TRANSCRIPTOME-WIDE STUDY OF ALTERNATIVE SPLICING ACROSS MULTIPLE CANCER TYPES

Yi-hsuan Tsai

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in particial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Bioinformatics and Computational Biology.

Chapel Hill
2015

Approved by

Shawn M. Gomez,

Scott M. Hammond,

Alain Laederach,

Jan Prins,

Zefeng Wang.

ABSTRACT

Yi-hsuan Tsai: Transcriptome-wide study of alternative splicing across multiple cancer types
(Under the direction of Zefeng Wang and Shawn M. Gomez)

Alternative splicing (AS) is a very important cellular process in eukaryote, which contribute to both the proteome diversity and control of gene expression levels. It is tightly regulated in different tissues and developmental stages and dysregulation of splicing can lead to many human diseases. Cancer is one of the extreme examples where splicing is highly altered. However the underline mechanism responsible for such dysregulation is largely unclear. Moreover, identification of cancer specific AS events is complicated by the large noise of different tissues-specific splicing. In the chapters that follow, we first explore the evolution and functionality of RNA binding proteins which is one of the key regulators of AS. We then use the RNA-seq data from TCGA (The Cancer Genome Atlas) to identify AS events that are significantly altered between tumor and normal samples across multiple cancer types. We also show that these cancer-specific AS events are highly conserved, more likely to maintain protein reading frame, and mainly function in cell cycle, cell adhesion/migration, and insulin signaling pathway. Furthermore, these events can serve as new molecular biomarkers to distinguish cancer from normal tissues, to separate cancer subtypes, and to predict patient survival. We also demonstrate that most genes whose expression is closely associated with cancer-specific splicing are key regulators of the cell cycle. Our study uncovers a common set of cancer-specific AS events altered across multiple cancers, providing mechanistic insight into how splicing is mis-regulated in cancers. Lastly

we show that kidney tumors harbor significantly higher intron retention than other tumor types and such increase of splicing alteration in kidney cancer is highly correlated with patient survival. Together our works have helped to better understand AS across multiple human cancers and how it might be regulated and its connection to some important cancer pathways.

ACKNOWLEDGMENTS

This work would not have been possible without the support of my two awesome advisors, Zefeng Wang and Shawn Gomez. All the advice and guidance they have given me over the past years has been incredible. Both Zefeng and Shawn are brilliant scientists with keen insights. Their mentorship, knowledge has taught me how to think, how to ask research questions and how to conduct scientific research in my scientific and personal development. I really appreciate that they give me a lot of flexibility in my research and have always being supportive about my research projects, both financially and mentally.

I'm thankful for the past and present members of the Wang lab and Gomez lab for their support, encouragement and expertise: Yang Wang, Daniel Dominguez, Wenjing Zhang, Rajarshi Choudhury, Shuyun Dong, Russell Maxwell, HuanHuan Wei, Yun Yang, Miaowei Mao, Kyla Collins, Janet Doolittle, Matt Berginski, Alicia Midland, Kwangbom Choi, Jennifer Staab. Yang and Daniel have left the lab last year, but they have been there since I joined the lab. Although the collaboration with Daniel can sometimes be stressful, it gave me the chance to my first high-throughput sequencing data analysis which prepare me for the skills I need for the future. During the past years, those lunch time together, holiday parties and birthday celebrations, all have made my stressful working day (sometimes), more cheerful, and I thank them for the accompany.

I am also grateful for my thesis committee members: Scott Hammond, Alain Laederach, and Jan Prins for their guidance and suggestions on my research projects. Thanks

to Dr. Prins inviting me to their weekly RNA-seq meeting in computer science department. I have learned many techniques I need for my research there and learned how to see research questions from a different angle in those meeting.

Another collaborator I would like to thank to is Joel Parker, who has been really helpful in providing the TCGA data and answers to all the questions I have about the data. His special skills and knowledge in breast cancer have been a great help.

Finally I would like to thank my family members who have been always supportive no matter what. Although my parents and brothers in Taiwan have no clue what I'm doing in the US, they always remember to reminds me not to work too hard and sleep more through skype. I'm especially grateful for my husband who has been always standing by my side, helping me through many difficult times and helping celebrate every joyful moments with me. Without their love and support, this dissertation would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

