ratio in a probabilistic framework. Although possible assembly errors in very complicated genes may affect the inclusion ratio estimate of each single isoform, the impact will be mitigated since the test target of Bayesian approach is AR, not specific isoform.

To detect genes with these pathogenic regions, we used a program named MISO v0.4.3 [156] to test the inclusion ratio difference of alternative regions in 105 patients. MISO models the generative process of RNA-Seq reads from different isoforms and uses Markov Chain Monte Carlo Sampling to estimate the distribution of the expression level of each isoform relative to the total gene expression level of the gene. After the sampling process, the relative expression level Ψ_i of isoform i (ranging from 0 to 1) is estimated from its distribution that is determined by a group of posterior sampling values. Since we are interested in the inclusion ratio difference of a particular AR, the distribution of the inclusion ratio of one AR is estimated from $\hat{\Psi}_{AR} = \sum_{i \in I_{incl.}} \hat{\Psi}_i$, where $\hat{\Psi}_i$ is the sampling values of Ψ_i. To detect differentially expressed AR, the difference of $\hat{\Psi}_{AR}$ in tumor and normal data, $\triangle\hat{\Psi}_{AR} = \hat{\Psi}_{AR,T} - \hat{\Psi}_{AR,N}$, can be evaluated statistically using the Bayes factor (BF), where $\hat{\Psi}_{AR,T}$ and $\hat{\Psi}_{AR,N}$ are the sampling values of Ψ_{AR} in tumor and normal data. The BF, as a measurement of test significance, is defined as the ratio of the posterior probability of the alternative hypothesis, $\triangle\Psi_{AR} \neq 0$, to the null hypothesis, $\triangle\Psi_{AR} = 0$ and calculated using MISO "–compare-samples" option [156].

The differentially included ARs were then selected at certain significance level based on the following criteria. First, the absolute value of inclusion ratio difference $\triangle\hat{\Psi}_{AR}$ must exceed 10%, which is motivated by the fact that change in exon inclusion level of $\sim$ 10% can have important consequences in diseases [140]. Second, the BF must exceed 20, since experimental validation by qRT-PCR has shown that $\sim$ 100% of exons with BF$\geq$20 were detected as differentially expressed and show good agreement on the magnitude of $\triangle\hat{\Psi}_{AR}$, compared to 21% of exons with BF<20 [156]. ARs satisfying both criteria were selected in the following analysis and named as aberrant alternative regions (AARs).

The more individuals possess the AAR, the more likely the AAR to be associated with cancer development. In the next step, we integrated all the AARs detected in 105 individuals, selected one most aberrant AR for each gene, and sorted the selected genes according to their occurrence frequency. In total, we have identified 5088 genes with its AAR identified in at leat 10 patients (see Appendix table C.2 for details).

4.5 Results and Discussions

To systematically examine the quality of our AARs and choose reliable candidate for further analysis, we ranked all the AAR by their occurring frequency in 105 patients. The top 25 AARs and the relevant experimental evidence reported in the literature are listed in Table 4.1. Column "genomic coordinate" is the AAR's location in human reference genome; Column "# diff." is the number of patients where the AAR was identified from; Column "ave. diff." is the average inclusion ratio difference in the identified patients; The last column is the name of the gene where the AAR resides in. These AARs were found aberrantly included with high significance in at leat 75% (79/105) breast cancer patients. In the following sections, we examined 4 top AARs and analyzed their possible pathogenic mechanisms by comparing related evidences and studies in the literature. The consistency with experimental evidence and interesting findings show that our method are highly effective to identify important genes related to cancer development and provide important insight in cancer pathogenicity.

Table 4.1: 25 most frequent AARs in 105 breast cancer patients

genomic coordinate	# diff.	ave. diff.	RefGene
chr14:69345175-69345240	103	-0.30	ACTN1
chr3:13663275-13663415	102	-0.49	FBLN2
chr9:124043748-124043840	96	0.53	GSN
chr15:74466087-74466360	92	-0.39	ISLR
chr2:174123427-174123543	92	0.25	ZAK
chr10:93000241-93000337	91	-0.30	PCGF5
chr5:33751303-33751508	91	-0.40	ADAMTS12
chr9:123631853-123632122	90	0.30	PHF19
chr1:207963598-207963690	88	-0.26	CD46
chr6:56507420-56507694	88	-0.36	DST
chr2:238678586-238678635	87	-0.28	LRRFIP1
chr9:117808689-117808961	86	0.38	TNC
chr3:37132958-37133029	84	-0.37	LRRFIP2
chr3:57911572-57911661	84	-0.28	SLMAP
chr14:73745989-73746132	84	0.33	NUMB
chr12:56558153-56558431	84	0.29	SMARCC2
chr2:64069014-64069338	82	-0.28	UGP2
chr5:38445578-38445780	82	-0.37	EGFLAM
chr19:49605431-49605442	81	-0.24	SNRNP70
chr15:89422649-89423841	81	-0.23	HAPLN3
chr15:64429767-64430240	81	-0.33	SNX1
chrX:102942917-102943086	81	-0.23	MORF4L2
chr17:48828003-48828055	81	0.28	LUC7L3
chr8:95470496-95470664	80	0.41	RAD54B
chr11:131240371-131240783	79	-0.49	NTM

4.5.1 ACTN1

ACTN1 codes for homo sapiens alpha actinins, an actin-binding protein with multiple roles in different cell types. Figure 4.3 depicts the RNA-Seq read coverage profile in 5 patients with largest inclusion ratio difference. The AAR is an alternatively spliced exon included in three isoforms. It can be seen from Figure 4.3 that the isoforms with the AAR are almost not expressed in the 5 Tumor samples, but the inclusion ratio of this AAR are around 0.5 in 5 Normal samples.

Our result indicated that the inclusion ratio difference of this AAR is greater than 0.1 in 103 patients, which means that skipping of this ASE is prevalent in breast cancer. Moreover, the reduced inclusion of the same exon has also been identified in colon, bladder, and prostate cancer by exon arrays [158]. These results indicate a general loss of this exon in different cancers. Further examination of expressed isoforms indicates that the AAR and upstream alternative exons are generally expressed in a mutual exclusive manner, which implies that the actual expression level of the isoform with both exons included is near zero (Figure 4.4).

To study the potential function of this AAR in ACTN1, we further analyzed the conserved functional domain near the AAR. As shown in Figure 4.4 and documented in the UniProt protein database [159], ACTN1 contains two calcium-binding domains known as the EF-hand domain, which is coded by the upstream and downstream exons of the AAR. In most currently known proteins, the EF-hand domain always occur in adjacent pairs, and the pairing results in cooperativity as a functional consequence of Ca^{2+} binding [160]. Moreover, disruption of each of the two EF-hand doamins in Dictyostelium ortholog of ACTN1 has distinguishable consequences on its regulatory activities [161]. Our result indicates that most isoforms of ACTN1 expressed in breast cancer cells contain both two EF-hands, and thus have higher binding affinities to

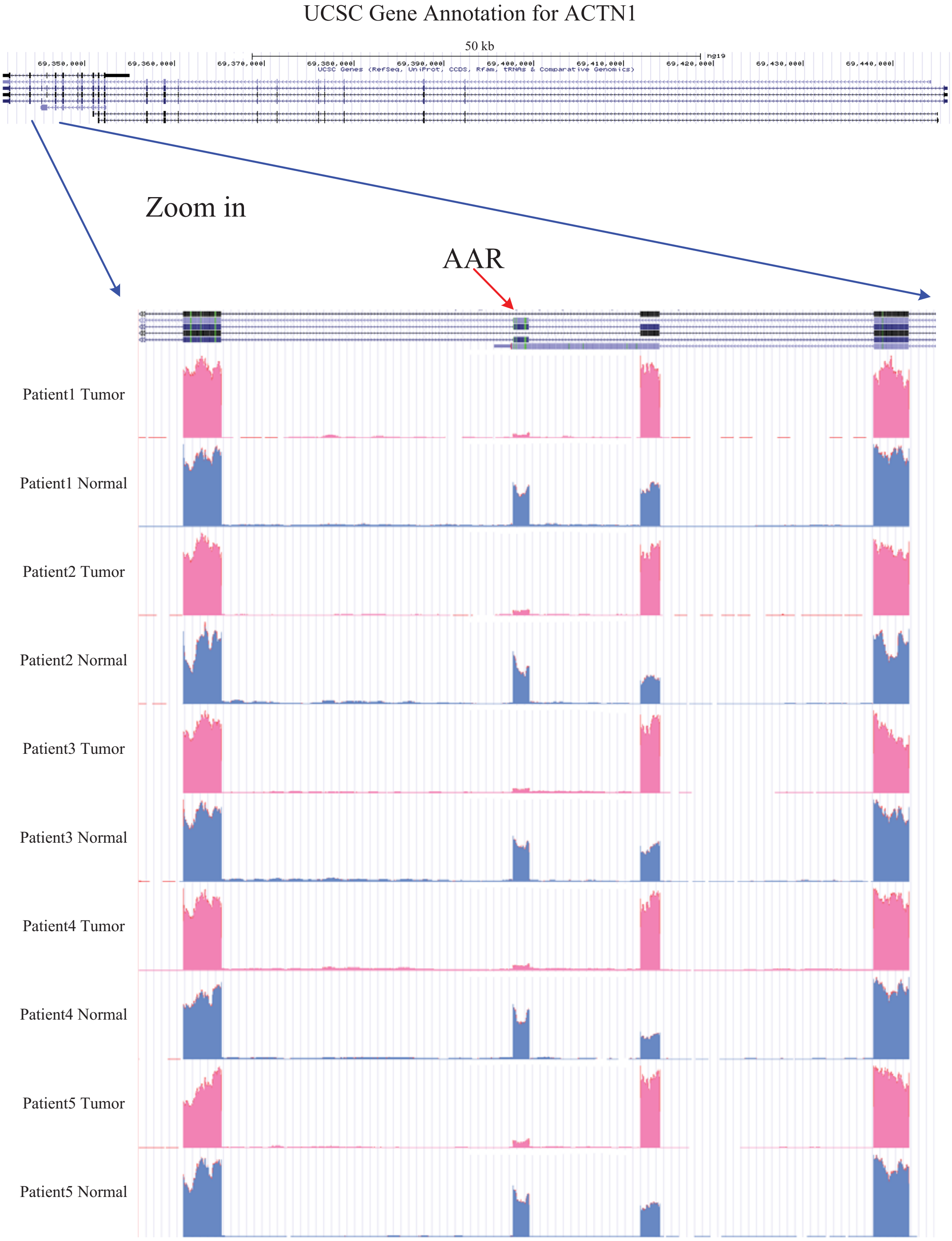


Figure 4.3: RNA-Seq read coverage profile for ACTN1 in 5 patients. UCSC gene annotation was visualized using UCSC genome browser. Blocks indicates exons. The AAR is the exon indicated by the arrow.
]

Ca^{2+}, while in normal cells, only about half of the ACTN1 isoforms contain both EF-hand. Since Ca^{2+} is a crucial regulator of cell migration and tumor metastasis [162], this AAR may modulate ACTN1's functions in metastasis by changing its binding sensitivity.

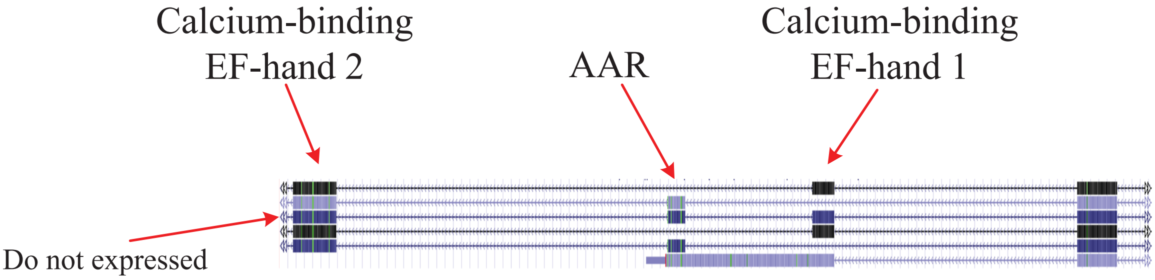


Figure 4.4: Two EF-hand domains near the AAR in ACTN1. Since the gene is on the reverse stand, the upstream (downstream) exon is located at the right-hand (left-hand) side of the AAR.

4.5.2 FBLN2

FBLN2 is calcium dependent, extracellular matrix protein. Our results indicate that only two isoforms are expressed in both tumor and normal cells. However, the composition of the two isoforms are reversed in tumor compared to normal tissues (Figure 4.5B). This aberrant region was identified in 102 breast cancer patients.

Domain analysis indicate that FBLN2 contains a tandem of epidermal growth factor-like (EGF) repeat, most of which have a calcium-binding domain (Figure 4.5A). The AAR codes for one Calcium-binding EGF-like domain (EGF_CA), upstream and downstream exons of the AAR also code for one EGF and one EGF_CA domain. EGF domains in tandem can interact with each other to affect Ca^{2+} binding affinity [163]. In nasopharyngeal carcinoma, it has been shown that the short FBLN2 isoform without AAR (FBLN2S) may be a candidate tumor-suppressor that can inhibit cell proliferation, migration, invasion and angiogenesis and down-regulated in tumor cells, while the long isoform with AAR (FBLN2L) is either not detectable or is expressed only at low levels in both normal and tumor tissues [164]. This is on contrary to our breast cancer result, where FBLN2S is up-regulated and FBLN2L is down-regulated in breast cancer cells. It may indicates that a different function or pathogenic mechanism of FBLN2 in breast cancer.

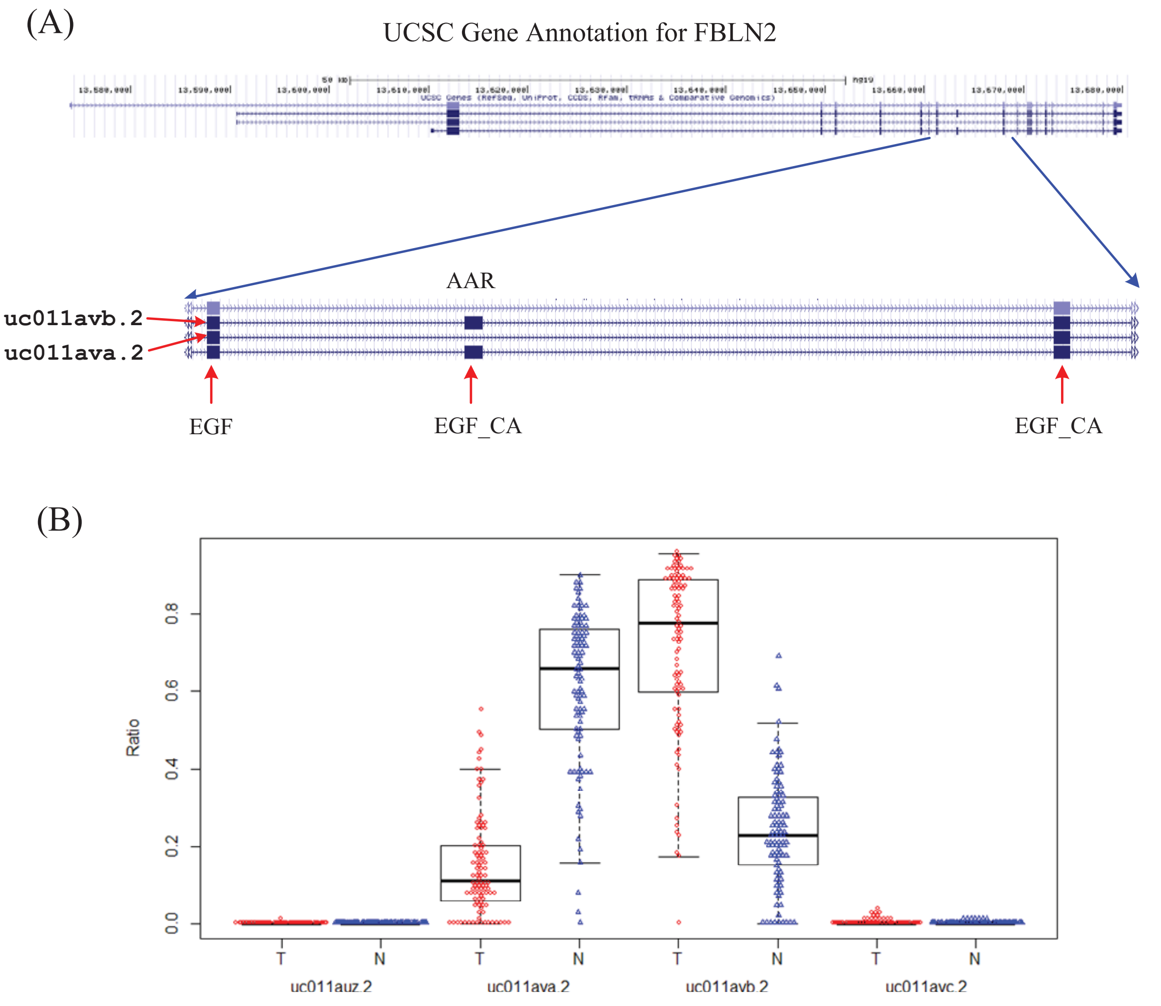


Figure 4.5: (A) Domain analysis near the AAR of FBLN2. (B) Box and scatter plots for the inclusion ratio of each isoform in tumor and normal samples. The inclusion ratio is calculated as the expression ratio between the isoform and the whole gene. Note that only two isoforms are actually expressed.

4.5.3 ISLR

ISLR (immunoglobulin superfamily containing leucine-rich repeat) gene has alternative promoters in human genome. Aberrant usage of the promoters were identified in our results (Figure 4.6). Note that the alternative first exon do not codes for amino acids, thus the protein sequences of the two isoforms are identical.

Aberrant usage of promoters has been found to be related to various diseases, including cancer [136]. It has been found that the protein level of ISLR released into the extracellular media was down-regulated in a malignant breast-cancer cell line compared to its matched normal cell line. [165]. However, our result indicates that the overall gene expression in mRNA level do not show difference in tumor and normal samples. Thus, it is highly possible that the downstream promoter affects the stability or translation efficiency of the mRNA variants, and thus down-regulates the protein level of ISLR, which in turn regulates the tumor cell's immune system response.

4.5.4 ZAK

The AAR for ZAK is alternative last exons. There are two kind of isoforms expressed from this gene (Figure 4.7), one with alternative last several exons (ZAK-α), the other one without the last several exons (ZAK-β). The encoded proteins also have distinct C-terminus. Although both isoforms have similar functions in activation of the ERK, JNK/SAPK, p38, and ERK5 pathways, ZAK-α, but not ZAK-β, in Swiss 3T3 cells can cause disruption of actin stress fibers and dramatic morphological changes [166]. Moreover, over-expression of ZAK-α can induce neoplastic cell transformation and tumorigenesis in athymic nude mice [167]. This is consistent with our result, where inclusion ratio of ZAK-α was significantly up-regulated (uc002uhz.3 in

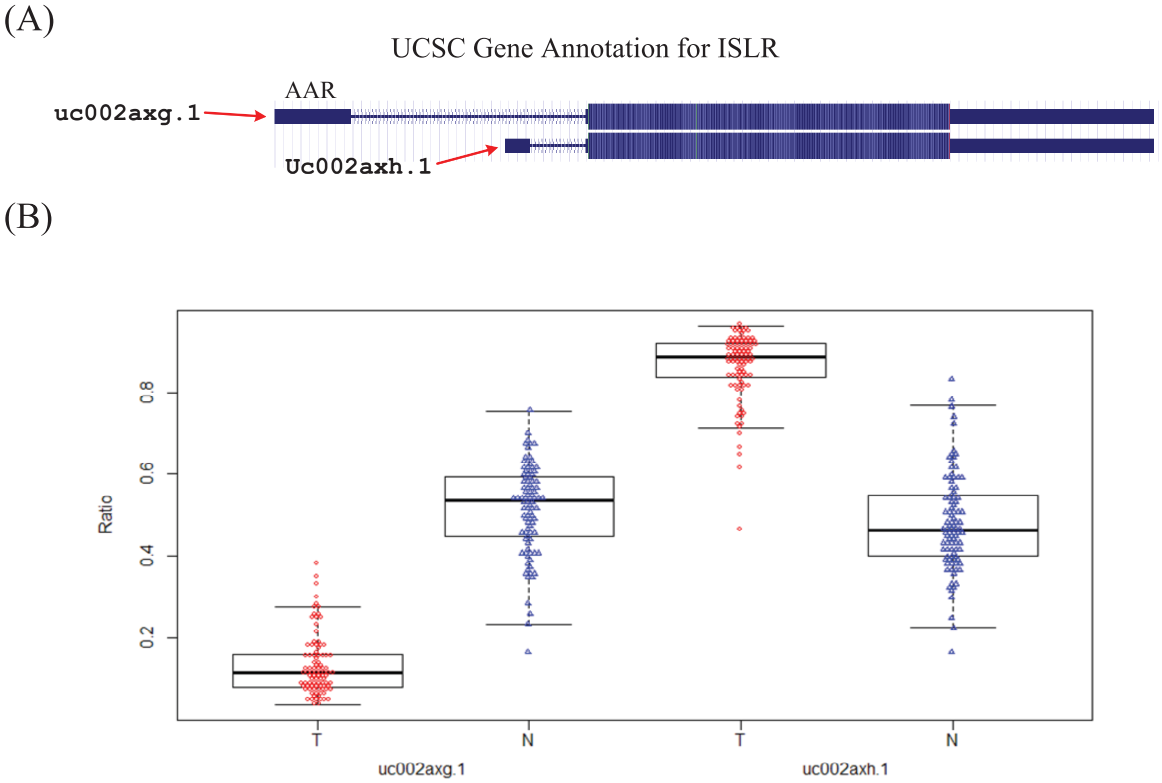


Figure 4.6: (A) Domain analysis near the AAR of ISLR. (B) The inclusion ratio of each isoform in tumor and normal samples.

Figure 4.7B) in tumor, even though the overall gene expression level is down-regulated (TotalG Figure 4.7C) in breast cancer cells. However, when comparing the log2 expression ratio of ZAK-α in tumor and normal samples (uc002uhz.3 in Figure 4.7C), there is no obvious evidence of over-expression. One possible explanation is that it is the relative proportion of ZAK-α and ZAK-β, rather than the absolute difference of ZAK-α, to influence cell transformation and cancer development.

4.5.5 Comparison with results using microarray profiling

Before the advent of RNA-Seq, Microarray is the major tool for large-scale genomic study. However, it is not the best platform for splicing analysis since it is confined by the experimental design and detection sensitivity. We compared our result with two microarray based analysis in breast cancer to show the higher sensitivity of RNA-Seq data and our method.

Misquitta-Ali *et al.* [140] surveyed 5183 alternative exons in lung and breast cancers, using patient-matched normal as controls. Four genes (VEGFA, MACF1, APP, and NUMB genes) were identified aberrantly spliced in at least 5 lung cancer patients (50% of profiled patients) with at least 10% inclusion ratio difference, and similar aberrant splicing in NUMB and APP were also identified in breast cancers. The function of the most frequent aberrant gene NUMB was further analyzed to show that the inclusion of alternative region is capable of promoting cell proliferation. The exact same exon in NUMB was also identified as an AAR in our RNA-Seq result with same inclusion ratio change direction. In our dataset, 84 out of 105 (80%) patients have the NUMB AAR whose inclusion ratio is at least 10% higher in breast cancer samples than that in matched normal samples. This AAR was ranked as the 15th most aberrant regions in our AAR list. Gene VEGFA, MACF1, APP were also iden-

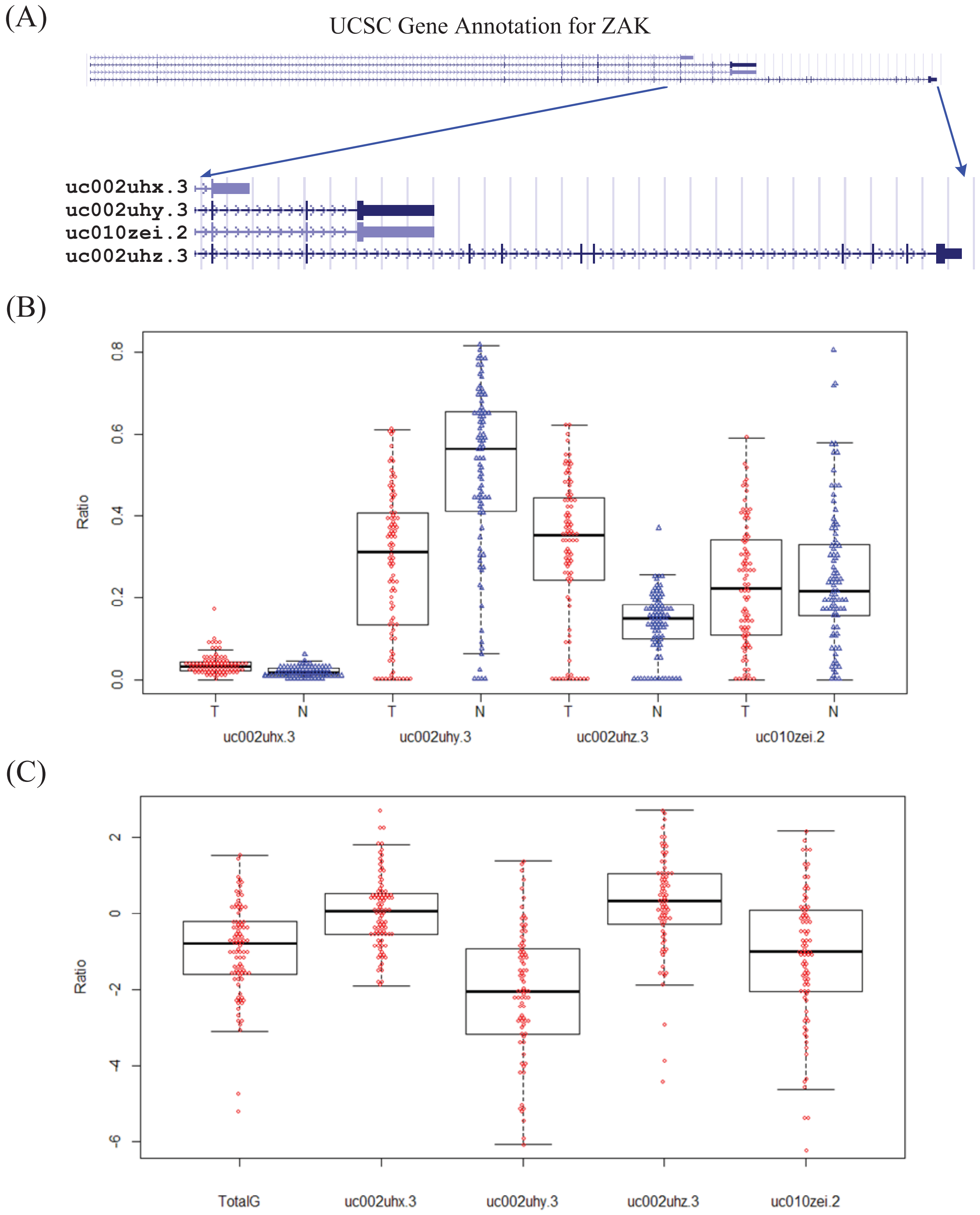


Figure 4.7: (A) Gene structure of ZAK. (B) The inclusion ratio of each isoform in tumor and normal samples. (C) The expression ratio of each isoform in tumor and normal samples. The ratio is calculated as the expression level in tumor divided by the expression level in its matched normal samples. “TotalG” indicates the overall gene expression ratio.

tified aberrantly spliced in 41, 49, 71 patients with same alternative regions except MACF1, where a different alternative region was identified as the most aberrant one for this gene. Since Misquitta-Ali *et al.*'s work used different patients's data and micro-array analysis, the consistency with their results indicate the robustness of our AAR based method. In addition, since we have used RNA-Seq data of more patients, our results have higher sensitivity and power. Thus the best hit in Misquitta-Ali *et al.*'s work (NUMB), although identified with high frequency, only ranked 15 in our result.

Venables *et al.* [168] studied the alternative splicing profiles of 600 cancer-associated genes in 21 normal and 26 cancerous breast tissues and identified 41 ASEs in 40 genes that significantly differed in breast tumors relative to normal breast tissues. Among the 40 genes, 30 were also identified in our 5088 gene list. The significant overlap between these two results (fisher exact test p-value=1.3e-17) indicates the validity of our result. Note that the tumor and normal samples used in this study are not always from the same patient, and thus pairwise test are not applicable. The result may be affected by high gene expression variation in tumor and polymorphism among individuals. None of the 30 genes are ranked top 25 in our gene list 4.1.

4.5.6 Functional Enrichment Analysis

We have analyzed top AARs identified with our methods and found that they always have specific cellular functions or are related to specific pathways. To systematically evaluate general functions of all the identified AARs, we applied functional enrichment analysis using the Gene Ontology (GO) database. The GO project [169] is a collaborative effort to describe gene product attributes across species and databases. GO consists of three hierarchically structured annotations that describe gene products in terms of their associated biological processes, cellular components and molecular functions and can be represented by a directed acyclic graph.

The functional enrichment analysis was perform using GOrilla [170] with default parameters, a web-based application that identifies enriched GO (Gene Ontology) terms. Specifically, we used our 5088 genes as target gene set, and all the other genes as the background gene set. GOrilla calculated the significance of the overlap between the target gene set and the gene members in each GO term by hypergeometric test. The output of GOrilla consists of a hierarchical tree graph to indicate the relationship between each enriched GO term (p-value<1e-3). Since nodes of the tree near the root level, such as cellular process, are a very broad biological concept, and contain little information about cancer development, we extracted all the leaf nodes, and sorted them based on their p-values. The top 10 GO terms are listed in Table 4.2.

The GO enrichment analysis implies that AAR occurs frequently in several important pathways. The nerve growth factor (NGF) is known to play a major role in cancer development and metastasis [171] and has been proposed as a potential therapeutic target in breast cancer recently [172]. Its signalling pathway has been shown to regulate cell proliferation in rat chromaffin cells [173]. Axon guidance molecules control neuronal migration and neuronal survival, and their expression are not confined to the nervous system. Recently, several studies have suggested that they regulate key pathways involved in cell proliferation and migration [174], and are gaining more and more attention in many pathological processes, particularly in cancers [175–177]. Epidermal growth factor receptor (EGFR), as the first receptor targeted for cancer therapy [178], its signalling pathway has been studied extensively in cancer biology [179, 180]. These results indicate that mutations or gene expression changes in these pathways are associated with cancer development, however, to the best of our knowledge, this is the first time that these pathways are also found enriched with aberrant isoform level expression.

Table 4.2: Top 10 enriched biological process

GO term	Description	p-value
GO:0048011	nerve growth factor receptor signaling pathway	2.02E-17
GO:0007411	axon guidance	2.99E-15
GO:0007173	epidermal growth factor receptor signaling pathway	1.40E-12
GO:0035023	regulation of Rho protein signal transduction	8.79E-10
GO:0051017	actin filament bundle assembly	1.46E-09
GO:0006915	apoptotic process	9.49E-09
GO:0007507	heart development	1.22E-08
GO:0006281	DNA repair	3.31E-08
GO:0019048	virus-host interaction	3.79E-08
GO:0042787	protein ubiquitination involved in ubiquitin-dependent protein catabolic process	4.85E-08

4.6 Concluding Remarks

Alternative splicing, alternative promoter usage and alternative polyadenylation have been shown important to cancer development. However, due to restrictions in experimental platform, only aberrant gene expression is extensively studied. Recent advances in sequencing technology have provided new opportunity for cancer research and therapies. In this chapter, we applied state-of-the-art algorithms to the RNA-Seq data of 105 breast cancer patients. We have identified several important genes that are highly possible to be correlated to breast cancer development with strong experimental evidence. Special attention has not been paid to most of the identified AARs in breast cancer. For the first time, we also reported that several cancer related pathway are significantly enriched with AARs. These exciting results will provide more insight to cancer biology and new directions for cancer therapies.

CHAPTER 5

Summary and Future Work

5.1 Summary

Alternative splicing (AS) of precursor mRNA (pre-mRNA) is a crucial step in the expression of most eukaryotic genes. It provides an important means of regulating gene expression and generating transcriptomic and proteomic diversity. Disruption of AS regulation can lead to diseases such as cancer. A key mechanism of AS regulation is to influence the spliceosome to recognize splice sites via binding or unbinding of splicing factors to SREs. It is therefore important to identify these SREs and their combinatorial effects on regulating AS. Toward this end, we developed two different methods from different points of view in Chapter 2 and 3 to search for SREs. In Chapter 2, we employed a traditional enrichment-based approach but incorporated expression data into the search process and used a discriminative method that compared a positive data set with a more reliable negative data set to increase the detection power. We also did some in-depth analysis on the position bias of the identified SREs to reveal their regulatory mechanisms. However, the results found in Chapter 2 did not uncover the mechanisms of tissue-specific alternative splicing. To understand tissue-specific alternative splicing, we employed principles of thermodynamics to derive a theoretical model in Chapter 3. In this study, we included both combinatorial regulation of different SREs and interactions between two SREs to model the splicing system. We also developed a framework to identify the regulatory SREs and SRE pairs, whose performance was comparable to the best nonlinear model for transcription. We demonstrated that our model could reveal various cooperative mechanisms between SREs for AS regulation. In Chapter 4, we explored the alternative splicing in human invasive breast cancer. In order to study the alternative first exon (AFE) and alternative last exon (ALE) at the same time, we generalized the aberrant alternative splicing problem into an aberrant alternative region (AAR) problem. The proposed AAR concept is helpful in designing a systematic and robust framework or workflow to study alternative splicing. The identified aberrant regions not only provided good candidates for cancer therapy, but also provided more insight into the pathological mechanisms.

There are several major differences between the methods and results in Chapter 2 and 3. First, the thermodynamics-based method used in Chapter 3 integrated all the candidate SREs into the multiple regression model, thus it could study the combinatorial effects of the SREs in one framework, whereas the enrichment-based method used in Chapter 2 tested the effect of each SRE separately. Second, the interpretation of the results are different. In Chapter 2, we only considered SREs that are effective (enhancer or silencer) in one tissue. If an SRE is effective in one tissue but not in another tissue, it is recognized as a tissue-specific SRE. In Chapter 3, we explained the tissue specific splicing more explicitly. It is the tissue-specific expression of the SF, not the SRE that causes tissue-specific splicing or differential splicing. Without the expression level of the SF, we can only conclude that an SRE is responsible for different splicing levels in different tissues, but we do not know if it is an enhancer of silencer.

There are also relationships between these two methods. In the thermodynamics model used in Chapter 3, we identified SREs by testing the correlations between the presence of SREs and the splicing changes in different tissues. This is based on the model that tissue-specific expression of the SFs is in essence the reason of differential splicing, rather than SRE. For example, if an enhancing SF is expressed in tissue 1 but not expressed in tissue 2, every gene transcribed in tissue 1 with the SF's binding site (SRE) will tend to have higher inclusion ratio, but this tendency should not exist in tissue 2 since the SRE do not work in tissue 2. From another point of view, the presence of the SRE tend to correlate to the higher splicing level in tissue 1, whereas the presence of the SRE is irrelevant to the splicing result in tissue 2. The discriminative method used in Chapter 2 is designed for capturing such correlation and identify this SRE as an enhancer in tissue 1 and will not identify it as an SRE in tissue 2.

In chapter 4, we focused on the differential region usage (generalization of differential splicing) and their consequences in cancer development. Although splicing regulation is not the emphasis of chapter 4, it is natural to integrate the efforts in 2 and 3 into cancer biology to study the aberrant splicing regulation in cancer. Combining with the regulation information, we can gain more insight into the complexity of cancer development.

5.2 Future Work

5.2.1 Identification of Cancer Driver Genes

Cancer is a disease of genome aberration [131]. Although cancers arise as a result of alteration in DNA sequences, not all the somatic mutations present in a cancer genome are involved in the development of the cancer [131]. Therefore, the somatic mutations can be classed into two categories, namely driver and passenger mutations [181]. A driver mutation can cause cancer or confer growth advantage on the cancer cells, whereas a passenger mutation refers to a mutation in the cancer genome but without obvious growth advantage to the cancerous cells. The driver mutations reside, by definition, in the subset of genes known as cancer or driver genes. A central objective of cancer genome analysis is to distinguish driver mutations from passenger mutations and to identify driver genes.

Two strategies have been developed to search for driver genes. The first enrichment-based strategy assumes that driver mutations are enriched in driver genes whereas passenger mutations are more or less randomly distributed. This strategy has been successfully applied in the past to identify most driver genes that have altered protein sequence in cancer. For example, Greenman *et. al.* [182] identified driver genes as those for which the ratio of non-synonymous vs synonymous mutations is significantly higher than expected, where a non-synonymous mutation alters the amino acid sequence of a protein and a synonymous mutation does not change the amino acid. Several methods have been proposed to estimate the background silent mutation rate. Youn *et. al.* [183] reviewed these methods and developed a new method to identify driver genes by improving the accuracy of the estimation of the background mutation rate.

The second strategy combined gene expression profiles and somatic copy number alterations (CNA) information of the same patient to identify driver genes. The somatic copy number alterations can promote tumorigenesis without changing the protein sequence, but by increasing the expression level of oncogenes or decreasing the expression level of tumor suppressor genes. For example, CONEXIC [184] integrates tumor gene expression and copy number data into a single framework to identify likely drivers in cancer. For a recent review of this strategy, see [185].

It has been indicated that aberrant splicing contributes to all aspects of tumor biology [21, 186]. Driver mutations can occur near splice sites and it has been shown that the selection pressure for nonsense mutations (change an amino acid to a stop codon) and mutations around splice sites are much stronger than that for missense mutations (change an amino acid to another amino acid) in breast cancer [187]. Moreover, synonymous mutations or intronic mutations not considered in previous algorithms for the detection of driver mutations may be also pathogenic, because they can disrupt splicing enhancers or silencers [188, 189]. Therefore, some driver mutations are pathogenic since they alter the isoform expression patterns in cancer cells, rather than change amino acids or overall gene expression levels. However, currently there is no systematic computational method to search for this kind of driver genes in large scale.

The introduction of the concept of aberrant alternative region and frequently identified AAR in breast cancer make the exploration of driver genes in AAR context possible. Although we have identified a group of genes with AAR in chapter 4, we do not know what is the cause of the AAR. According to the models in Chapter 3, it can either result from mutations of the cis-regulatory elements on the pre-mRNA, or from the change of the expression levels or amino acid mutations of trans-acting splicing factors. In the former case, we can easily identify putative driver genes in AAR context by integrating the mutation information (considering both exon and intron) around the identified AAR.

5.2.2 AAR regulation

There are two kinds of possible regulatory networks in cancer under AAR context. First, aberrant inclusion of one alternative region may be a key regulatory step in biological pathways. In this case, the existence of a small subset of AARs may be used to explain a large group of aberrantly expressed genes, and they are good candidates for cancer therapy.

On the other hand, several AARs may be regulated by a single regulator (such as splicing factor or transcription factor). Aberrant expression or mutation of the regulator may cause a comprehensive aberrant inclusion of the downstream genes it regulates. In this case, the different inclusion ratio of a large set of AARs may be explained by only a small subset of regulators, and these regulators should be better targets for cancer therapy than individual AARs. Correlation or regression analysis by incorporating *cis*-acting binding sites will provide more exiting insights to the regulatory network in cancer.

5.2.3 Tumor Classification

Tumor classification methods that distinguish tumor and normal tissues or different tumor subtypes are important in cancer diagnosis, and are usually designed to help identify patient groups that may have similar prognosis or response to treatment. Many works in tumor classification used microarry as the gene expression profiling tool and classified tumor samples by clustering genes based on their expression levels [190, 191].

Although the change of gene expression levels change is an important signature in cancer, a larger study in prostate cancer found that 30% of the genes studied on the microarray show significant differential splicing in the absence of detectable changes in overall gene expression [192]. The difference of splicing ratio may be also a good signature for classification. Therefore, we can develop a systematic method to combine the gene expression and splicing profiles to search for a more robust signature that could be used to classify tumor and normal tissues or tumor subtypes.

APPENDIX A

Support Materials for Chapter 2

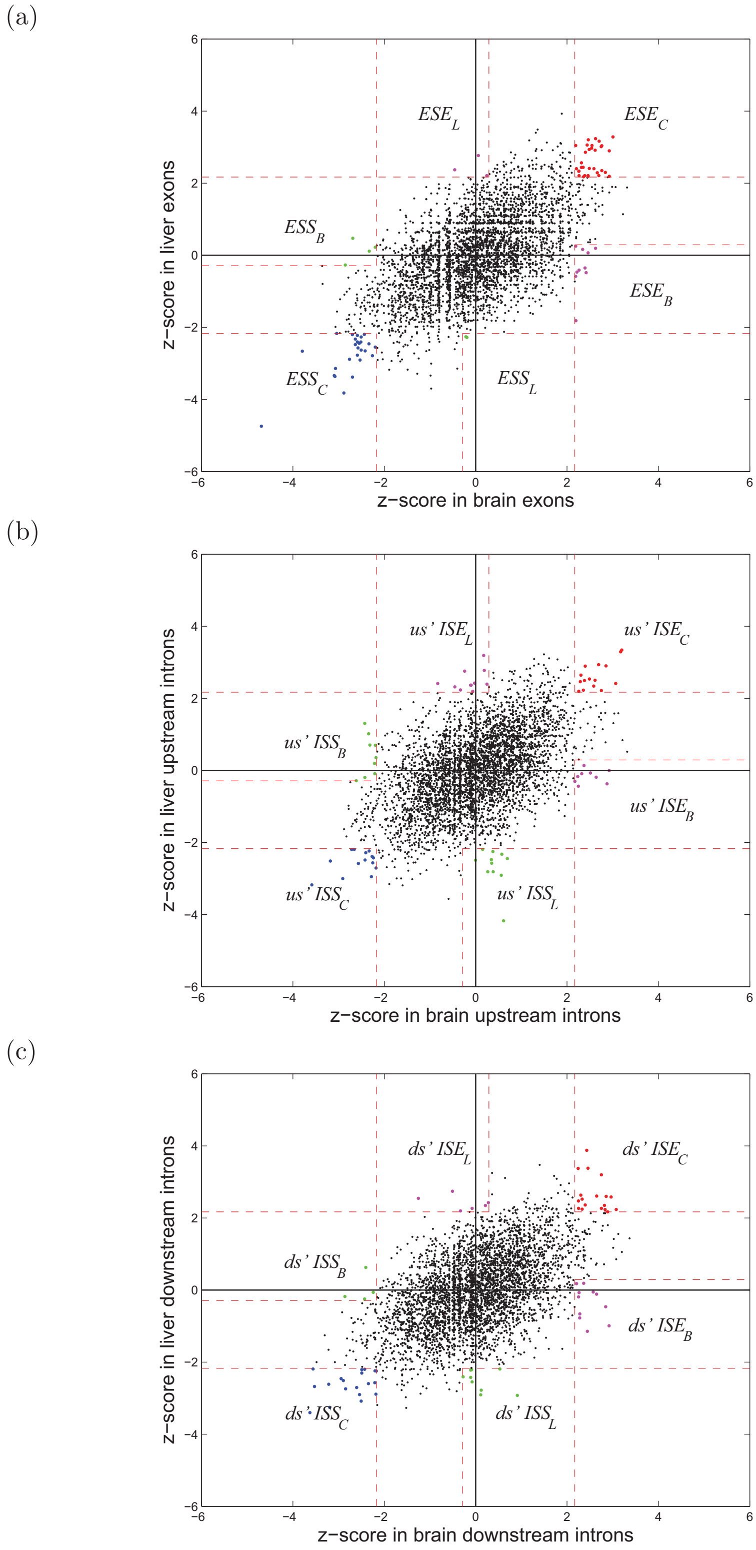


Figure A.1: z-scores for all hexamers in brain and liver. (a) z-scores in exons. ESE_C, ESE_B and ESE_L stand for common ESE, brain-specific ESE and liver-specific ESE, respectively. (b) z-scores in 400 nt intronic sequences upstream of the exons. (c) z-scores in 400 nt intronic sequences downstream of the exons.

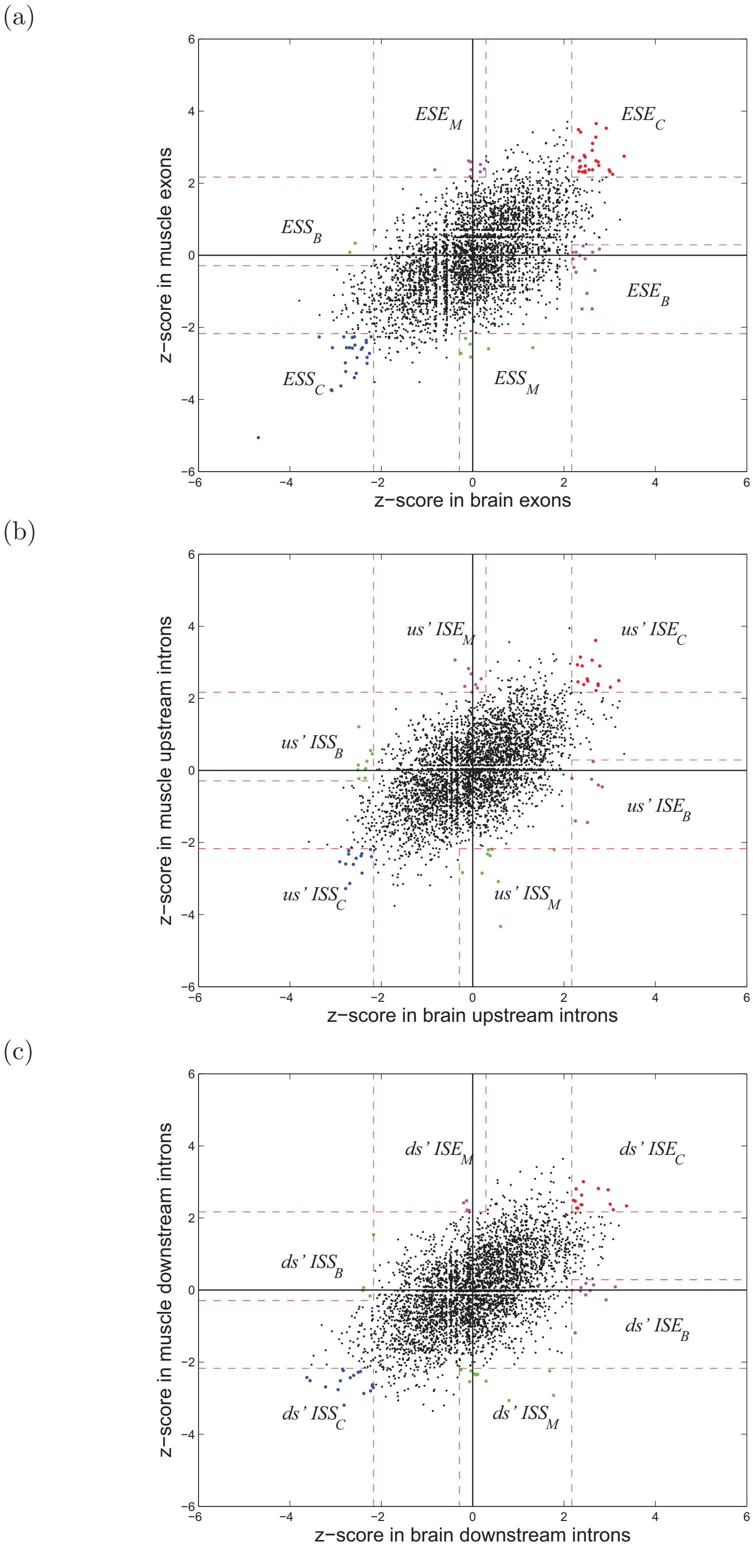

(a)
(b)
(c)

Figure A.2: z-scores for all hexamers in brain and muscle. (a) z-scores in exons. ESE_C, ESE_B and ESE_M stand for common ESE, brain-specific ESE and muscle-specific ESE, respectively. (b) z-scores in 400 nt intronic sequences upstream of the exons. (c) z-scores in 400 nt intronic sequences downstream of the exons.

Table A.1: First page of the list of 456 common and tissue-specific SREs and comparison with existing results (see supplementary table 1 in [1] for the full version of the table). Columns 2 to 6 specify the type of the SREs. Column 7 indicates whether an SRE is also a RESCUE-ESE. Last column contains the information of an SRE in SpliceAid database, which includes the binding factors, PMID reference and the number of the sequences (in parentheses) that our SRE can match to. If more than two records can be found in SpliceAid, only first two are kept.

Motif	ESE	ESS	5’ISE	5’ISS	3’ISE	3’ISS	RescueESE	Selected SpliceAid
AACUGC	BL?	.	.	.	.	.	-	
AAGAAG	BLM	.	.	.	.	.	AAGAAG	HTra2α,9546399(2);HTra2β1,9546399(2);
AAGCAG	BLM	.	.	.	.	.	AAGCAG	SC35,10094314(2);
AUCUAU	BL?	.	.	.	BL-	.	-	
ACUUCG	BL?	.	.	.	.	.	-	9G8,10094314(1);SRp20,10094314(1);
ACGGCA	BLM	.	.	.	.	.	-	
AGAAGC	BLM	.	.	.	.	.	AGAAGC	SC35,7543047(1);
AGCAGC	BL?	.	.	.	.	.	AGCAGC	FMRP,15805463(1);
AGCUGC	BLM	.	.	.	.	.	-	
AGGAAC	BLM	.	.	.	.	.	AGGAAC	SF2/ASF,7543047(2);FMRP,15805463(1);
UAUGAC	BLM	.	.	.	.	.	-	
UCGACU	BL?	.	.	.	.	.	-	SRp20,10094314(4);
UGUUAG	BLM	.	.	.	.	.	-	
UGGAGC	BLM	.	.	.	.	.	-	
UGGUGU	BL?	.	.	.	.	.	-	
UGGGCA	BL?	.	.	.	.	.	-	
CACGGC	BLM	.	.	.	.	.	-	
CAGCAA	BL?	.	.	.	.	.	-	
CUCAUA	BL?	.	.	.	.	.	-	
CUGGUG	BL?	.	.	.	.	.	-	FMRP,15805463(1);
CGACUG	BL?	.	.	.	-L-	.	-	FMRP,15805463(4);
CGGCAC	BLM	.	.	.	.	.	-	SF2/ASF,16825284(1);
CGGCCA	BL?	.	.	.	.	.	-	
GACUAU	BLM	.	.	.	.	.	-	
GUGAUA	BLM	.	.	.	.	.	-	
GUGGCU	BLM	.	.	.	.	.	-	FMRP,15805463(1);
GCAGAA	BL?	.	.	.	.	.	GCAGAA	
GCCAAC	BL?	.	.	.	.	.	-	
GGAUUU	BL?	.	.	.	.	.	-	
GGAUGA	BL?	.	.	.	.	.	GGATGA	SF2/ASF,7543047(1);
GGAGCA	BLM	.	.	.	.	.	-	SRp40,9037021(4);
GGUCAG	BL?	.	.	.	.	.	-	SC35,10629063(1);
AUGAGC	B--	.	.	.	.	.	-	
ACGCGC	B--	.	.	.	.	.	-	SF2/ASF,7543047(4);
AGCCUG	B--	.	.	.	.	.	-	
UCUUGC	B-?	.	.	.	.	.	-	
CUGAAA	B--	.	.	.	.	.	CTGAAA	
CGCUGC	B--	.	.	.	.	.	-	SC35,10629063(1);
GCAUUC	B--	.	.	.	.	.	-	
GCGCGC	B--	.	.	.	.	--M	-	SF2/ASF,7543047(1);
GGGGAC	B-?	.	.	.	.	.	-	
GGGGCC	B--	.	.	B?M	.	.	-	
AAAUCA	-L?	.	.	.	.	.	-	
AUGUAC	-L?	.	.	.	.	.	-	
UCAGGC	-L?	.	.	.	.	.	-	
AGGGUG	B?M	.	.	.	.	.	-	
UUUUGG	B?M	.	.	.	.	.	-	
UUCGAG	B?M	.	.	.	.	.	-	SC35,7543047(3);SC35,10094314(3);et al.
CAAUGA	B?M	.	.	.	.	.	-	
CAACCA	B?M	.	.	.	.	.	-	

Continued in supplementary table 1 of reference [1]

Table A.2: List of 71 SREs with p-value < 0.01 in the position bias test. Column 2 and 3 represent the annotation of the SRE. Column 4 gives the total number of occurrence of the SRE in the data analyzed. Column 5 gives the total number of intronic or exonic sequences used in analyses. The last three columns list p-value, chi-square statistic and degree of freedom, respectively.

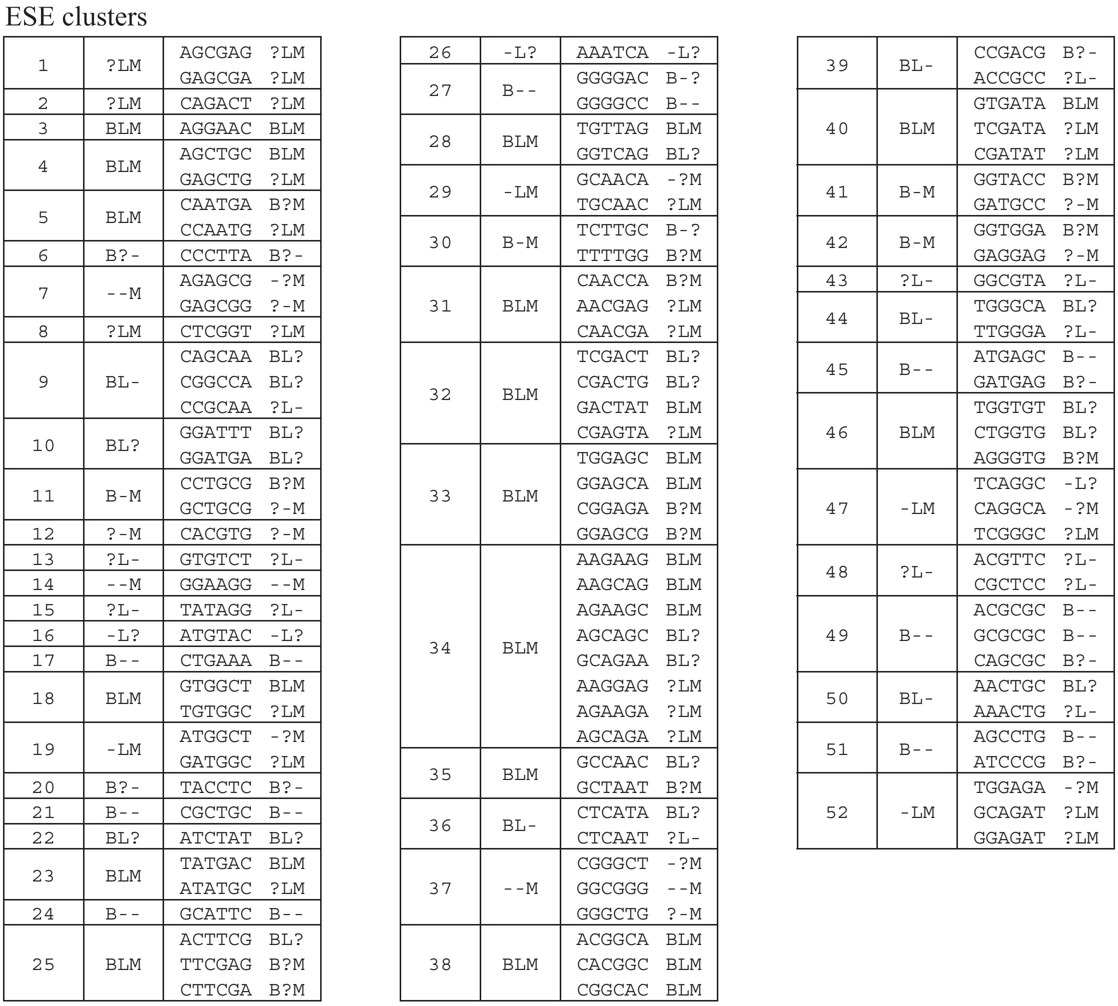
Elements	Anno.	Anno.	No.Occu.	No.Seq.	p-value	statistic	DF
CTCTCT	5ISS	-LM	301	721	1.33E-16	158.1739	38
TCTCTC	5ISS	-LM	279	721	2.59E-14	144.3971	38
ATGGAG	3ISE	BLM	142	994	1.83E-13	139.1915	38
AAGATT	5ISS	BLM	168	1030	4.69E-12	130.4024	38
TTGTAC	5ISE	BLM	93	964	6.85E-11	122.9677	38
TCTCTT	5ISS	?LM	263	721	1.58E-10	120.6236	38
TTCTTC	3ISS	BLM	316	1044	1.61E-09	113.9873	38
AGATAA	3ISS	B?M	101	658	4.69E-08	104.0396	38
CTGCCT	ESS	BLM	43	454	5.26E-08	51.65116	9
CCTAAA	5ISE	BLM	116	964	1.20E-07	101.1897	38
TGATTT	5ISE	?LM	196	639	1.51E-07	100.4796	38
TTACTG	3ISE	BLM	177	994	2.83E-07	98.55172	38
CTTGGG	5ISE	BLM	202	964	5.63E-07	96.4	38
CTAAAA	3ISS	BLM	189	1044	1.41E-06	93.49206	38
TTAAAG	ESS	BLM	33	454	1.63E-06	43.66667	9
GTGATT	5ISE	B?M	113	663	2.21E-06	92.05556	38
TTCATG	5ISE	B?M	136	663	2.41E-06	91.77778	38
AGGAAC	ESE	BLM	38	483	2.63E-06	42.52632	9
CTCTTT	5ISS	?LM	226	721	3.50E-06	90.57399	38
AGTTTC	3ISS	BL?	136	702	4.21E-06	89.97059	38
TATGCA	5ISE	BL?	90	626	6.83E-06	88.38636	38
GGGAAA	5ISS	BL?	151	686	1.08E-05	86.87417	38
TGTCAC	5ISE	?LM	88	639	1.58E-05	85.58621	38
AGATAT	5ISE	BL?	89	626	1.97E-05	84.84091	38
GAGGGA	3ISE	BL?	144	643	2.81E-05	83.63636	38
GTTTGG	3ISS	BL?	134	702	3.27E-05	83.1194	38
TTGATT	3ISE	BL?	120	643	4.17E-05	82.28814	38
CTCTGT	3ISE	?LM	239	663	4.97E-05	81.68103	38
TAAAAG	3ISS	B?M	135	658	5.42E-05	81.37778	38
TCATCA	3ISS	BLM	142	1044	5.43E-05	81.36765	38
GAATAT	3ISS	BL?	107	702	8.62E-05	79.75	38
TCCTGC	5ISE	B?M	134	663	1.06E-04	79.00752	38
CTGGAG	3ISS	BLM	245	1044	1.11E-04	78.86885	38
AGACAC	5ISE	?LM	90	639	1.51E-04	77.75	38
GGGGGA	3ISE	B?M	117	682	1.78E-04	77.16814	38
GCAGCA	5ISE	BLM	163	964	1.87E-04	76.9816	38
CACTTA	5ISE	B?M	86	663	2.23E-04	76.34884	38
GGCCCT	3ISS	?LM	99	728	2.25E-04	76.30612	38
TAGAAC	3ISS	B?M	96	658	2.29E-04	76.25	38
TTGTAG	3ISE	B?M	118	682	2.94E-04	75.33333	38
TCCTCA	5ISS	B?M	135	653	2.95E-04	75.31298	38
GGAGGG	3ISE	BL?	189	643	3.28E-04	74.92064	38

Continued on next page

Table A.2 -- *Continued from previous page*

Elements	Anno.	Anno.	No.Occu.	No.Seq.	p-value	statistic	DF
GCAGAA	ESE	BL?	25	320	3.47E-04	30.6	9
CCTTCT	3ISS	BL?	163	702	4.21E-04	74	38
TGATTG	5ISE	BL?	89	626	5.04E-04	73.31818	38
CACTTG	5ISE	?LM	109	639	5.32E-04	73.11927	38
CTAGAA	5ISS	BLM	177	1030	5.37E-04	73.08475	38
CTTTCT	5ISS	BL?	223	686	6.37E-04	72.43243	38
ATTACT	3ISE	B?M	100	682	7.93E-04	71.6	38
TATAAA	5ISE	?LM	150	639	8.70E-04	71.24324	38
TGAGAA	5ISS	B?M	151	653	1.03E-03	70.60403	38
CATGGG	3ISS	?LM	161	728	1.10E-03	70.3354	38
GAGATT	5ISE	BL?	89	626	1.51E-03	69.09302	38
TAGAAG	3ISS	?LM	100	728	1.78E-03	68.42424	38
AAAAGC	3ISS	B?M	108	658	1.88E-03	68.22222	38
GATGGC	ESE	?LM	21	306	1.94E-03	26.14286	9
CCCACC	5ISS	BL?	123	686	2.21E-03	67.56098	38
GTTAAA	3ISE	B?M	109	682	2.30E-03	67.3945	38
ACTTAG	5ISE	B?M	95	663	2.99E-03	66.33684	38
AGCCAT	3ISS	B?M	102	658	3.13E-03	66.14	38
GGAAAC	5ISS	?LM	100	721	3.13E-03	66.14	38
ACTCAA	3ISS	?LM	85	728	3.28E-03	65.95294	38
ATTTTG	3ISS	BL?	207	702	3.47E-03	65.71707	38
GCATGG	3ISS	BL?	136	702	3.52E-03	65.65672	38
CTTCCT	5ISS	?LM	290	721	4.32E-03	64.80282	38
GACTCA	ESS	B?M	26	294	5.70E-03	23.23077	9
CATTTT	5ISS	BL?	256	686	6.19E-03	63.27059	38
TCAGCT	5ISE	?LM	102	639	6.55E-03	63.02	38
AAGAAG	ESE	BLM	45	478	6.72E-03	22.77778	9
CTAAAG	5ISE	BLM	131	964	6.88E-03	62.80916	38
CTTCTC	3ISS	BL?	169	702	8.49E-03	61.89157	38

Table A.3: List of clustering result for 6 types of SREs (ESE, ESS, upstream ISE, upstream ISS, downstream ISE and downstream ISS). Three columns give cluster ID, cluster annotation, and the SREs in a cluster and their original annotations, respectively.



ESE clusters

1	?LM	AGCGAG ?LM GAGCGA ?LM
2	?LM	CAGACT ?LM
3	BLM	AGGAAC BLM
4	BLM	AGCTGC BLM GAGCTG ?LM
5	BLM	CAATGA B?M CCAATG ?LM
6	B?-	CCCTTA B?-
7	--M	AGAGCG -?M GAGCGG ?-M
8	?LM	CTCGGT ?LM
9	BL-	CAGCAA BL? CGGCCA BL? CCGCAA ?L-
10	BL?	GGATTT BL? GGATGA BL?
11	B-M	CCTGCG B?M GCTGCG ?-M
12	?-M	CACGTG ?-M
13	?L-	GTGTCT ?L-
14	--M	GGAAGG --M
15	?L-	TATAGG ?L-
16	-L?	ATGTAC -L?
17	B--	CTGAAA B--
18	BLM	GTGGCT BLM TGTGGC ?LM
19	-LM	ATGGCT -?M GATGGC ?LM
20	B?-	TACCTC B?-
21	B--	CGCTGC B--
22	BL?	ATCTAT BL?
23	BLM	TATGAC BLM ATATGC ?LM
24	B--	GCATTC B--
25	BLM	ACTTCG BL? TTCGAG B?M CTTCGA B?M
26	-L?	AAATCA -L?
27	B--	GGGGAC B-? GGGGCC B--
28	BLM	TGTTAG BLM GGTCAG BL?
29	-LM	GCAACA -?M TGCAAC ?LM
30	B-M	TCTTGC B-? TTTTGG B?M
31	BLM	CAACCA B?M AACGAG ?LM CAACGA ?LM
32	BLM	TCGACT BL? CGACTG BL? GACTAT BLM CGAGTA ?LM
33	BLM	TGGAGC BLM GGAGCA BLM CGGAGA B?M GGAGCG B?M
34	BLM	AAGAAG BLM AAGCAG BLM AGAAGC BLM AGCAGC BL? GCAGAA BL? AAGGAG ?LM AGAAGA ?LM AGCAGA ?LM
35	BLM	GCCAAC BL? GCTAAT B?M
36	BL-	CTCATA BL? CTCAAT ?L-
37	--M	CGGGCT -?M GGCGGG --M GGGCTG ?-M
38	BLM	ACGGCA BLM CACGGC BLM CGGCAC BLM
39	BL-	CCGACG B?- ACCGCC ?L-
40	BLM	GTGATA BLM TCGATA ?LM CGATAT ?LM
41	B-M	GGTACC B?M GATGCC ?-M
42	B-M	GGTGGA B?M GAGGAG ?-M
43	?L-	GGCGTA ?L-
44	BL-	TGGGCA BL? TTGGGA ?L-
45	B--	ATGAGC B-- GATGAG B?-
46	BLM	TGGTGT BL? CTGGTG BL? AGGGTG B?M
47	-LM	TCAGGC -L? CAGGCA -?M TCGGGC ?LM
48	?L-	ACGTTC ?L- CGCTCC ?L-
49	B--	ACGCGC B-- GCGCGC B-- CAGCGC B?-
50	BL-	AACTGC BL? AAACTG ?L-
51	B--	AGCCTG B-- ATCCCG B?-
52	-LM	TGGAGA -?M GCAGAT ?LM GGAGAT ?LM

ESS clusters

1	BLM	TTTAAA BLM TTTTAA BL? TTTGAA BL? TTGAAA BL?
2	BLM	AAATAA BLM AATTAT BLM GTATAA BL? AATAAT B?M TAATAA B?M TTATAA ?LM
3	-LM	ACCTTT -L? ACCTCT ?LM
4	BLM	ACTCAC BL? GACTCA B?M
5	B?M	CGCTGA B?M
6	B-?	AGAGTG B-?
7	BLM	GGGTCT B?M GGTCTT ?LM
8	B?M	AACACT B?M
9	BL?	AGATAT BL?
10	?LM	CAGACC ?LM
11	-?M	AAGCCT -?M
12	?-M	GGGTTG ?-M
13	-L?	ACATCC -L?
14	BLM	ACCGTC BLM
15	--M	CCGCCC -?M CCGCGC -?M CCGCTC ?-M
16	BL?	TGTAGC BL?
17	BL?	AGGGGA BL?
18	B-M	ATAATA B-M ATAAGA B?M
19	--M	AATTCC --M
20	BLM	CCATGT BL? CATTTG B?M CCCATT B?M

21	BL-	TTCTTT BL? GTTCTG B?- ATTCTT ?L-
22	B--	ATGAAT B--
23	BLM	TTAAAG BLM CTTAAG B?M GTAAAG ?LM
24	B?M	TATTTA B?M GATTAA B?M
25	?LM	CACTCT ?LM CTCTGT ?LM
26	BLM	GAACTC BL? TAACTT ?LM
27	BLM	TCTTAG BLM TGCTTA BL?
28	B-M	AATGTT B-? ATATTG B?M
29	BLM	GTGTTT BL? TTGTTT B?M
30	BLM	ACAATA BL? ATACAA ?LM
31	BLM	TAAGAT BLM TAGGCT BL?
32	--M	TCCTTG --M CCCTGC -?M
33	BLM	TGCCTG BLM CTGCCT BLM CCTGCC BLM GCCTGC BLM CCTCCC ?LM
34	?L-	CCCCAG ?L- CGGCAG ?L-

5' ISE (upstream intronic enhancer) clusters

1	BLM	CTTGGG BLM
2	BLM	GCAGCA BLM
3	BL?	TATGCA BL?
4	-L?	CCGAAC -L?
5	-L?	TTGTAG -L?
6	BLM	TAACGT BLM
7	?L-	AAGTTT ?L-
8	-?M	AGGCCG -?M
9	B-?	TTGGTG B-?
10	B?M	AAGTGA B?M
11	BLM	TTGTAC BLM
12	--M	AGCTGA -?M
		GCTGAC --M
13	BL-	CACATA BL-
		CAAATG ?L-
14	B--	AACTCA B--
15	B-?	TAGTCG B-?
16	BL?	GGAGTC BL?
17	BL?	CGACAG BL?
18	B--	CATGCA B--
19	-LM	CTATAA -L?
		ATAAAT -?M
		TATAAA ?LM
20	-L?	AGGAGA -L?
21	B-M	GGTGCC B-?
		GCTCCC B?M
22	-LM	GACACA -L?
		AGACAC ?LM
		TGTCAC ?LM
23	B-M	CTGGGC B-?
		CAGGGG B?M
24	-L?	GATGGT -L?
25	--M	TCCCCT --M
26	--M	TAATAA ?-M
		TAATAT -?M

27	-L-	ACACTG -L?
		TACACT -L-
28	BLM	ACTTAG B?M
		CACTTA B?M
		CACTTG ?LM
29	BLM	AGCTTA B?M
		TCAGCT ?LM
30	BLM	TGTGCG BL?
		GTGTGC BL?
		TGCGTG ?LM
		GTGCGT ?LM
31	-LM	ACCGGA -?M
		CCAGAG ?LM
32	-LM	ACTTGA -L?
		CTTGAA ?LM
33	BLM	CTAAAG BLM
		CCTAAA BLM
34	BL-	AGATAT BL?
		GAGATT BL?
		TGAGAT B?-
35	?-M	CAGCCA ?-M
		GCAGCC ?-M
		GGCAGC ?-M
36	B-M	TCCTGA B-?
		TTCATG B?M
		TCCTGC B?M
37	BLM	TGATTG BL?
		GTGATT B?M
		TGATTT ?LM
38	B--	CCTACT B--
39	B--	ATATGT B--
		ATGTGT B?-
40	-LM	CGCATG -L?
		GCACGC -L?
		AGCAAG -?M

5’ ISS (upstream intronic silencer) clusters

1	-LM	TGTATG -L? GTATAT -?M CTATAT ?LM GTGTAT ?LM
2	B-M	CATCCG B-M
3	-?M	AATTGC -?M
4	B?-	GTTCCG B?-
5	?LM	TCTGCA ?LM TCTGCG ?LM
6	B?M	ATATGA B?M
7	-LM	CTTCTT -L? CTTCCT ?LM
8	BL-	CCCACC BL? CCGACC ?L- CGACCC ?L-
9	B?M	CAGGAC B?M
10	BLM	TTACAA BL? TGAGAA B?M
11	-LM	CCCCCA -L? CTCCCG ?LM
12	B?-	ATCGTT B?-
13	-L?	TAAGGC -L?
14	BLM	GCCGAT BLM TGCCGA B?M
15	B-?	ACTATT B-?
16	-LM	GGATAC -L? GGAAAC ?LM
17	B-?	CTTACA B-?
18	BLM	CATTTT BL? ATTTAG B?M
19	BL?	GGGAAA BL?
20	B?-	CATCAC B?-
21	-L?	GAGAAG -L?
22	B?-	ACAGGC B?- CAGGCT B?-
23	BLM	AAGATT BLM
24	BL-	CAGCGG B?- CTGCGG ?L-

25	B-M	GGGGCC B?M GGGTCT ?-M GGGGCG B?M
26	B--	CAAAAA B-- CAAGGA B?-
27	-L-	ATTCAT -L- ATCCCT -L?
28	--M	ATATAC --M ACACAC --M
29	-LM	TATGTG -?M CCTATG ?LM
30	B-M	TCCTCA B?M CATCAA ?-M
31	-LM	CTAATT -?M TCTAAT ?LM
32	B--	GTATCG B-- GGTAGC B?-
33	BLM	CTAGAA BLM CTGCAA BL? CTGGAA BL?
34	BLM	TTATCA BL? GTTATC ?LM
35	-LM	TCTCTC -LM CTCTCT -LM TCTCTT ?LM CTCTTT ?LM
36	BL?	CCCTTC BL? CCCCGC BL?
37	B-M	CAGTCA B-? GTCAGT B-? CAGTCC B?M
38	BL-	CTTTCT BL? GGCTTT ?L-
39	BLM	TCCAAT BL? CAATAC ?LM
40	B-?	AACCAG B-? AACCGT B-?
41	-LM	GCATGC -L? ATGCAG -?M

3’ ISE (downstream intronic enhancer) clusters

1	BL?	GGATGC BL?
2	BLM	GTTAAA B?M TTAAAA ?LM
3	B?-	GGAACT B?-
4	B-?	CTTCCT B-?
5	BL?	CTTGCC BL?
6	BL?	AATCAT BL?
7	BLM	ATTGAT BL? TTGATT BL? TTGCTT BLM
8	BLM	TGAGGG BL? GAGGGA BL? GGAGGG BL? GGGGGA B?M
9	B--	CCGACC B-? CGACAT B?-
10	?-M	AGGGCT ?-M
11	?L-	CGCGCT ?L-
12	BLM	GAACCG BLM
13	?LM	CTCTGT ?LM
14	-?M	TTTAGT -?M
15	B?-	CTAGTA B?-
16	B-?	TTGAGG B-?
17	B?-	CAGTTA B?-
18	-L-	ATGATC -L? TGAACA -L? GATCAG ?L-
19	B-M	GAGTTG B-M
20	-L?	GCGGAT -L?
21	--M	TGCATG --M
22	B-M	GCGTGC B?M GCGAAC ?-M
23	-L-	CGACTG -L- AACTGG ?L-
24	BLM	GGCAGC ?LM GGCAGG B?M GGCGGC B?M

25	BLM	CGCTAA B?M ACGTTA ?LM CGTTAT ?LM
26	B--	GGAATG B-? GGGATG B--
27	B-?	CACCCT B-? CACCGT B-?
28	BL-	CTGCGT BL? TTGCGA ?L-
29	BL-	GTCGTC BL? TCATCT ?L-
30	BLM	TTACTG BLM CATTAC BL? ATTACT B?M
31	BLM	TTGTAG B?M ATGGAG BLM TGTGGA ?LM GTGGAG ?LM
32	B--	ACGCCC B-- TGCGCC B-? GCGCCT B--
33	--M	TGGTTT -?M AATGGT ?-M
34	BL-	AGGGGC B?- CGGTGC ?L-
35	BLM	CCGAAG BL? CCAAAC B?M
36	BL?	GCCTTT BL?
37	--M	GTCTCG --M
38	BL-	ATCTAT BL- TTCTAT B?- TATCTA ?L- TCTATC ?L-
39	?-M	TGACAT ?-M CTGACA ?-M
40	-L-	ATTACA -L- GCTTAC -L?

3’ ISS (downstream intronic silencer) clusters

1	BLM	TCTAAA BL? CTAAAA BLM
2	BL?	ATTTTG BL?
3	BLM	CTTCTT BL? CTTCTC BL? CCTTCT BL? CCATCT B?M
4	BLM	CGAGAA BLM TAGAAC B?M TAGAAG ?LM
5	--M	TCTATC --M CTATCT --M CTAGCC -?M
6	B?-	ACGATG B?-
7	-L-	TGAAGG -L- TGAACG ?L-
8	B?M	TCGTTC B?M
9	-L?	GCTCGG -L?
10	?-M	TATATT ?-M TTTTAT ?-M
11	BLM	GAATAT BL? GCATAA B?M
12	B-M	AAGCCA B?M AGCCAT B?M GCCATG ?-M
13	BL?	AAACTT BL?
14	-LM	CCCCGC -LM
15	-L?	TGGGGA -L?
16	?L-	TACCTC ?L-
17	BL?	ACTCGG BL?
18	BLM	CTGGAG BLM
19	B-?	GAAGAG B-? GGAAGA B-?
20	-?M	ATCATC -?M AGCATT -?M
21	BL-	TCCTTA B?- TTCATT ?L-
22	--M	GAGGCT --M
23	BLM	TACGTA BLM
24	B?-	GGATCC B?-
25	BLM	CACATG BL? GCATGG BL? CATGGG ?LM
26	B-M	CTACCC B?M TTCCCC ?-M
27	BLM	CTTCAA B?M ACTCAA ?LM
28	BL-	AAGTTT BL? AGTTTC BL? GTTTGG BL? GTTTCT B?-
29	BLM	TTCTTC BLM TCATCA BLM
30	BLM	AAAAGC B?M TAAAAG B?M ATAAGT ?LM
31	--M	ACTATT --M ACTCTA ?-M
32	?L-	TATGGC ?L- GCTATG ?L-
33	-L-	TGCATG -L- GCATGT -L- GTGCAT ?L-
34	-L-	CGCCCC -L- GCCCCG -L? CCCCGA ?L-
35	B-M	AGATAA B?M AGCTAG ?-M AGGTAG ?-M
36	?L-	CTCTTC ?L-
37	?LM	GGCCCT ?LM
38	-LM	ACGGCT -L? TCGGGT ?LM CCGGAT ?LM
39	B-?	CACACA B-? GCACAG B-?
40	--M	GCGCGC --M CGCGCC -?M

APPENDIX B

Support Materials for Chapter 3

Table B.1: First two pages of the list of all the SREs and SRE pairs identified in 9 tissues (see Table S1 in [2] for the full version of the table). Column ‘Beta’, ‘p’ are the regression coefficient and p-value in the final multiple regression model. Column ‘supp.’ (support) is the number of SRE or SRE pair that is present in ASEs used for inference in each tissue. If an SRE is detected, Column ‘6mer1’ gives the SRE; if SRE pairs are detected, Columns ‘6mer1’ and ‘6mer2’ give the SREs in the pair. Column ‘tissue’ list the tissues where the SRE or SRE pair identified. The last two columns give the SF(s) that have been identified with either SELEX or RNAcompete to bind to the hexmer. References for SELEX or RNAcompete results are listed below the table.

beta	p	supp.	6mer1	6mer2	tissues	exp.6mer1	exp. 6mer2
1.27	1.58E-03	64	AATTGG_UU	-	adipose	-	-
0.79	5.98E-03	99	AATGTT_UU	-	adipose	-	-
0.94	3.90E-03	75	AAGTGG_UU	-	adipose	-	-
-1.21	6.90E-03	37	ATATAG_UU	-	adipose	-	-
-1.14	1.88E-03	57	ATTTAC_UU	-	adipose	-	-
1.29	9.07E-04	50	ACAAGA_UU	-	adipose	-	-
1.00	2.45E-03	74	ACCCCA_UU	-	adipose	hnRNPK;	-
0.71	1.27E-03	173	TTTGGG_UU	-	adipose	-	-
0.77	1.87E-03	137	TTGAAA_UU	-	adipose	-	-
1.52	3.16E-05	59	CAAGGC_UU	-	adipose	-	-
-1.71	6.27E-04	31	GATCCA_UU	-	adipose	-	-
1.39	4.03E-03	32	GACATC_UU	-	adipose	-	-
-1.02	2.57E-03	67	GACTGA_UU	-	adipose	-	-
-1.56	9.89E-09	110	GAGTGG_UU	-	adipose	-	-
0.64	7.65E-03	142	GAGGAG_UU	-	adipose	-	-
-1.01	7.51E-04	90	GGAAAT_UU	-	adipose	-	-
1.32	1.48E-04	62	AACAAG_UD	-	adipose	-	-
-1.08	1.19E-03	68	ATCCAG_UD	-	adipose	-	-
-1.46	1.65E-05	68	ACTCAT_UD	-	adipose	-	-
-0.70	9.33E-04	201	TTCTCT_UD	-	adipose	-	-
-1.16	2.22E-04	80	TCTGCA_UD	-	adipose	-	-
-1.17	1.66E-05	113	TCCCTC_UD	-	adipose	-	-
-1.05	1.37E-05	160	CTTCCT_UD	-	adipose	-	-
-1.30	7.55E-04	52	CTCTAG_UD	-	adipose	-	-
1.31	1.90E-03	43	CCAAGC_UD	-	adipose	-	-
0.75	5.59E-03	128	CCTTCC_UD	-	adipose	-	-
1.15	6.93E-04	67	CCTCAA_UD	-	adipose	-	-
1.84	3.48E-03	19	GCACGG_UD	-	adipose	-	-
-1.34	1.63E-03	41	AGATCT_EXON	-	adipose	9G8;	-
-1.41	5.70E-03	28	AGGTCC_EXON	-	adipose	-	-
-0.88	5.88E-03	79	TGCAGG_EXON	-	adipose	PSF;SRp40;	-
1.09	9.44E-05	119	CTGGGA_EXON	-	adipose	PSF;	-

Continued on next page

Table B.1 -- *Continued from previous page*

beta	p	supp.	6mer1	6mer2	tissues	exp.6mer1	exp. 6mer2
-1.90	5.45E-03	16	GGCATA_EXON	–	adipose	–	–
1.14	2.74E-05	107	AAATTG_DU	–	adipose	–	–
-1.52	1.01E-06	82	AAGCTT_DU	–	adipose	–	–
0.99	1.24E-03	84	ATCTGT_DU	–	adipose	–	–
0.90	6.02E-03	75	ACATCT_DU	–	adipose	–	–
1.16	2.49E-03	52	TATGGA_DU	–	adipose	–	–
-1.51	5.15E-06	98	TTAAGT_DU	–	adipose	–	–
0.89	2.65E-03	88	TTGGTG_DU	–	adipose	–	–
1.45	4.50E-08	136	TCTCTG_DU	–	adipose	–	–
1.52	4.20E-05	56	TCCTAA_DU	–	adipose	–	–
-1.03	8.13E-05	122	TCCCTT_DU	–	adipose	hnRNPK;	–
-1.34	1.02E-04	65	CACTCC_DU	–	adipose	–	–
1.47	3.79E-04	45	CCGGGC_DU	–	adipose	–	–
-2.32	6.66E-05	22	CGTCTT_DU	–	adipose	–	–
0.91	2.47E-03	89	GAAAGG_DU	–	adipose	–	–
1.51	6.61E-05	53	GATTCA_DU	–	adipose	–	–
-1.26	1.15E-03	50	AAGAGC_DD	–	adipose	PSF;	–
1.16	1.66E-04	83	ATAACT_DD	–	adipose	–	–
1.02	4.21E-04	92	ATTCCT_DD	–	adipose	–	–
3.15	7.02E-06	15	ATTGCG_DD	–	adipose	–	–
-1.00	4.13E-04	98	TAATTG_DD	–	adipose	–	–
1.51	9.06E-05	51	TAAGAG_DD	–	adipose	PSF;	–
-1.13	8.17E-04	67	TACAGT_DD	–	adipose	–	–
-1.04	6.27E-03	52	TAGGTT_DD	–	adipose	–	–
0.97	5.22E-04	105	TCTTGA_DD	–	adipose	–	–
-0.95	1.05E-03	92	TGCCAC_DD	–	adipose	–	–
0.80	2.76E-03	110	CTTGTT_DD	–	adipose	–	–
-1.63	8.90E-08	85	GAGGGG_DD	–	adipose	–	–
-3.45	7.43E-08	21	AAAATG_UU	TCTCTG_DU	adipose	SAM68;	–
-3.00	1.09E-04	17	AATTGG_UU	TTTTCT_UU	adipose	–	TIA1/TIAR;PTB;
3.55	1.50E-07	17	ATTACA_UU	ATATTT_UD	adipose	–	SAM68;
2.78	5.58E-06	28	TTTTCT_UU	TTAAGT_DU	adipose	TIA1/TIAR;PTB;	–

-1.30	7.87E-05	75	TTTTCT_UU	TATTTT_DD	adipose	TIA1/TIAR;PTB;	TIA1/TIAR;HuR(Rc);
							SAM68;

3.04	1.50E-05	15	TCTGTG_UU	GGGCTT_UD	adipose	–	–
2.17	6.15E-05	26	TCCCTG_UU	CCAGGG_UU	adipose	–	–
-2.33	6.38E-04	19	TCCCTG_UU	CTGGGA_EXON	adipose	–	PSF;
-2.04	1.19E-03	20	TCCCTG_UU	CTGTCC_DD	adipose	–	–
2.26	2.76E-05	26	ATATTT_UD	TAGTTT_DU	adipose	SAM68;	–
-3.00	5.79E-06	17	TAGTTT_UD	TACTTT_DD	adipose	–	PTB(Rc);
-2.67	6.26E-06	21	TTAAAT_UD	TTAAGA_DU	adipose	SAM68;	–
-4.20	1.24E-08	14	TTTAAT_UD	GTATAG_UD	adipose	SAM68;	DAZAP1;
-1.46	1.45E-04	57	TTTAAT_UD	TTTTAT_DD	adipose	SAM68;	TIA1/TIAR;
-1.94	8.93E-04	23	TTTTAA_UD	CATTAT_DU	adipose	SAM68;	–
-2.00	2.64E-03	17	TTTTAC_UD	AGATTT_DU	adipose	–	–
-1.66	7.01E-03	20	TTTGTG_UD	TATGAA_DD	adipose	–	–
2.15	1.24E-03	17	TCTCCT_UD	GGCTGT_DU	adipose	–	–
-2.36	4.81E-05	22	CAAAAG_UD	TTAAAA_DD	adipose	–	SAM68;
2.19	3.54E-04	20	CCTGGG_UD	GGTCTC_UD	adipose	–	–
-1.39	1.90E-03	42	GTTTTA_UD	TTTTAT_DD	adipose	TIA1/TIAR;	TIA1/TIAR;
2.61	3.93E-04	14	GGCTGG_UD	AGTCCT_DU	adipose	–	–
1.90	5.65E-03	16	AGCTGG_EXON	ACCCTG_DD	adipose	–	–
1.83	5.10E-04	27	CACTGC_EXON	CTGCTG_EXON	adipose	–	MBNL1;
1.77	7.08E-03	17	ACCAGT_DU	TTGTTT_DU	adipose	–	TIA1/TIAR;HuR(Rc);
1.71	8.70E-04	29	ACCCCT_DU	CCCCTC_DU	adipose	–	–
-2.24	7.46E-04	17	AGATAT_DU	ATTTGT_DD	adipose	–	–
3.31	4.90E-06	14	CAGCTG_DU	AAGTGC_DD	adipose	–	–
-1.21	2.08E-03	52	TTAATT_DD	TTTTAT_DD	adipose	SAM68;	TIA1/TIAR;
-1.27	3.20E-03	41	TGTCCT_DD	GTCCTT_DD	adipose	–	–
-1.07	4.89E-03	90	ATTGAA_UU	–	brain	–	–

Continued on next page

Table B.1 -- *Continued from previous page*

beta	p	supp.	6mer1	6mer2	tissues	exp.6mer1	exp. 6mer2
-1.48	6.29E-04	67	AGTTGG_UU	-	brain	-	-
-1.83	4.42E-06	86	AGTGCA_UU	-	brain	-	-
2.38	1.30E-07	61	TACACA_UU	-	brain	-	-
1.68	6.41E-05	69	TTGGGC_UU	-	brain	-	-
-1.21	3.33E-03	76	TCATAT_UU	-	brain	-	-
1.55	2.24E-05	99	TGCCCC_UU	-	brain	-	-
2.42	1.11E-04	31	CATACC_UU	-	brain	-	-
2.28	2.13E-05	42	CACCAA_UU	-	brain	-	-
-1.15	4.54E-04	127	CTATTT_UU	-	brain	-	-
1.75	4.06E-04	56	CTAGTT_UU	-	brain	-	-
-1.05	3.64E-04	153	CTTCCT_UU	-	brain	-	-
-1.35	1.47E-04	103	CTGCCA_UU	-	brain	MBNL1;	-
2.76	1.35E-04	23	CCAATG_UU	-	brain	SF2;	-
1.37	1.14E-03	70	GAAATC_UU	-	brain	-	-
-2.89	3.40E-04	18	GACGCT_UU	-	brain	SC35;	-
1.10	2.24E-03	99	GAGGGT_UU	-	brain	-	-
1.63	3.67E-04	60	GCGGGT_UU	-	brain	-	-
-1.12	2.41E-03	96	ATAAAG_UD	-	brain	-	-
1.94	3.40E-03	27	ACAATC_UD	-	brain	-	-
1.35	1.35E-03	71	AGTTGA_UD	-	brain	-	-
2.32	4.69E-03	18	TCCGGG_UD	-	brain	-	-
1.13	4.14E-04	125	TGAAAG_UD	-	brain	-	-
-1.21	1.32E-03	89	TGTAGT_UD	-	brain	DAZAP1;	-
-1.72	3.43E-04	53	CAAGGG_UD	-	brain	-	-
1.74	8.98E-04	45	CTATAG_UD	-	brain	-	-
1.07	1.41E-03	116	CTTTCC_UD	-	brain	-	-
-1.73	1.58E-03	40	CCATAG_UD	-	brain	-	-
-1.99	2.08E-05	56	CCATTG_UD	-	brain	-	-
2.02	1.20E-08	106	CCTCTG_UD	-	brain	-	-
-2.19	7.95E-04	28	CGTTTT_UD	-	brain	-	-
1.12	6.13E-03	74	GATTGT_UD	-	brain	-	-

References for the last two columns:

SF name	# of 6mer in Ref.	Reference
hnRNPC	7	[90]
hnRNPA1	8	[91]
TIA1/TIAR	28	[92]
SAM68	47	[93]
NOVA	43	[94]
SRp40	45	[95]
SRp55	16	[95]
TRA2	13	[96]
9G8	98	[97]
SC35	90	[98]
hnRNPK	32	[99]
TLS	12	[100]
PSF	65	[101]
CUGBP2	20	[102]
QKI	2	[103]
DAZAP1	19	[104]
FOX1	2	[51]
SF2	104	[105]
SRp30c	38	[106]
PTB	17	[107]
MBNL1	49	[108]
SLM2(Rc)	9	[109]
RBM4(Rc)	6	[109]
SF2(Rc)	6	[109]
FUSIP1(Rc)	7	[109]
HuR(Rc)	7	[109]
PTB(Rc)	8	[109]

** "Rc" indicates that the binding site is obtained from RNAcompete experiment, while SF name without (Rc) indicates SELEX experiment*

APPENDIX C

Support Materials for Chapter 4

Table C.1: List of samples used in Chapter 4. Each patient has two sample available for RNA-Seq experiment. One is from the tumor tissue, the other one is derived from the same organ site as the tumor from the same patient. The first two columns list the 105 patient id and sample id given in TCGA database. The third column indicates the sample type. Column 'size' is the raw RNA-Seq data file size in unit of gigabyte. The last column is the number of millions of paired-end reads that can map to the transcriptome in each sample.

Patient ID	Sample ID	Tumor/Normal	size(GB)	#Reads(M)
TCGA-A7-A0D9	TCGA-A7-A0D9-01A-31R-A056-07	Tumor	6.2	56.5
TCGA-A7-A0D9	TCGA-A7-A0D9-11A-53R-A089-07	Normal	6.6	51.5
TCGA-A7-A0DB	TCGA-A7-A0DB-01A-11R-A00Z-07	Tumor	6.3	62.8
TCGA-A7-A0DB	TCGA-A7-A0DB-11A-33R-A089-07	Normal	5.9	48.4
TCGA-A7-A0DC	TCGA-A7-A0DC-01A-11R-A00Z-07	Tumor	7.4	72.9
TCGA-A7-A0DC	TCGA-A7-A0DC-11A-41R-A089-07	Normal	6.3	49.9
TCGA-A7-A13G	TCGA-A7-A13G-01A-11R-A13Q-07	Tumor	8.4	80.5
TCGA-A7-A13G	TCGA-A7-A13G-11A-51R-A13Q-07	Normal	8.1	65.3
TCGA-AC-A23H	TCGA-AC-A23H-01A-11R-A157-07	Tumor	7.8	78.1
TCGA-AC-A23H	TCGA-AC-A23H-11A-12R-A157-07	Normal	7.9	70.5
TCGA-AC-A2FB	TCGA-AC-A2FB-01A-11R-A17B-07	Tumor	7.5	69.9
TCGA-AC-A2FB	TCGA-AC-A2FB-11A-13R-A17B-07	Normal	8.4	78.7
TCGA-AC-A2FF	TCGA-AC-A2FF-01A-11R-A17B-07	Tumor	7.7	71.3
TCGA-AC-A2FF	TCGA-AC-A2FF-11A-13R-A17B-07	Normal	8.8	80.0
TCGA-BH-A0AY	TCGA-BH-A0AY-01A-21R-A00Z-07	Tumor	5.1	52.6
TCGA-BH-A0AY	TCGA-BH-A0AY-11A-23R-A089-07	Normal	6.4	52.5
TCGA-BH-A0AZ	TCGA-BH-A0AZ-01A-21R-A12P-07	Tumor	4.6	53.3
TCGA-BH-A0AZ	TCGA-BH-A0AZ-11A-22R-A12P-07	Normal	7.0	70.2
TCGA-BH-A0B2	TCGA-BH-A0B2-01A-11R-A10J-07	Tumor	4.8	58.0
TCGA-BH-A0B2	TCGA-BH-A0B2-11A-11R-A10J-07	Normal	6.0	58.4
TCGA-BH-A0B3	TCGA-BH-A0B3-01A-11R-A056-07	Tumor	6.2	66.9
TCGA-BH-A0B3	TCGA-BH-A0B3-11B-21R-A089-07	Normal	7.9	62.7
TCGA-BH-A0B5	TCGA-BH-A0B5-01A-11R-A12P-07	Tumor	5.9	62.5
TCGA-BH-A0B5	TCGA-BH-A0B5-11A-23R-A12P-07	Normal	6.9	59.2
TCGA-BH-A0B7	TCGA-BH-A0B7-01A-12R-A115-07	Tumor	6.0	64.0
TCGA-BH-A0B7	TCGA-BH-A0B7-11A-34R-A115-07	Normal	5.3	51.0
TCGA-BH-A0B8	TCGA-BH-A0B8-01A-21R-A056-07	Tumor	4.7	49.3
TCGA-BH-A0B8	TCGA-BH-A0B8-11A-41R-A089-07	Normal	7.1	55.5
TCGA-BH-A0BC	TCGA-BH-A0BC-01A-22R-A084-07	Tumor	6.7	75.7
TCGA-BH-A0BC	TCGA-BH-A0BC-11A-22R-A089-07	Normal	6.3	52.1
TCGA-BH-A0BJ	TCGA-BH-A0BJ-01A-11R-A056-07	Tumor	5.7	61.4
TCGA-BH-A0BJ	TCGA-BH-A0BJ-11A-23R-A089-07	Normal	6.6	55.2
TCGA-BH-A0BM	TCGA-BH-A0BM-01A-11R-A056-07	Tumor	3.7	37.2
TCGA-BH-A0BM	TCGA-BH-A0BM-11A-12R-A089-07	Normal	5.7	54.6
TCGA-BH-A0BQ	TCGA-BH-A0BQ-01A-21R-A115-07	Tumor	4.0	50.3
TCGA-BH-A0BQ	TCGA-BH-A0BQ-11A-33R-A115-07	Normal	5.0	50.6

Continued on next page

Table C.1 -- *Continued from previous page*

Patient ID	Sample ID	Tumor/Normal	size(GB)	#Reads(M)
TCGA-BH-A0BS	TCGA-BH-A0BS-01A-11R-A12P-07	Tumor	6.9	68.4
TCGA-BH-A0BS	TCGA-BH-A0BS-11A-11R-A12P-07	Normal	6.3	63.8
TCGA-BH-A0BT	TCGA-BH-A0BT-01A-11R-A12P-07	Tumor	6.5	59.2
TCGA-BH-A0BT	TCGA-BH-A0BT-11A-21R-A12P-07	Normal	5.2	55.6
TCGA-BH-A0BV	TCGA-BH-A0BV-01A-11R-A00Z-07	Tumor	5.1	57.0
TCGA-BH-A0BV	TCGA-BH-A0BV-11A-31R-A089-07	Normal	8.0	74.2
TCGA-BH-A0BW	TCGA-BH-A0BW-01A-11R-A115-07	Tumor	6.1	59.2
TCGA-BH-A0BW	TCGA-BH-A0BW-11A-12R-A115-07	Normal	5.8	52.2
TCGA-BH-A0C0	TCGA-BH-A0C0-01A-21R-A056-07	Tumor	4.1	50.1
TCGA-BH-A0C0	TCGA-BH-A0C0-11A-21R-A089-07	Normal	7.8	72.1
TCGA-BH-A0C3	TCGA-BH-A0C3-01A-21R-A12P-07	Tumor	5.1	52.0
TCGA-BH-A0C3	TCGA-BH-A0C3-11A-23R-A12P-07	Normal	5.7	57.5
TCGA-BH-A0DD	TCGA-BH-A0DD-01A-31R-A12P-07	Tumor	5.9	60.5
TCGA-BH-A0DD	TCGA-BH-A0DD-11A-23R-A12P-07	Normal	5.5	57.2
TCGA-BH-A0DG	TCGA-BH-A0DG-01A-21R-A12P-07	Tumor	5.1	55.0
TCGA-BH-A0DG	TCGA-BH-A0DG-11A-43R-A12P-07	Normal	7.7	67.9
TCGA-BH-A0DH	TCGA-BH-A0DH-01A-11R-A084-07	Tumor	5.6	60.1
TCGA-BH-A0DH	TCGA-BH-A0DH-11A-31R-A089-07	Normal	6.2	60.5
TCGA-BH-A0DK	TCGA-BH-A0DK-01A-21R-A056-07	Tumor	4.8	53.6
TCGA-BH-A0DK	TCGA-BH-A0DK-11A-13R-A089-07	Normal	7.8	76.4
TCGA-BH-A0DO	TCGA-BH-A0DO-01B-11R-A12D-07	Tumor	6.2	61.1
TCGA-BH-A0DO	TCGA-BH-A0DO-11A-22R-A12D-07	Normal	6.4	60.3
TCGA-BH-A0DP	TCGA-BH-A0DP-01A-21R-A056-07	Tumor	4.5	53.6
TCGA-BH-A0DP	TCGA-BH-A0DP-11A-12R-A089-07	Normal	6.7	65.7
TCGA-BH-A0DQ	TCGA-BH-A0DQ-01A-11R-A084-07	Tumor	5.3	54.9
TCGA-BH-A0DQ	TCGA-BH-A0DQ-11A-12R-A089-07	Normal	8.6	79.3
TCGA-BH-A0DT	TCGA-BH-A0DT-01A-21R-A12D-07	Tumor	5.6	56.3
TCGA-BH-A0DT	TCGA-BH-A0DT-11A-12R-A12D-07	Normal	6.2	58.5
TCGA-BH-A0DV	TCGA-BH-A0DV-01A-21R-A12P-07	Tumor	7.0	80.8
TCGA-BH-A0DV	TCGA-BH-A0DV-11A-22R-A12P-07	Normal	7.5	66.0
TCGA-BH-A0E0	TCGA-BH-A0E0-01A-11R-A056-07	Tumor	3.5	39.5
TCGA-BH-A0E0	TCGA-BH-A0E0-11A-13R-A089-07	Normal	5.4	53.8
TCGA-BH-A0E1	TCGA-BH-A0E1-01A-11R-A056-07	Tumor	4.6	57.6
TCGA-BH-A0E1	TCGA-BH-A0E1-11A-13R-A089-07	Normal	4.9	49.8
TCGA-BH-A0H7	TCGA-BH-A0H7-01A-13R-A056-07	Tumor	7.5	76.7
TCGA-BH-A0H7	TCGA-BH-A0H7-11A-13R-A089-07	Normal	5.9	59.4
TCGA-BH-A0HA	TCGA-BH-A0HA-01A-11R-A12P-07	Tumor	5.5	54.1
TCGA-BH-A0HA	TCGA-BH-A0HA-11A-31R-A12P-07	Normal	5.9	56.0
TCGA-BH-A0HK	TCGA-BH-A0HK-01A-11R-A056-07	Tumor	4.6	45.9
TCGA-BH-A0HK	TCGA-BH-A0HK-11A-11R-A089-07	Normal	6.2	61.8
TCGA-BH-A18J	TCGA-BH-A18J-01A-11R-A12D-07	Tumor	6.4	66.0
TCGA-BH-A18J	TCGA-BH-A18J-11A-31R-A12D-07	Normal	4.8	47.1
TCGA-BH-A18K	TCGA-BH-A18K-01A-11R-A12D-07	Tumor	6.8	63.5
TCGA-BH-A18K	TCGA-BH-A18K-11A-13R-A12D-07	Normal	4.8	45.7
TCGA-BH-A18L	TCGA-BH-A18L-01A-32R-A12D-07	Tumor	7.1	73.3
TCGA-BH-A18L	TCGA-BH-A18L-11A-42R-A12D-07	Normal	5.1	52.3
TCGA-BH-A18M	TCGA-BH-A18M-01A-11R-A12D-07	Tumor	5.8	61.2
TCGA-BH-A18M	TCGA-BH-A18M-11A-33R-A12D-07	Normal	4.9	49.5
TCGA-BH-A18N	TCGA-BH-A18N-01A-11R-A12D-07	Tumor	6.0	61.5
TCGA-BH-A18N	TCGA-BH-A18N-11A-43R-A12D-07	Normal	6.8	65.0
TCGA-BH-A18P	TCGA-BH-A18P-01A-11R-A12D-07	Tumor	5.1	56.8
TCGA-BH-A18P	TCGA-BH-A18P-11A-43R-A12D-07	Normal	6.2	54.7
TCGA-BH-A18Q	TCGA-BH-A18Q-01A-12R-A12D-07	Tumor	5.3	52.1
TCGA-BH-A18Q	TCGA-BH-A18Q-11A-34R-A12D-07	Normal	7.2	67.7
TCGA-BH-A18R	TCGA-BH-A18R-01A-11R-A12D-07	Tumor	8.1	76.2
TCGA-BH-A18R	TCGA-BH-A18R-11A-42R-A12D-07	Normal	7.5	70.5
TCGA-BH-A18S	TCGA-BH-A18S-01A-11R-A12D-07	Tumor	6.3	66.0
TCGA-BH-A18S	TCGA-BH-A18S-11A-43R-A12D-07	Normal	4.6	45.1
TCGA-BH-A18U	TCGA-BH-A18U-01A-21R-A12D-07	Tumor	6.4	67.8
TCGA-BH-A18U	TCGA-BH-A18U-11A-23R-A12D-07	Normal	4.5	46.0

Continued on next page

Table C.1 -- *Continued from previous page*

Patient ID	Sample ID	Tumor/Normal	size(GB)	#Reads(M)
TCGA-BH-A18V	TCGA-BH-A18V-01A-11R-A12D-07	Tumor	6.7	68.0
TCGA-BH-A18V	TCGA-BH-A18V-11A-52R-A12D-07	Normal	5.2	52.2
TCGA-BH-A1EN	TCGA-BH-A1EN-01A-11R-A13Q-07	Tumor	7.0	73.1
TCGA-BH-A1EN	TCGA-BH-A1EN-11A-23R-A13Q-07	Normal	8.5	79.1
TCGA-BH-A1ET	TCGA-BH-A1ET-01A-11R-A137-07	Tumor	10.4	98.8
TCGA-BH-A1ET	TCGA-BH-A1ET-11B-23R-A137-07	Normal	9.1	82.4
TCGA-BH-A1EU	TCGA-BH-A1EU-01A-11R-A137-07	Tumor	11.9	105.9
TCGA-BH-A1EU	TCGA-BH-A1EU-11A-23R-A137-07	Normal	3.7	35.2
TCGA-BH-A1EV	TCGA-BH-A1EV-01A-11R-A137-07	Tumor	6.1	61.8
TCGA-BH-A1EV	TCGA-BH-A1EV-11A-24R-A137-07	Normal	10.0	90.0
TCGA-BH-A1F0	TCGA-BH-A1F0-01A-11R-A137-07	Tumor	12.0	106.9
TCGA-BH-A1F0	TCGA-BH-A1F0-11B-23R-A137-07	Normal	7.7	73.2
TCGA-BH-A1F2	TCGA-BH-A1F2-01A-31R-A13Q-07	Tumor	8.0	76.6
TCGA-BH-A1F2	TCGA-BH-A1F2-11A-32R-A13Q-07	Normal	9.3	82.3
TCGA-BH-A1F8	TCGA-BH-A1F8-01A-11R-A13Q-07	Tumor	8.5	81.4
TCGA-BH-A1F8	TCGA-BH-A1F8-11B-21R-A13Q-07	Normal	3.6	38.6
TCGA-BH-A1FC	TCGA-BH-A1FC-01A-11R-A13Q-07	Tumor	8.4	83.1
TCGA-BH-A1FC	TCGA-BH-A1FC-11A-32R-A13Q-07	Normal	3.5	36.8
TCGA-BH-A1FD	TCGA-BH-A1FD-01A-11R-A13Q-07	Tumor	8.0	77.0
TCGA-BH-A1FD	TCGA-BH-A1FD-11B-21R-A13Q-07	Normal	8.2	67.9
TCGA-BH-A1FE	TCGA-BH-A1FE-01A-11R-A13Q-07	Tumor	8.4	83.5
TCGA-BH-A1FE	TCGA-BH-A1FE-11B-14R-A13Q-07	Normal	8.5	66.0
TCGA-BH-A1FG	TCGA-BH-A1FG-01A-11R-A13Q-07	Tumor	6.5	62.9
TCGA-BH-A1FG	TCGA-BH-A1FG-11B-12R-A13Q-07	Normal	7.2	56.4
TCGA-BH-A1FH	TCGA-BH-A1FH-01A-12R-A13Q-07	Tumor	8.8	75.5
TCGA-BH-A1FH	TCGA-BH-A1FH-11B-42R-A13Q-07	Normal	4.0	44.0
TCGA-BH-A1FJ	TCGA-BH-A1FJ-01A-11R-A13Q-07	Tumor	8.7	77.2
TCGA-BH-A1FJ	TCGA-BH-A1FJ-11B-42R-A13Q-07	Normal	10.5	87.3
TCGA-BH-A1FM	TCGA-BH-A1FM-01A-11R-A13Q-07	Tumor	8.1	67.4
TCGA-BH-A1FM	TCGA-BH-A1FM-11B-23R-A13Q-07	Normal	8.0	72.4
TCGA-BH-A1FN	TCGA-BH-A1FN-01A-11R-A13Q-07	Tumor	4.9	47.1
TCGA-BH-A1FN	TCGA-BH-A1FN-11A-34R-A13Q-07	Normal	8.5	82.0
TCGA-BH-A1FU	TCGA-BH-A1FU-01A-11R-A14D-07	Tumor	5.3	52.5
TCGA-BH-A1FU	TCGA-BH-A1FU-11A-23R-A14D-07	Normal	4.0	39.6
TCGA-BH-A203	TCGA-BH-A203-01A-12R-A169-07	Tumor	8.0	68.5
TCGA-BH-A203	TCGA-BH-A203-11A-42R-A169-07	Normal	7.9	61.7
TCGA-BH-A204	TCGA-BH-A204-01A-11R-A157-07	Tumor	10.8	101.1
TCGA-BH-A204	TCGA-BH-A204-11A-53R-A157-07	Normal	7.7	66.6
TCGA-BH-A208	TCGA-BH-A208-01A-11R-A157-07	Tumor	6.5	67.1
TCGA-BH-A208	TCGA-BH-A208-11A-51R-A157-07	Normal	7.8	73.2
TCGA-BH-A209	TCGA-BH-A209-01A-11R-A157-07	Tumor	8.6	80.1
TCGA-BH-A209	TCGA-BH-A209-11A-42R-A157-07	Normal	7.6	72.0
TCGA-E2-A153	TCGA-E2-A153-01A-12R-A12D-07	Tumor	4.0	44.7
TCGA-E2-A153	TCGA-E2-A153-11A-31R-A12D-07	Normal	4.9	47.7
TCGA-E2-A15I	TCGA-E2-A15I-01A-21R-A137-07	Tumor	6.7	60.2
TCGA-E2-A15I	TCGA-E2-A15I-11A-32R-A137-07	Normal	5.3	46.4
TCGA-E2-A1IG	TCGA-E2-A1IG-01A-11R-A144-07	Tumor	5.6	56.4
TCGA-E2-A1IG	TCGA-E2-A1IG-11A-22R-A144-07	Normal	4.4	44.1
TCGA-E2-A1L7	TCGA-E2-A1L7-01A-11R-A144-07	Tumor	4.4	45.3
TCGA-E2-A1L7	TCGA-E2-A1L7-11A-33R-A144-07	Normal	5.6	54.7
TCGA-E2-A1LB	TCGA-E2-A1LB-01A-11R-A144-07	Tumor	5.4	54.3
TCGA-E2-A1LB	TCGA-E2-A1LB-11A-22R-A144-07	Normal	5.8	55.5
TCGA-E2-A1LH	TCGA-E2-A1LH-01A-11R-A14D-07	Tumor	3.8	34.9
TCGA-E2-A1LH	TCGA-E2-A1LH-11A-22R-A14D-07	Normal	4.3	44.1
TCGA-E2-A1LS	TCGA-E2-A1LS-01A-12R-A157-07	Tumor	7.8	55.8
TCGA-E2-A1LS	TCGA-E2-A1LS-11A-32R-A157-07	Normal	8.2	67.7
TCGA-E9-A1N4	TCGA-E9-A1N4-01A-11R-A14M-07	Tumor	9.6	93.1
TCGA-E9-A1N4	TCGA-E9-A1N4-11A-33R-A14M-07	Normal	8.2	72.4
TCGA-E9-A1N5	TCGA-E9-A1N5-01A-11R-A14D-07	Tumor	4.3	42.9
TCGA-E9-A1N5	TCGA-E9-A1N5-11A-41R-A14D-07	Normal	3.2	32.3

Continued on next page

Table C.1 -- *Continued from previous page*

Patient ID	Sample ID	Tumor/Normal	size(GB)	#Reads(M)
TCGA-E9-A1N6	TCGA-E9-A1N6-01A-11R-A144-07	Tumor	5.0	50.0
TCGA-E9-A1N6	TCGA-E9-A1N6-11A-32R-A144-07	Normal	5.6	49.2
TCGA-E9-A1N9	TCGA-E9-A1N9-01A-11R-A14D-07	Tumor	5.9	48.8
TCGA-E9-A1N9	TCGA-E9-A1N9-11A-71R-A14D-07	Normal	7.6	58.8
TCGA-E9-A1NA	TCGA-E9-A1NA-01A-11R-A144-07	Tumor	7.1	76.3
TCGA-E9-A1NA	TCGA-E9-A1NA-11A-33R-A144-07	Normal	5.8	59.4
TCGA-E9-A1ND	TCGA-E9-A1ND-01A-11R-A144-07	Tumor	5.0	48.2
TCGA-E9-A1ND	TCGA-E9-A1ND-11A-43R-A144-07	Normal	5.0	52.3
TCGA-E9-A1NF	TCGA-E9-A1NF-01A-11R-A14D-07	Tumor	5.3	45.1
TCGA-E9-A1NF	TCGA-E9-A1NF-11A-73R-A14D-07	Normal	6.5	53.2
TCGA-E9-A1NG	TCGA-E9-A1NG-01A-21R-A14M-07	Tumor	7.6	79.9
TCGA-E9-A1NG	TCGA-E9-A1NG-11A-52R-A14M-07	Normal	5.2	52.6
TCGA-E9-A1R7	TCGA-E9-A1R7-01A-11R-A14M-07	Tumor	8.1	63.3
TCGA-E9-A1R7	TCGA-E9-A1R7-11A-42R-A14M-07	Normal	8.0	60.7
TCGA-E9-A1RB	TCGA-E9-A1RB-01A-11R-A157-07	Tumor	9.0	87.9
TCGA-E9-A1RB	TCGA-E9-A1RB-11A-33R-A157-07	Normal	8.3	77.2
TCGA-E9-A1RC	TCGA-E9-A1RC-01A-11R-A157-07	Tumor	9.6	92.4
TCGA-E9-A1RC	TCGA-E9-A1RC-11A-33R-A157-07	Normal	8.0	73.5
TCGA-E9-A1RD	TCGA-E9-A1RD-01A-11R-A157-07	Tumor	9.1	86.7
TCGA-E9-A1RD	TCGA-E9-A1RD-11A-33R-A157-07	Normal	8.4	78.4
TCGA-E9-A1RF	TCGA-E9-A1RF-01A-11R-A157-07	Tumor	7.6	71.6
TCGA-E9-A1RF	TCGA-E9-A1RF-11A-32R-A157-07	Normal	6.7	60.7
TCGA-E9-A1RH	TCGA-E9-A1RH-01A-21R-A169-07	Tumor	8.7	77.5
TCGA-E9-A1RH	TCGA-E9-A1RH-11A-34R-A169-07	Normal	7.8	59.7
TCGA-E9-A1RI	TCGA-E9-A1RI-01A-11R-A169-07	Tumor	8.5	70.3
TCGA-E9-A1RI	TCGA-E9-A1RI-11A-41R-A169-07	Normal	8.0	67.9
TCGA-GI-A2C8	TCGA-GI-A2C8-01A-11R-A16F-07	Tumor	7.6	72.2
TCGA-GI-A2C8	TCGA-GI-A2C8-11A-22R-A16F-07	Normal	8.3	81.5
TCGA-A7-A13E	TCGA-A7-A13E-01A-11R-A12P-07	Tumor	5.4	50.7
TCGA-A7-A13E	TCGA-A7-A13E-11A-61R-A12P-07	Normal	5.4	48.8
TCGA-A7-A13F	TCGA-A7-A13F-01A-11R-A12P-07	Tumor	5.5	57.0
TCGA-A7-A13F	TCGA-A7-A13F-11A-42R-A12P-07	Normal	5.8	52.4
TCGA-BH-A0AU	TCGA-BH-A0AU-01A-11R-A12P-07	Tumor	5.3	52.1
TCGA-BH-A0AU	TCGA-BH-A0AU-11A-11R-A12P-07	Normal	5.9	53.6
TCGA-BH-A0BZ	TCGA-BH-A0BZ-01A-31R-A12P-07	Tumor	7.0	75.3
TCGA-BH-A0BZ	TCGA-BH-A0BZ-11A-61R-A12P-07	Normal	6.2	62.8
TCGA-BH-A0DL	TCGA-BH-A0DL-01A-11R-A115-07	Tumor	8.6	78.0
TCGA-BH-A0DL	TCGA-BH-A0DL-11A-13R-A115-07	Normal	7.9	71.7
TCGA-BH-A0H5	TCGA-BH-A0H5-01A-21R-A115-07	Tumor	5.9	63.0
TCGA-BH-A0H5	TCGA-BH-A0H5-11A-62R-A115-07	Normal	5.4	52.5
TCGA-BH-A1EO	TCGA-BH-A1EO-01A-11R-A137-07	Tumor	3.5	39.8
TCGA-BH-A1EO	TCGA-BH-A1EO-11A-31R-A137-07	Normal	10.6	105.2
TCGA-BH-A1EW	TCGA-BH-A1EW-01A-11R-A137-07	Tumor	5.5	58.1
TCGA-BH-A1EW	TCGA-BH-A1EW-11B-33R-A137-07	Normal	11.0	99.2
TCGA-BH-A1F6	TCGA-BH-A1F6-01A-11R-A13Q-07	Tumor	4.7	48.7
TCGA-BH-A1F6	TCGA-BH-A1F6-11B-94R-A13Q-07	Normal	8.8	78.2
TCGA-BH-A1FB	TCGA-BH-A1FB-01A-11R-A13Q-07	Tumor	9.2	76.5
TCGA-BH-A1FB	TCGA-BH-A1FB-11A-33R-A13Q-07	Normal	4.6	44.4
TCGA-BH-A1FR	TCGA-BH-A1FR-01A-11R-A13Q-07	Tumor	4.0	43.5
TCGA-BH-A1FR	TCGA-BH-A1FR-11B-42R-A13Q-07	Normal	4.8	44.1
TCGA-E2-A158	TCGA-E2-A158-01A-11R-A12D-07	Tumor	7.3	67.8
TCGA-E2-A158	TCGA-E2-A158-11A-22R-A12D-07	Normal	4.0	39.0
TCGA-E2-A15M	TCGA-E2-A15M-01A-11R-A12D-07	Tumor	12.9	107.7
TCGA-E2-A15M	TCGA-E2-A15M-11A-22R-A12D-07	Normal	4.8	49.3

Table C.2: The first page of the list of 5088 genes with AAR in at least 10 patients identified in Chapter 4. The 1st column lists the genomic coordinates of AAR, the 2nd (4th) column lists the number of patients with the higher (lower) AAR inclusion ration in tumor than normal. The 3rd and 5th columns give the average inclusion ratio difference between tumor and normal samples in the patients. The last column lists the gene name provided in Refseq database. The table was sorted by the maximum value of column 2 and 4.

region	# IT>IN	ave(IT-IN)	# IT<IN	ave(IT-IN)	RefGene name
chr14:69345175-69345240	0	0.00	103	-0.30	ACTN1;
chr3:13663275-13663415	0	0.00	102	-0.49	FBLN2;
chr9:124043748-124043840	96	0.53	2	-0.14	GSN;
chr15:74466087-74466360	2	0.36	92	-0.39	ISLR;
chr2:174123427-174123543	92	0.25	0	0.00	ZAK;
chr10:93000241-93000337	1	0.13	91	-0.30	PCGF5;
chr5:33751303-33751508	0	0.00	91	-0.40	ADAMTS12;
chr9:123631853-123632122	90	0.30	1	-0.15	PHF19;
chr1:207963598-207963690	0	0.00	88	-0.26	CD46;
chr6:56507420-56507694	7	0.21	88	-0.36	DST;
chr2:238678586-238678635	1	0.14	87	-0.28	LRRFIP1;
chr9:117808689-117808961	86	0.38	3	-0.15	TNC;
chr3:37132958-37133029	2	0.16	84	-0.37	LRRFIP2;
chr3:57911572-57911661	2	0.19	84	-0.28	SLMAP;
chr14:73745989-73746132	84	0.33	4	-0.19	NUMB;AX747833;
chr12:56558153-56558431	84	0.29	0	0.00	SMARCC2;
chr2:64069014-64069338	1	0.14	82	-0.28	UGP2;
chr5:38445578-38445780	3	0.24	82	-0.37	EGFLAM;
chr19:49605431-49605442	8	0.15	81	-0.24	SNRNP70;
chr15:89422649-89423841	1	0.16	81	-0.23	HAPLN3;
chr15:64429767-64430240	0	0.00	81	-0.33	SNX1;
chrX:102942917-102943086	0	0.00	81	-0.23	MORF4L2;
chr17:48828003-48828055	81	0.28	4	-0.14	LUC7L3;
chr8:95470496-95470664	80	0.41	0	0.00	RAD54B;
chr11:131240371-131240783	0	0.00	79	-0.49	NTM;
chrX:154124352-154124507	3	0.17	79	-0.22	F8;
chr14:36157665-36157746	3	0.26	78	-0.36	RALGAPA1;
chr2:173366500-173366629	3	0.22	78	-0.27	ITGA6;
chr12:56554410-56554454	9	0.14	77	-0.26	MYL6;
chr6:42016239-42016610	4	0.22	77	-0.37	CCND3;
chr2:120885264-120885427	2	0.23	77	-0.38	EPB41L5;
chr7:138738711-138738841	1	0.22	77	-0.25	ZC3HAV1;
chr10:105770574-105770666	77	0.43	7	-0.16	SLK;
chr1:51435642-51436029	1	0.10	77	-0.22	CDKN2C;
chr3:100549422-100549478	0	0.00	77	-0.27	ABI3BP;
chr5:134686517-134686603	1	0.17	77	-0.23	H2AFY;AX747819;
chr1:25570125-25570715	4	0.23	76	-0.32	C1orf63;
chr2:161993466-161993574	3	0.33	76	-0.43	TANK;
chr5:177635540-177635916	3	0.17	76	-0.25	AGXT2L2;
chr11:47493655-47493742	1	0.13	76	-0.22	CELF1;
chr11:85342189-85342360	0	0.00	76	-0.26	TMEM126B;
chr1:54723742-54723822	76	0.25	1	-0.20	SSBP3;
chr1:155294368-155294378	76	0.24	1	-0.25	RUSC1;
chr3:123401071-123401157	76	0.32	14	-0.22	MYLK;
chrX:15843929-15845495	4	0.16	75	-0.29	AP1S2;
chr10:95152674-95152712	3	0.23	75	-0.29	MYOF;
chr11:1874200-1874427	75	0.43	4	-0.13	LSP1;
chr8:103032465-103032517	75	0.38	2	-0.24	NCALD;
chr13:28891635-28891734	0	0.00	74	-0.20	FLT1;BC048278;
chr9:116353613-116353677	74	0.33	6	-0.19	RGS3;
chr15:63353397-63353472	74	0.31	4	-0.16	TPM1;AK055197;
chr3:37402734-37402796	74	0.27	2	-0.33	GOLGA4;
chr11:111835273-111835402	74	0.33	4	-0.13	DIXDC1;
chr8:15977927-15978115	0	0.00	74	-0.23	MSR1;
chr17:49053224-49053262	2	0.19	73	-0.41	SPAG9;
chr11:82745315-82745606	0	0.00	73	-0.25	RAB30;
chr17:30693684-30693776	0	0.00	73	-0.24	ZNF207;MIR632;
chr19:7150508-7150543	0	0.00	73	-0.34	INSR;

Bibliography

[1] Wen, J., Chiba, A., and Cai, X., "Computational identification of tissue-specific alternative splicing elements in mouse genes from RNA-Seq," *Nucleic Acids Res*, Vol. 38, No. 22, 2010, pp. 7895–7907.

[2] Wen, J., Chen, Z., and Cai, X., "A biophysical model for identifying splicing regulatory elements and their Interactions," *PLoS ONE*, Vol. 8, No. 1, 01 2013, pp. e54885.

[3] Crick, F., "Central dogma of molecular biology," *Nature*, Vol. 227, No. 5258, Aug. 1970, pp. 561–563.

[4] Consortium, M. G. S., "Initial sequencing and comparative analysis of the mouse genome," *Nature*, Vol. 420, No. 6915, Dec. 2002, pp. 520–562.

[5] Nilsen, T. W. and Graveley, B. R., "Expansion of the eukaryotic proteome by alternative splicing," *Nature*, Vol. 463, No. 7280, Jan. 2010, pp. 457–463.

[6] Wahl, M. C., Will, C. L., and Lührmann, R., "The spliceosome: design principles of a dynamic RNP Machine," *Cell*, Vol. 136, No. 4, 2009, pp. 701–718.

[7] Ast, G., "How did alternative splicing evolve?" *Nat Rev Genet*, Vol. 5, No. 10, Oct. 2004, pp. 773–782.

[8] Chasin, L. A., "Searching for splicing motifs," *Adv. Exp. Med. Biol.*, Vol. 623, 2007, pp. 85–106.

[9] Wang, Z. and Burge, C. B., "Splicing regulation: from a parts list of regulatory elements to an integrated splicing code," *RNA*, Vol. 14, No. 5, 2008, pp. 802–813.

[10] Matlin, A. J., Clark, F., and Smith, C. W., "Understanding alternative splicing: Towards a cellular code." *Nat Rev Mol Cell Biol*, Vol. 6, No. 5, May 2005, pp. 386–398.

[11] Izquierdo, J. M., Majos, N., Bonnal, S., Martinez, C., Castelo, R., Guigó, R., Bilbao, D., and Valcárcel, J., "Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition," *Mol Cell*, Vol. 19, No. 4, Aug. 2005, pp. 475–484.

[12] Chen, M. and Manley, J. L., “Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches,” *Nat Rev Mol Cell Biol*, Vol. 10, No. 11, September 2009, pp. 741–754.

[13] Sharma, S., Kohlstaedt, L. A., Damianov, A., Rio, D. C., and Black, D. L., “Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome,” *Nat Struct Mol Biol*, Vol. 15, No. 2, Jan. 2008, pp. 183–191.

[14] Wang, Z., Gerstein, M., and Snyder, M., “RNA-Seq: a revolutionary tool for transcriptomics.” *Nat Rev Genet*, Vol. 10, No. 1, January 2009, pp. 57–63.

[15] Cuperlovic-Culf, M., Belacel, N., Culf, A. S., and Ouellette, R. J., “Microarray analysis of alternative splicing.” *OMICS*, Vol. 10, No. 3, 2006, pp. 344–357.

[16] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B., “Mapping and quantifying mammalian transcriptomes by RNA-Seq,” *Nat. Methods*, Vol. 5, No. 7, May 2008, pp. 621–628.

[17] Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B., “Alternative isoform regulation in human tissue transcriptomes,” *Nature*, Vol. 456, No. 7221, November 2008, pp. 470–476.

[18] Jiang, H. and Wong, W. H., “Statistical inferences for isoform expression in RNA-Seq,” *Bioinformatics*, Vol. 25, No. 8, 2009, pp. 1026–1032.

[19] Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J., “Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing,” *Nat Genet*, Vol. 40, No. 12, Nov. 2008, pp. 1413–1415.

[20] Cooper, T. A., Wan, L., and Dreyfuss, G., “RNA and disease.” *Cell*, Vol. 136, No. 4, Feb. 2009, pp. 777–793.

[21] David, C. J. and Manley, J. L., “Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged,” *Genes Dev*, Vol. 24, No. 21, 2010, pp. 2343–2364.

[22] Djordjevic, M., “SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways,” *Biomol. Eng.*, Vol. 24, No. 2, 2007, pp. 179–189.

[23] Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R. B., “CLIP identifies nova-regulated RNA networks in the brain,” *Science*, Vol. 302, No. 5648, 2003, pp. 1212–1215.

[24] Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M., and Burge, C. B., “Systematic identification and analysis of exonic splicing silencers,” *Cell*, Vol. 119, No. 6, 2004, pp. 831–845.

[25] Witten, J. T. and Ule, J., "Understanding splicing regulation through RNA splicing maps," *Trends Genet*, Vol. 27, No. 3, 2011, pp. 89–97.

[26] Network, T. C. G. A. R., "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, Vol. 455, No. 7216, Sept. 2008, pp. 1061–1068.

[27] Fairbrother, W. G., Yeh, R.-F., Sharp, P. A., and Burge, C. B., "Predictive identification of exonic splicing enhancers in human genes," *Science*, Vol. 297, No. 5583, 2002, pp. 1007–1013.

[28] Zhang, X. H. and Chasin, L. A., "Computational definition of sequence motifs governing constitutive exon splicing," *Genes Dev.*, Vol. 18, No. 11, June 2004, pp. 1241–1250.

[29] Sironi, M., Menozzi, G., Riva, L., Cagliani, R., Comi, G. P., Bresolin, N., Giorda, R., and Pozzoli, U., "Silencer elements as possible inhibitors of pseudoexon splicing," *Nucleic Acids Res.*, Vol. 32, No. 5, 2004, pp. 1783–1791.

[30] Castle, J. C., Zhang, C., Shah, J. K., Kulkarni, A. V., Kalsotra, A., Cooper, T. A., and Johnson, J. M., "Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines," *Nat Genet*, Vol. 40, No. 12, December 2008, pp. 1416–1425.

[31] Hartmann, B. and Valcrcel, J., "Decrypting the genome's alternative messages," *Curr. Opin. Cell Biol.*, Vol. 21, No. 3, 2009, pp. 377–386.

[32] Brudno, M., Gelfand, M. S., Spengler, S., Zorn, M., Dubchak, I., and Conboy, J. G., "Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing," *Nucleic Acids Res.*, Vol. 29, No. 11, 2001, pp. 2338–2348.

[33] Wang, X., Wang, K., Radovich, M., Wang, Y., Wang, G., Feng, W., Sanford, J., and Liu, Y., "Genome-wide prediction of cis-acting RNA elements regulating tissue-specific pre-mRNA alternative splicing," *BMC Genomics*, Vol. 10, No. Suppl 1, 2009, pp. S4.

[34] Sinha, S., "Discriminative motifs," *J. Comput. Biol.*, Vol. 10, No. 3-4, 2003, pp. 599–615.

[35] Smith, A. D., Sumazin, P., and Zhang, M. Q., "Identifying tissue-selective transcription factor binding sites in vertebrate promoters," *Proc Natl Acad Sci U S A*, Vol. 102, No. 5, 2005, pp. 1560–1565.

[36] Redhead, E. and Bailey, T., "Discriminative motif discovery in DNA and protein sequences using the DEME algorithm," *BMC Bioinformatics*, Vol. 8, No. 1, 2007, pp. 385.

[37] Kim, E., Goren, A., and Ast, G., "Alternative splicing: current perspectives," *BioEssays*, Vol. 30, No. 1, 2008, pp. 38–47.

[38] Piva, F., Giulietti, M., Nocchi, L., and Principato, G., "SpliceAid: a database of experimental RNA target motifs bound by splicing proteins in humans," *Bioinformatics*, Vol. 25, No. 9, 2009, pp. 1211–1213.

[39] Holste, D. and Ohler, U., "Strategies for identifying RNA splicing regulatory motifs and predicting alternative splicing events," *PLoS Comput Biol*, Vol. 4, No. 1, 01 2008, pp. e21.

[40] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S., "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol*, Vol. 10, No. 3, 2009, pp. R25.

[41] Sun, H. and Chasin, L. A., "Multiple splicing defects in an intronic false exon," *Mol Cell Biol*, Vol. 20, No. 17, 2000, pp. 6414–6425.

[42] Zar, J. H., *Biostatistical analysis (4th Edition)*, Prentice Hall, October 1998.

[43] Chan, R. C. and Black, D. L., "Conserved intron elements repress splicing of a neuron-specific c-src exon in vitro," *Mol Cell Biol*, Vol. 15, No. 11, 1995, pp. 6377–6385.

[44] Singh, R., Valcarcel, J., and Green, M. R., "Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins," *Science*, Vol. 268, No. 5214, 1995, pp. 1173–1176.

[45] Kino, Y., Mori, D., Oma, Y., Takeshita, Y., Sasagawa, N., and Ishiura, S., "Muscleblind protein, MBNL1/EXP, binds specifically to CHHG repeats," *Hum. Mol. Genet.*, Vol. 13, No. 5, 2004, pp. 495–507.

[46] Kabat, J. L., Barberan-Soler, S., McKenna, P., Clawson, H., Farrer, T., and Zahler, A. M., "Intronic alternative splicing regulators identified by comparative genomics in nematodes," *PLoS Comput Biol*, Vol. 2, No. 7, 07 2006, pp. e86.

[47] Kabat, J. L., Barberan-Soler, S., and Zahler, A. M., "HRP-2, the c. elegans homolog of mammalian heterogeneous nuclear ribonucleoproteins Q and R, is an alternative splicing factor that binds to UCUAUC splicing regulatory elements," *J Biol Chem*, Vol. 284, No. 42, 2009, pp. 28490–28497.

[48] Spellman, R. and Smith, C. W., "Novel modes of splicing repression by PTB," *Trends Biochem. Sci.*, Vol. 31, No. 2, 2006, pp. 73 – 76.

[49] Sauliere, J., Sureau, A., Expert-Bezancon, A., and Marie, J., "The polypyrimidine tract binding protein (PTB) represses splicing of exon 6B from the β-tropomyosin pre-mRNA by directly interfering with the binding of the U2AF65 subunit," *Mol Cell Biol*, Vol. 26, No. 23, 2006, pp. 8755–8769.

[50] Minovitsky, S., Gee, S. L., Schokrpur, S., Dubchak, I., and Conboy, J. G., "The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons," *Nucleic Acids Res.*, Vol. 33, No. 2, 2005, pp. 714–724.

[51] Ponthier, J. L., Schluepen, C., Chen, W., Lersch, R. A., Gee, S. L., Hou, V. C., Lo, A. J., Short, S. A., Chasis, J. A., Winkelmann, J. C., and Conboy, J. G., "Fox-2 splicing factor binds to a conserved intron motif to promote inclusion of protein 4.1R alternative exon 16," *J Biol Chem*, Vol. 281, No. 18, 2006, pp. 12468–12474.

[52] Zhou, H.-L., Baraniak, A. P., and Lou, H., "Role for Fox-1/Fox-2 in mediating the neuronal pathway of calcitonin/calcitonin gene-related peptide alternative RNA processing," *Mol Cell Biol*, Vol. 27, No. 3, 2007, pp. 830–841.

[53] Hui, J., Hung, L. H., Heiner, M., Schreiner, S., Neumuller, N., Reither, G., Haas, S. A., and Bindereif, A., "Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing," *EMBO J.*, Vol. 24, 2005, pp. 1988–1998.

[54] Zuccato, E., Buratti, E., Stuani, C., Baralle, F. E., and Pagani, F., "An intronic polypyrimidine-rich element downstream of the donor site modulates cystic fibrosis transmembrane conductance regulator exon 9 alternative splicing," *J Biol Chem*, Vol. 279, No. 17, 2004, pp. 16980–16988.

[55] Timchenko, N. A., Patel, R., Iakova, P., Cai, Z.-J., Quan, L., and Timchenko, L. T., "Overexpression of CUG triplet repeat-binding protein, CUGBP1, in mice inhibits myogenesis," *J Biol Chem*, Vol. 279, No. 13, 2004, pp. 13129–13139.

[56] Fairbrother, W. G., Yeo, G. W., Yeh, R., Goldstein, P., Mawson, M., Sharp, P. A., and Burge, C. B., "Rescue-ese identifies candidate exonic splicing enhancers in vertebrate exons," *Nucleic Acids Res*, Vol. 32, No. suppl_2, 2004, pp. W187–W190.

[57] Yeo, G., Hoon, S., Venkatesh, B., and Burge, C. B., "Variation in sequence and organization of splicing regulatory elements in vertebrate genes," *Proc Natl Acad Sci U S A*, Vol. 101, No. 44, 2004, pp. 15700–15705.

[58] Sugnet, C. W., Srinivasan, K., Clark, T. A., O'Brien, G., Cline, M. S., Wang, H., Williams, A., Kulp, D., Blume, J. E., Haussler, D., and Ares, Jr., M., "Unusual intron conservation near tissue-regulated exons found by splicing microarrays," *PLoS Comput Biol*, Vol. 2, No. 1, 1 2006, pp. e4.

[59] Graveley, B. R., Hertel, K. J., and Maniatis, T., "A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers," *EMBO J.*, Vol. 17, No. 22, 1998, pp. 6747–6756.

[60] Kechris, K., Yang, Y. H., and Yeh, R.-F., "Prediction of alternatively skipped exons and splicing enhancers from exon junction arrays," *BMC Genomics*, Vol. 9, No. 1, 2008, pp. 551.

[61] Kuroyanagi, H., "Fox-1 family of RNA-binding proteins." *Cell Mol. Life Sci.*, Vol. 66, August 2009, pp. 3895–3907.

[62] Underwood, J. G., Boutz, P. L., Dougherty, J. D., Stoilov, P., and Black, D. L., "Homologues of the caenorhabditis elegans Fox-1 protein are neuronal splicing regulators in mammals," *Mol Cell Biol*, Vol. 25, No. 22, 2005, pp. 10005–10016.

[63] Yeo, G. W., Van Nostrand, E. L., and Liang, T. Y., "Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements," *PLoS Genet*, Vol. 3, No. 5, 05 2007, pp. e85.

[64] Voelker, R. B. and Berglund, J. A., "A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing," *Genome Res*, Vol. 17, No. 7, 2007, pp. 1023–1033.

[65] Das, D., Clark, T. A., Schweitzer, A., Yamamoto, M., Marr, H., Arribere, J., Minovitsky, S., Poliakov, A., Dubchak, I., Blume, J. E., and Conboy, J. G., "A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing," *Nucleic Acids Res*, Vol. 35, No. 14, 2007, pp. 4845–4857.

[66] Ke, S. and Chasin, L., "Intronic motif pairs cooperate across exons to promote pre-mRNA splicing," *Genome Biol*, Vol. 11, No. 8, 2010, pp. R84.

[67] Friedman, B. A., Stadler, M. B., Shomron, N., Ding, Y., and Burge, C. B., "Ab initio identification of functionally interacting pairs of cis-regulatory elements," *Genome Res*, Vol. 18, No. 10, 2008, pp. 1643–1651.

[68] Suyama, M., Harrington, E. D., Vinokourova, S., von Knebel Doeberitz, M., Ohara, O., and Bork, P., "A network of conserved co-occurring motifs for the regulation of alternative splicing," *Nucleic Acids Res*, Vol. 38, No. 22, 2010, pp. 7916–7926.

[69] Bussemaker, H. J., Li, H., and Siggia, E. D., "Regulatory element detection using correlation with expression," *Nat Genet*, Vol. 27, No. 2, Feb. 2001, pp. 167–174.

[70] Conlon, E. M., Liu, X. S., Lieb, J. D., and Liu, J. S., "Integrating regulatory motif discovery and genome-wide expression analysis," *Proc Natl Acad Sci U S A*, Vol. 100, No. 6, 2003, pp. 3339–3344.

[71] Das, D. and Zhang, M. Q., "Predictive models of gene regulation: application of regression methods to microarray data." *Methods Mol Biol*, Vol. 377, 2007, pp. 95–110.

[72] Das, D., Pellegrini, M., and Gray, J. W., "A primer on regression methods for decoding cis-regulatory logic," *PLoS Comput Biol*, Vol. 5, No. 1, 01 2009, pp. e1000269.

[73] Gertz, J., Siggia, E. D., and Cohen, B. A., "Analysis of combinatorial cis-regulation in synthetic and genomic promoters," *Nature*, Vol. 457, No. 7226, 2009, pp. 215–218.

[74] Shea, M. A. and Ackers, G. K., "The or control system of bacteriophage lambda: a physical-chemical model for gene regulation," *Journal of Molecular Biology*, Vol. 181, No. 2, 1985, pp. 211 – 230.

[75] Buchler, N. E., Gerland, U., and Hwa, T., "On schemes of combinatorial transcription logic," *Proc Natl Acad Sci U S A*, Vol. 100, No. 9, 2003, pp. 5136–5141.

[76] Das, D., Banerjee, N., and Zhang, M. Q., "Interacting models of cooperative gene regulation," *Proc Natl Acad Sci U S A*, Vol. 101, No. 46, 2004, pp. 16234–16239.

[77] Tibshirani, R., "The lasso method for variable selection in the cox model," *Stat Med*, Vol. 16, No. 4, 1997, pp. 385–395.

[78] Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D., and Kent, W. J., "The UCSC genome browser database: update 2011," *Nucleic Acids Res*, Vol. 39, No. suppl 1, 2011, pp. D876–D882.

[79] Hsu, F., Kent, W. J., Clawson, H., Kuhn, R. M., Diekhans, M., and Haussler, D., "The UCSC Known Genes," *Bioinformatics*, Vol. 22, No. 9, 2006, pp. 1036–1046.

[80] Fan, J. and Lv, J., "Sure independence screening for ultrahigh dimensional feature space," *J R Stat Soc B*, Vol. 70, No. 5, 2008, pp. 849–911.

[81] Tibshirani, R., "Regression shrinkage and selection via the lasso," *J R Stat Soc B*, Vol. 58, No. 1, 1996, pp. 267–288.

[82] Zou, H., "The adaptive lasso and its oracle properties," *J Am Stat Assoc*, Vol. 101, 2006, pp. 1418–1429.

[83] Yi, N. and Xu, S., "Bayesian lasso for quantitative trait loci mapping," *Genetics*, Vol. 179, No. 2, 2008, pp. 1045–1055.

[84] Han, D., Cai, X., Wen, J., Kenyon, N., and Chen, Z., "From biomarkers to a clue of biology: a computation-aided perspective of immune gene expression profiles in human type 1 diabetes," *Frontiers in Immunology*, Vol. 3, No. 320, 2012.

[85] Gustafsson, M., Hornquist, M., Lundstrom, J., Bjorkegren, J., and Tegner, J., "Reverse engineering of gene networks with lasso and nonlinear basis functions," *Annals of the New York Academy of Sciences*, Vol. 1158, No. 1, 2009, pp. 265–275.

[86] Hastie, T., Tibshirani, R., and Friedman, J., *The elements of statistical learning: data mining, inference, and prediction, second edition*, Springer Series in Statistics, Springer, Sept. 2009.

[87] Friedman, J., Hastie, T., and Tibshirani, R., "Regularization paths for generalized linear models via coordinate descent," *J Stat Software*, Vol. 33, No. 1, 2010, pp. 1–22.

[88] Fan, J., Guo, S., and Hao, N., "Variance estimation using refitted cross-validation in ultrahigh dimensional regression," *J R Stat Soc B*, Vol. 74, No. 1, 2012, pp. 37–65.

[89] Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J R Stat Soc B*, Vol. 57, No. 1, 1995, pp. 289–300.

[90] Görlach, M., Burd, C. G., and Dreyfuss, G., "The determinants of RNA-binding specificity of the heterogeneous nuclear ribonucleoprotein C proteins." *J Biol Chem*, Vol. 269, No. 37, 1994, pp. 23074–23078.

[91] Burd, C. G. and Dreyfuss, G., "RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing." *EMBO J.*, Vol. 13, No. 5, 1994, pp. 1197–1204.

[92] Dember, L. M., Kim, N. D., Liu, K.-Q., and Anderson, P., "Individual RNA recognition motifs of TIA-1 and TIAR have different RNA binding specificities," *J Biol Chem*, Vol. 271, No. 5, 1996, pp. 2783–2788.

[93] Lin, Q., Taylor, S. J., and Shalloway, D., "Specificity and determinants of sam68 RNA binding," *J Biol Chem*, Vol. 272, No. 43, 1997, pp. 27274–27280.

[94] Buckanovich, R. J. and Darnell, R. B., "The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo." *Mol Cell Biol*, Vol. 17, No. 6, 1997, pp. 3194–3201.

[95] Liu, H.-X., Zhang, M., and Krainer, A. R., "Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins," *Genes Dev*, Vol. 12, No. 13, 1998, pp. 1998–2012.

[96] Tacke, R., Tohyama, M., Ogawa, S., and Manley, J. L., "Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing," *Cell*, Vol. 93, No. 1, 1998, pp. 139–148.

[97] Cavaloc, Y., Bourgeois, C. F., Kister, L., and Stvenin, J., "The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers." *RNA*, Vol. 5, No. 3, 1999, pp. 468–483.

[98] Liu, H.-X., Chew, S. L., Cartegni, L., Zhang, M. Q., and Krainer, A. R., "Exonic splicing enhancer motif recognized by human SC35 under splicing conditions," *Mol Cell Biol*, Vol. 20, No. 3, 2000, pp. 1063–1071.

[99] Thisted, T., Lyakhov, D. L., and Liebhaber, S. A., "Optimized RNA targets of two closely related triple KH domain proteins, heterogeneous nuclear ribonucleoprotein K and αCP-2KL, suggest distinct modes of RNA recognition," *J Biol Chem*, Vol. 276, No. 20, 2001, pp. 17484–17496.

[100] Lerga, A., Hallier, M., Delva, L., Orvain, C., Gallais, I., Marie, J., and Moreau-Gachelin, F., "Identification of an RNA binding specificity for the potential splicing factor TLS," *J Biol Chem*, Vol. 276, No. 9, 2001, pp. 6807–6816.

[101] Peng, R., Dye, B. T., Pérez, I., Barnard, D. C., Thompson, A. B., and Patton, J. G., "PSF and p54nrb bind a conserved stem in U5 snRNA." *RNA*, Vol. 8, No. 10, 2002, pp. 1334–1347.

[102] Faustino, N. A. and Cooper, T. A., "Identification of putative new splicing targets for ETR-3 using sequences identified by systematic evolution of ligands by exponential enrichment," *Mol Cell Biol*, Vol. 25, No. 3, 2005, pp. 879–887.

[103] Galarneau, A. and Richard, S., "Target RNA motif and target mRNAs of the quaking star protein," *Nat Struct Mol Biol*, Vol. 12, No. 8, July 2005, pp. 691–698.

[104] Hori, T., Taguchi, Y., Uesugi, S., and Kurihara, Y., "The RNA ligands for mouse proline-rich RNA-binding protein (mouse Prrp) contain two consensus sequences in separate loop structure," *Nucleic Acids Res*, Vol. 33, No. 1, 2005, pp. 190–200.

[105] Smith, P. J., Zhang, C., Wang, J., Chew, S. L., Zhang, M. Q., and Krainer, A. R., "An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers," *Hum Mol Genet*, Vol. 15, No. 16, 2006, pp. 2490–2508.

[106] Paradis, C., Cloutier, P., Shkreta, L., Toutant, J., Klarskov, K., and Chabot, B., "hnRNP I/PTB can antagonize the splicing repressor activity of SRp30c," *RNA*, Vol. 13, No. 8, 2007, pp. 1287–1300.

[107] Reid, D. C., Chang, B. L., Gunderson, S. I., Alpert, L., Thompson, W. A., and Fairbrother, W. G., "Next-generation SELEX identifies sequence and structural determinants of splicing factor binding in human pre-mRNA sequence," *RNA*, Vol. 15, No. 12, 2009, pp. 2385–2397.

[108] Goers, E. S., Purcell, J., Voelker, R. B., Gates, D. P., and Berglund, J. A., "MBNL1 binds GC motifs embedded in pyrimidines to regulate alternative splicing," *Nucleic Acids Res*, Vol. 38, No. 7, 2010, pp. 2467–2484.

[109] Ray, D., Kazan, H., Chan, E., Castillo, L. P., Chaudhry, S., Talukder, S., Blencowe, B., Morris, Q., and Hughes, T., "Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins," *Nat Biotechnol*, Vol. 27, No. 7, June 2009, pp. 667–670.

[110] Auweter, S. D., Fasan, R., Reymond, L., Underwood, J. G., Black, D. L., Pitsch, S., and Allain, F. H.-T., "Molecular basis of RNA recognition by the human alternative splicing factor Fox-1." *EMBO J*, Vol. 25, No. 1, Jan 2006, pp. 163–173.

[111] Zhang, C., Zhang, Z., Castle, J., Sun, S., Johnson, J., Krainer, A. R., and Zhang, M. Q., "Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2," *Genes Dev*, Vol. 22, No. 18, 2008, pp. 2550–2563.

[112] Perez, I., Lin, C. H., McAfee, J. G., and Patton, J. G., "Mutation of PTB binding sites causes misregulation of alternative 3' splice site selection in vivo." *RNA*, Vol. 3, No. 7, 1997, pp. 764–778.

[113] Llorian, M., Schwartz, S., Clark, T. A., Hollander, D., Tan, L.-Y., Spellman, R., Gordon, A., Schweitzer, A. C., de la Grange, P., Ast, G., and Smith, C. W. J., "Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB," *Nat Struct Mol Biol*, Vol. 17, No. 9, Aug. 2010, pp. 1114–1123.

[114] Zearfoss, N. R., Clingman, C. C., Farley, B. M., McCoig, L. M., and Ryder, S. P., "Quaking regulates hnrnpa1 expression through its 3' utr in oligodendrocyte precursor cells," *PLoS Genet*, Vol. 7, No. 1, 01 2011, pp. e1001269.

[115] Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J., and Frey, B. J., "Deciphering the splicing code," *Nature*, Vol. 465, No. 7294, May 2010, pp. 53–59.

[116] Wu, J. I., Reed, R. B., Grabowski, P. J., and Artzt, K., "Function of quaking in myelination: regulation of alternative splicing," *Proc Natl Acad Sci U S A*, Vol. 99, No. 7, 2002, pp. 4233–4238.

[117] Ho, T. H., Charlet-B, N., Poulos, M. G., Singh, G., Swanson, M. S., and Cooper, T. A., "Muscleblind proteins regulate alternative splicing." *EMBO J.*, Vol. 23, No. 15, 2004, pp. 3103–3112.

[118] Timchenko, L. T., Miller, J. W., Timchenko, N. A., DeVore, D. R., Datar, K. V., Lin, L., Roberts, R., Caskey, C. T., and Swanson, M. S., "Identification of a (CUG)n triplet repeat RNA-binding protein and its expression in myotonic dystrophy," *Nucleic Acids Res*, Vol. 24, No. 22, 1996, pp. 4407–4414.

[119] Miller, J. W., Urbinati, C. R., Teng-Umnuay, P., Stenberg, M. G., Byrne, B. J., Thornton, C. A., and Swanson, M. S., "Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy." *EMBO J*, Vol. 19, No. 17, Sept. 2000, pp. 4439–4448.

[120] Oberstrass, F. C., Auweter, S. D., Erat, M., Hargous, Y., Henning, A., Wenter, P., Reymond, L., Amir-Ahmady, B., Pitsch, S., Black, D. L., and Allain, F. H.-T., "Structure of PTB bound to RNA: specific binding and implications for splicing regulation," *Science*, Vol. 309, No. 5743, 2005, pp. 2054–2057.

[121] Amir-ahmady, B., Boutz, P. L., Markovtsov, V., Phillips, M. L., and Black, D. L., "Exon repression by polypyrimidine tract binding protein," *RNA*, Vol. 11, No. 5, 2005, pp. 699–716.

[122] Caputi, M. and Zahler, A. M., "Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family," *J Biol Chem*, Vol. 276, No. 47, 2001, pp. 43850–43859.

[123] Martinez-Contreras, R., Fisette, J.-F., Nasim, F.-u. H., Madden, R., Cordeau, M., and Chabot, B., "Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing," *PLoS Biol*, Vol. 4, No. 2, 01 2006, pp. e21.

[124] Chaudhury, A., Chander, P., and Howe, P. H., "Heterogeneous nuclear ribonucleoproteins (hnRNPs) in cellular processes: focus on hnRNP E1's multifunctional regulatory roles," *RNA*, Vol. 16, No. 8, 2010, pp. 1449–1462.

[125] Paziewska, A., Wyrwicz, L. S., Bujnicki, J. M., Bomsztyk, K., and Ostrowski, J., "Cooperative binding of the hnRNP K three KH domains to mRNA targets," *FEBS Lett.*, Vol. 577, No. 1-2, 2004, pp. 134–140.

[126] Brown, J. W. S. and Simpson, C. G., "Splice site selection in plant pre-mRNA splicing," *Ann Rev Plant Physiol and Plant Mol Biol*, Vol. 49, No. 1, 1998, pp. 77–95.

[127] Merritt, H., McCullough, A. J., and Schuler, M. A., "Internal AU-rich elements modulate activity of two competing 3' splice sites in plant nuclei," *Plant J*, Vol. 12, No. 4, 1997, pp. 937–943.

[128] Zhu, H., Hinman, M. N., Hasman, R. A., Mehta, P., and Lou, H., "Regulation of neuron-specific alternative splicing of neurofibromatosis type 1 pre-mRNA," *Mol Cell Biol*, Vol. 28, No. 4, 2008, pp. 1240–1251.

[129] Wang, H., Molfenter, J., Zhu, H., and Lou, H., "Promotion of exon 6 inclusion in hud pre-mRNA by hu protein family members," *Nucleic Acids Res*, Vol. 38, No. 11, 2010, pp. 3760–3770.

[130] Djordjevic, M., Sengupta, A. M., and Shraiman, B. I., “A biophysical approach to transcription factor binding site discovery,” *Genome Res*, Vol. 13, No. 11, 2003, pp. 2381–2390.

[131] Stratton, M. R., Campbell, P. J., and Futreal, P. A., “The cancer genome,” *Nature*, Vol. 458, No. 7239, April 2009, pp. 719–724.

[132] Meyerson, M., Gabriel, S., and Getz, G., “Advances in understanding cancer genomes through second-generation sequencing,” *Nat Rev Genet*, Vol. 11, No. 10, Oct. 2010, pp. 685–696.

[133] Ding, L., Wendl, M. C., Koboldt, D. C., and Mardis, E. R., “Analysis of next-generation genomic data in cancer: accomplishments and challenges,” *Hum Mol Genet*, Vol. 19, No. R2, 2010, pp. R188–R196.

[134] Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., and Botstein, D., “Molecular portraits of human breast tumours.” *Nature*, Vol. 406, No. 6797, Aug. 2000, pp. 747–752.

[135] van Kouwenhove, M., Kedde, M., and Agami, R., “MicroRNA regulation by RNA-binding proteins and its implications for cancer,” *Nat Rev Cancer*, Vol. 11, No. 9, Sept. 2011, pp. 644–656.

[136] Davuluri, R. V., Suzuki, Y., Sugano, S., Plass, C., and Huang, T. H.-M., “The functional consequences of alternative promoter use in mammalian genomes,” *Trends Genet*, Vol. 24, No. 4, 2008, pp. 167 – 177.

[137] Mayr, C. and Bartel, D. P., “Widespread shortening of 3UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells,” *Cell*, Vol. 138, No. 4, 2009, pp. 673–684.

[138] Gardina, P., Clark, T., Shimada, B., Staples, M., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S., Davies, C., Williams, A., and Turpaz, Y., “Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array,” *BMC Genomics*, Vol. 7, No. 1, 2006, pp. 325.

[139] Lapuk, A., Marr, H., Jakkula, L., Pedro, H., Bhattacharya, S., Purdom, E., Hu, Z., Simpson, K., Pachter, L., Durinck, S., Wang, N., Parvin, B., Fontenay, G., Speed, T., Garbe, J., Stampfer, M., Bayandorian, H., Dorton, S., Clark, T. A., Schweitzer, A., Wyrobek, A., Feiler, H., Spellman, P., Conboy, J., and Gray, J. W., “Exon-level microarray analyses identify alternative splicing programs in breast cancer,” *Molecular Cancer Research*, Vol. 8, No. 7, 2010, pp. 961–974.

[140] Misquitta-Ali, C. M., Cheng, E., O’Hanlon, D., Liu, N., McGlade, C. J., Tsao, M. S., and Blencowe, B. J., “Global profiling and molecular characterization of

alternative splicing events misregulated in lung cancer," *Molecular and Cellular Biology*, Vol. 31, No. 1, January 1, 2011, pp. 138–150.

[141] Gilbert, W., "Why genes in pieces?" *Nature*, Vol. 271, No. 5645, Feb. 1978, pp. 501.

[142] Liu, M. and Grigoriev, A., "Protein domains correlate strongly with exons in multiple eukaryotic genomes–evidence of exon shuffling?" *Trends Genet*, Vol. 20, No. 9, 2004, pp. 399 – 403.

[143] Keren, H., Lev-Maor, G., and Ast, G., "Alternative splicing and evolution: diversification, exon definition and function," *Nat Rev Genet*, Vol. 11, No. 5, May 2010, pp. 345–355.

[144] Chin, L., Hahn, W. C., Getz, G., and Meyerson, M., "Making sense of cancer genomic data," *Genes Dev*, Vol. 25, No. 6, 2011, pp. 534–555.

[145] Network, T. C. G. A. R., "Comprehensive molecular portraits of human breast tumours," *Nature*, Vol. 490, No. 7418, Sept. 2012, pp. 61–70.

[146] Trapnell, C., Pachter, L., and Salzberg, S. L., "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, Vol. 25, No. 9, 2009, pp. 1105–1111.

[147] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L., "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nat Biotechnol*, Vol. 28, No. 5, May 2010, pp. 511–515.

[148] Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L., "Identification of novel transcripts in annotated genomes using RNA-Seq," *Bioinformatics*, Vol. 27, No. 17, 2011, pp. 2325–2329.

[149] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L., "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," *Nat Protoc*, Vol. 7, No. 3, March 2012, pp. 562–578.

[150] Anders, S. and Huber, W., "Differential expression analysis for sequence count data," *Genome Biol*, Vol. 11, No. 10, 2010, pp. R106.

[151] Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Lalo, D., Le Gall, C., Schaffer, B., Le Crom, S., Guedj, M., and Jaffrzic, F., "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis," *Briefings in Bioinformatics*, 2012.

[152] Tusher, V. G., Tibshirani, R., and Chu, G., “Significance analysis of microarrays applied to the ionizing radiation response,” *Proc Natl Acad Sci U S A*, Vol. 98, No. 9, 2001, pp. 5116–5121.

[153] Griffith, M., Griffith, O. L., Mwenifumbo, J., Goya, R., Morrissy, A. S., Morin, R. D., Corbett, R., Tang, M. J., Hou, Y.-C. C., Pugh, T. J., Robertson, G., Chittaranjan, S., Ally, A., Asano, J. K., Chan, S. Y., Li, H. I., McDonald, H., Teague, K., Zhao, Y., Zeng, T., Delaney, A., Hirst, M., Morin, G. B., Jones, S. J., Tai, I. T., and Marra, M. A., “Alternative expression analysis by RNA sequencing.” *Nature methods*, Vol. 7, No. 10, Oct. 2010, pp. 843–847.

[154] Richard, H., Schulz, M. H., Sultan, M., Nürnberger, A., Schrinner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., Haas, S. A., and Yaspo, M.-L., “Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments,” *Nucleic Acids Res*, Vol. 38, No. 10, Feb. 2010, pp. e112.

[155] Brooks, A. N., Yang, L., Duff, M. O., Hansen, K. D., Park, J. W., Dudoit, S., Brenner, S. E., and Graveley, B. R., “Conservation of an RNA regulatory map between Drosophila and mammals,” *Genome Res*, Vol. 21, No. 2, 2011, pp. 193–202.

[156] Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B., “Analysis and design of RNA sequencing experiments for identifying isoform regulation,” *Nature Methods*, Vol. 7, No. 12, Nov. 2010, pp. 1009–1015.

[157] Shen, S., Park, J. W., Huang, J., Dittmar, K. A., Lu, Z.-x., Zhou, Q., Carstens, R. P., and Xing, Y., “MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data,” *Nucleic Acids Res*, Vol. 40, No. 8, 2012, pp. e61.

[158] Thorsen, K., Sorensen, K. D., Brems-Eskildsen, A. S., Modin, C., Gaustadnes, M., Hein, A.-M. K., Kruhufer, M., Laurberg, S., Borre, M., Wang, K., Brunak, S., Krainer, A. R., Trring, N., Dyrskjt, L., Andersen, C. L., and rntoft, T. F., “Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis,” *Molecular & Cellular Proteomics*, Vol. 7, No. 7, July 2008, pp. 1214–1224.

[159] Consortium, T. U., “Reorganizing the protein space at the universal protein resource (UniProt),” *Nucleic Acids Research*, Vol. 40, No. D1, Jan. 2012, pp. D71–D75.

[160] Lewit-Bentley, A. and Rty, S., “EF-hand calcium-binding proteins,” *Current Opinion in Structural Biology*, Vol. 10, No. 6, 2000, pp. 637 – 643.

[161] Witke, W., Hofmann, A., Koppel, B., Schleicher, M., and Noegel, A. A., “The Ca(2+)-binding domains in non-muscle type alpha-actinin: biochemical and genetic analysis.” *The Journal of Cell Biology*, Vol. 121, No. 3, 1993, pp. 599–606.

[162] Prevarskaya, N., Skryma, R., and Shuba, Y., "Calcium in tumour metastasis: new roles for known actors," *Nat Rev Cancer*, Vol. 11, No. 8, July 2011, pp. 609–618.

[163] Liu, Y., Annis, D. S., and Mosher, D. F., "Interactions among the epidermal growth factor-like modules of thrombospondin-1," *J Biol Chem*, Vol. 284, No. 33, 2009, pp. 22206–22212.

[164] Law, E. W. L., Cheung, A. K. L., Kashuba, V. I., Pavlova, T. V., Zabarovsky, E. R., Lung, H. L., Cheng, Y., Chua, D., Kwong, D. L.-w., Tsao, S. W., Sasaki, T., Stanbridge, E. J., and Lung, M. L., "Anti-angiogenic and tumor-suppressive roles of candidate tumor-suppressor gene, Fibulin-2, in nasopharyngeal carcinoma," *Oncogene*, Vol. aop, No. current, July 2011.

[165] Liang, X., Huuskonen, J., Hajivandi, M., Manzanedo, R., Predki, P., Amshey, J. R., and Pope, R. M., "Identification and quantification of proteins differentially secreted by a pair of normal and malignant breast-cancer cell lines," *PROTEOMICS*, Vol. 9, No. 1, 2009, pp. 182–193.

[166] Gotoh, I., Adachi, M., and Nishida, E., "Identification and characterization of a novel MAP kinase kinase kinase, MLTK," *J Biol Chem*, Vol. 276, No. 6, 2001, pp. 4276–4286.

[167] Cho, Y.-Y., Bode, A. M., Mizuno, H., Choi, B. Y., Choi, H. S., and Dong, Z., "A novel role for mixed-lineage kinase-like mitogen-activated protein triple kinase alpha in neoplastic cell transformation and tumor development," *Cancer Research*, Vol. 64, No. 11, 2004, pp. 3855–3864.

[168] Venables, J. P., Klinck, R., Bramard, A., Inkel, L., Dufresne-Martin, G., Koh, C., Gervais-Bird, J., Lapointe, E., Froehlich, U., Durand, M., Gendron, D., Brosseau, J.-P., Thibault, P., Lucier, J.-F., Tremblay, K., Prinos, P., Wellinger, R. J., Chabot, B., Rancourt, C., and Elela, S. A., "Identification of alternative splicing markers for breast cancer," *Cancer Research*, Vol. 68, No. 22, 2008, pp. 9525–9531.

[169] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nat Genet*, Vol. 25, No. 1, May 2000, pp. 25–29.

[170] Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z., "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists," *BMC Bioinformatics*, Vol. 10, No. 1, 2009, pp. 48.

[171] Warrington, R. and Lewis, K., "Natural antibodies against nerve growth factor inhibit in vitro prostate cancer cell metastasis," *Cancer Immunology, Immunotherapy*, Vol. 60, 2011, pp. 187–195.

[172] Adriaenssens, E., Vanhecke, E., Saule, P., Mougel, A., Page, A., Romon, R., Nurcombe, V., Le Bourhis, X., and Hondermarck, H., "Nerve growth factor is a potential therapeutic target in breast cancer," *Cancer Research*, Vol. 68, No. 2, 2008, pp. 346–351.

[173] Powers, J. F., Shahsavari, M., Tsokas, P., and Tischler, A. S., "Nerve growth factor receptor signaling in proliferation of normal adult rat chromaffin cells," *Cell and Tissue Research*, Vol. 295, 1999, pp. 21–32.

[174] Harburg, G. and Hinck, L., "Navigating breast cancer: axon guidance molecules as breast cancer tumor suppressors and oncogenes," *Journal of Mammary Gland Biology and Neoplasia*, Vol. 16, 2011, pp. 257–270.

[175] Chedotal, A., Kerjan, G., and Moreau-Fauvarque, C., "The brain within the tumor: new roles for axon guidance molecules in cancers," *Cell Death & Differentiation*, Vol. 12, No. 8, Aug. 2005, pp. 1044–1056.

[176] Mehlen, P., Delloye-Bourgeois, C. A., and ChÃ©dotal, A., "Novel roles for Slits and netrins: axon guidance cues as anticancer targets?" *Nat Rev Cancer*, Vol. 11, No. 3, March 2011, pp. 188–197.

[177] Biankin, A. V., Waddell, N., Kassahn, K. S., Gingras, M.-C., Muthuswamy, L. B., Johns, A. L., Miller, D. K., Wilson, P. J., Patch, A.-M., Wu, J., Chang, D. K., Cowley, M. J., Gardiner, B. B., Song, S., Harliwong, I., Idrisoglu, S., Nourse, C., Nourbakhsh, E., Manning, S., Wani, S., Gongora, M., Pajic, M., Scarlett, C. J., Gill, A. J., Pinho, A. V., Rooman, I., Anderson, M., Holmes, O., Leonard, C., Taylor, D., Wood, S., Xu, C., Nones, K., Lynn Fink, J., Christ, A., Bruxner, T., Cloonan, N., Kolle, G., Newell, F., Pinese, M., Scott Mead, R., Humphris, J. L., Kaplan, W., Jones, M. D., Colvin, E. K., Nagrial, A. M., Humphrey, E. S., Chou, A., Chin, V. T., Chantrill, L. A., Mawson, A., Samra, J. S., Kench, J. G., Lovell, J. A., Daly, R. J., Merrett, N. D., Toon, C., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Kakkar, N., Zhao, F., Qing Wu, Y., Wang, M., Muzny, D. M., Fisher, W. E., Charles Brunicardi, F., Hodges, S. E., Reid, J. G., Drummond, J., Chang, K., Han, Y., Lewis, L. R., Dinh, H., Buhay, C. J., Beck, T., Timms, L., Sam, M., Begley, K., Brown, A., Pai, D., Panchal, A., Buchner, N., De Borja, R., Denroche, R. E., Yung, C. K., Serra, S., Onetto, N., Mukhopadhyay, D., Tsao, M.-S., Shaw, P. A., Petersen, G. M., Gallinger, S., Hruban, R. H., Maitra, A., Iacobuzio-Donahue, C. A., Schulick, R. D., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Capelli, P., Corbo, V., Scardoni, M., Tortora, G., Tempero, M. A., Mann, K. M., Jenkins, N. A., Perez-Mancera, P. A., Adams, D. J., Largaespada, D. A., Wessels, L. F. A., Rust, A. G., Stein, L. D., Tuveson, D. A., Copeland, N. G., Musgrove, E. A., Scarpa, A., Eshleman, J. R., Hudson, T. J., Sutherland, R. L., Wheeler, D. A., Pearson,

J. V., McPherson, J. D., Gibbs, R. A., and Grimmond, S. M., “Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes,” *Nature*, Vol. advance online publication, Oct. 2012.

[178] Scaltriti, M. and Baselga, Jose, t. . T. v. . . n. . . p. . . y. . . j. . C.

[179] Lo, H. W. and Hung, M. C., “Nuclear EGFR signalling network in cancers: linking EGFR pathway to cell cycle progression, nitric oxide pathway and patient survival.” *Br J Cancer*, Vol. 94, 2006, pp. 184–188.

[180] Shigematsu, H. and Gazdar, A. F., “Somatic mutations of epidermal growth factor receptor signaling pathway in lung cancers,” *International Journal of Cancer*, Vol. 118, No. 2, 2006, pp. 257–262.

[181] Haber, D. A. and Settleman, J., “Cancer: drivers and passengers,” *Nature*, Vol. 446, No. 7132, March 2007, pp. 145–146.

[182] Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O/’Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Menzies, A., Mironenko, T., Perry, J., Raine, K., Richardson, D., Shepherd, R., Small, A., Tofts, C., Varian, J., Webb, T., West, S., Widaa, S., Yates, A., Cahill, D. P., Louis, D. N., Goldstraw, P., Nicholson, A. G., Brasseur, F., Looijenga, L., Weber, B. L., Chiew, Y.-E., deFazio, A., Greaves, M. F., Green, A. R., Campbell, P., Birney, E., Easton, D. F., Chenevix-Trench, G., Tan, M.-H., Khoo, S. K., Teh, B. T., Yuen, S. T., Leung, S. Y., Wooster, R., Futreal, P. A., and Stratton, M. R., “Patterns of somatic mutation in human cancer genomes,” *Nature*, Vol. 446, No. 7132, March 2007, pp. 153–158.

[183] Youn, A. and Simon, R., “Identifying cancer driver genes in tumor genome sequencing studies,” *Bioinformatics*, Vol. 27, No. 2, 2011, pp. 175–181.

[184] Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., Pochanard, P., Mozes, E., Garraway, L. A., and Pe’er, D., “An integrated approach to uncover drivers of cancer,” *Cell*, Vol. 143, No. 6, 2010, pp. 1005 – 1017.

[185] Huang, N., Shah, P. K., and Li, C., “Lessons from a decade of integrating cancer copy number alterations with gene expression profiles,” *Briefings in Bioinformatics*, 2011.

[186] Srebrow, A. and Kornblihtt, A. R., “The connection between splicing and cancer,” *Journal of Cell Science*, Vol. 119, No. 13, 2006, pp. 2635–2641.

[187] Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R., and Easton, D. F., “Statistical analysis of pathogenicity of somatic mutations in cancer,” *Genetics*, Vol. 173, No. 4, August 2006, pp. 2187–2198.

[188] Sauna, Z. E. and Kimchi-Sarfaty, C., “Understanding the contribution of synonymous mutations to human disease,” *Nat Rev Genet*, Vol. 12, No. 10, Oct. 2011, pp. 683–691.

[189] Kong-Beltran, M., Seshagiri, S., Zha, J., Zhu, W., Bhawe, K., Mendoza, N., Holcomb, T., Pujara, K., Stinson, J., Fu, L., Severin, C., Rangell, L., Schwall, R., Amler, L., Wickramasinghe, D., and Yauch, R., “Somatic mutations lead to an oncogenic deletion of met in lung cancer,” *Cancer Research*, Vol. 66, No. 1, 2006, pp. 283–289.

[190] Segal, E., Friedman, N., Koller, D., and Regev, A., “A module map showing conditional activity of expression modules in cancer.” *Nat Genet*, Vol. 36, No. 10, Oct. 2004, pp. 1090–1098.

[191] de Souto, M., Costa, I., de Araujo, D., Ludermir, T., and Schliep, A., “Clustering cancer gene expression data: a comparative study,” *BMC Bioinformatics*, Vol. 9, No. 1, 2008, pp. 497.

[192] Zhang, C., Li, H.-R., Fan, J.-B., Wang-Rodriguez, J., Downs, T., Fu, X.-D., and Zhang, M., “Profiling alternatively spliced mRNA isoforms for prostate cancer classification,” *BMC Bioinformatics*, Vol. 7, No. 1, 2006, pp. 202.