RNA – Ribonucleic acid

mRNA – messenger RNA

AS – Alternative splicing

TCGA – The Cancer Genome Atlas

BPS – Branch Point Sequence

3'SS – 3' Splice Site

5'SS – 5' Splice Site

SRE – Splicing Regulatory Elementary

ESEs – Exonic Splicing Enhancers

ESSs – Exonic Splicing Silencers

ISEs – Intronic Splicing Enhancers

ISSs – Intronic Splicing Silencers

RBP – RNA Binding Protein

RRM – RNA Recognition Motif

KH – K Homology

PAZ – Piwi/Argonaute/Zwille

RBD – RNA Binding Domains

hnRNP – Heterogeneous nuclear riboncleoprotein

SR – Serine-Argine

ELAV – embryonic lethal abnormal vision

CELF – CUG-BP and ETR-like factor

PPR – Pentatricopeptide Repeat

PUF – Pumilio/FBF

CSD – Cold-Shock Domain

nt – nucleotides

DAZL – DAZ like

PABPs – Polyadenylate-Binding Proteins

UTR – UnTranslated Regions

GO – Gene Ontology

PSI – Percent Spliced In

MISO – Mixture of isoforms

BRCA – BReast invasive CArcinoma

LUSC – LUng Squamous cell Carcinoma

LIHC – LIver Hepatocellular Carcinoma

SE – Skipped Exon

RI – Retained Intron

A3SS – Alternative 3' Splice Site

A5SS – Alternative 5' Splice Site

NMD – Nonsense Mediated Decay

PCA – Principal Component Analysis

WBP1 – WW domain binding protein 1

BLCA – Bladder Urothelial Carcinoma

COAD – Colon adenocarcinoma

HNSC – Head and Neck squamous cell carcinoma

KICH – Kidney Chromophobe

KIRC – Kidney renal clear cell carcinoma

KIRP – Kidney renal papillary cell carcinoma

LUAD – Lung adenocarcinoma

PRAD – Prostate adenocarcinoma

THCA – Thyroid carcinoma

UCEC – Uterine Corpus Endometrial Carcinoma

PPI – Protein Protein Interaction

ccRCC – clear-cell renal carcinoma

IR – Intron Retention

lncRNA – long non-coding RNA

HIR – High Intron Retention

NIR – Normal Intron Retention

CLIP-seq – CrossLinking ImmunoPrecipitation sequencing

CHAPTER 1

INTRODUCTION

With the very first draft of the human genome announced in 2000 [1, 2], and the following full completion in 2003, people have once believed that the human genetic blueprint, such a big progress in biology, will bring us to the next era of medicine. Decoding of the human genomic sequences indeed took us to the next level of biological research. Hypothesis-driven science has been largely replaced by discovery-driven research because of the large amount of data. However, this is just the beginning of post-genomic revolution. Although we have all the sequence code for building a human being, there are still lots of unknown. What regions are coded for functional genes, what sequences contain regulatory elements and what else are just merely random noises. Genome annotation has progressed very slowly after the completion of human genome. After more than a decade, people are still debating about what percentage of the human genome is actually doing something important. To our surprise, only less than 2 percent of the human genome contains protein-coding sequences. Even when counting the regulatory sequences, the fraction of functional DNA is estimated less than 10-20% [3, 4].

Although deciphering the human genome wasn't as easy as we originally thought, many big advances in medical fields have been made. Predictive genetic tests for many human diseases are available now and personalized medicine is also on its way for future cancer therapy. All these would have not been possible without the first map of human genome.

1.1 Next-generation sequencing and RNA-seq

The next revolution occurred when the 'second-generation' (or next-generation) sequencing was developed. The new high-throughput sequencing technology allows us to sequence DNA/RNA much more quickly and with much lower cost. According to National Human Genome Research Institute's report, the cost of sequencing a genome was more than 10M in early 2000 and it has dropped to below 10K in around 2011 [5]. Such a dramatically drop in price has outpaced Moore's law since early 2008 when the next-generation sequencing technology took the place of the Sanger-based sequencing [5]. Because of the low cost, sequence data has been generated in a way that we cannot imagine and cannot finish analyzing them in time. Large consortium, such as ENCODE [6], 1000 genomes [7], and TCGA [8] all have accumulated large amount of data, and most of the data are public accessible. Even though some of these sequence data have been released for a while, we can still see new publications using the data every month. There are still a lot of treasures within these sequencing data waiting for us to discover.

The new technology has also made 'RNA-seq' (RNA sequencing) possible. The concept of RNA-seq is very simple. Short read sequences (typically about 25-50 nucleotides long) are obtained from random locations along RNA by either sequencing sheared double-stranded cDNA libraries (strandless RNA-seq) or sequencing directional cDNA libraries (stranded RNA-seq) [9]. After hundreds of millions of short sequences have been generated, they are then mapped back to a reference genome allowing gaps between exon-exon junctions using bioinformatics algorithms. RNA-seq can be used to study the dynamic of eukaryotic transcriptomes, make it possible to redefine the transcriptome content in different cell types,

different tissues and different developmental stages, not only qualitatively but also quantitatively.

1.2 Alternative splicing and its regulation

Since human transcriptome has been deeply sequenced, it's been shown that more than 90% of human genes undergo alternative splicing (AS) process to produce more than one splicing isoforms containing different combinations of exons [10, 11]. This important cellular process can control the expression level of genes and contribute to the diversity of proteome. Through AS, same gene can have protein isoforms with totally different functions. For example, the Fas receptor gene has one soluble and one membrane-bound isoforms with opposing effects on apoptosis [12]. Another example is the *Drosophila* fruitless (fru) gene, which is spliced differently in males and females to control its sexual orientation [13]. Splicing process is also tightly regulated in different tissues and developmental stages [10, 11], and dysregulation of AS is closely associated with various human diseases [14, 15].

The specificity of splicing is mainly determined by the core splicing signals including the 5' or 3' splice site (i.e. 5'ss or 3'ss) at each end of an intron and the branch point sequence (BPS) at the upstream of 3'ss. However, the core splice site motifs contain only about half of the information required to precisely define exon/intron boundaries [16]. In addition to these core splicing signals, splicing is regulated by multiple splicing regulatory elements (SREs) that specifically recruit *trans*-acting splicing factors to activate or repress the use of adjacent splice sites [17]. There are four types of SRE, two in the exonic region: exonic splicing enhancers (ESEs) or silencers (ESSs) that function to promote or inhibit inclusion of the exon they reside in, and two in the intronic region: intronic splicing enhancers (ISEs) or silencers (ISSs) that

enhance or inhibit usage of adjacent splice sites or exons (Figure. 1.1) [18]. It is known that the activities of SREs may depend on the relative locations of the elements in pre-mRNAs, e.g. G-triplets could act as an ISE when it's located in intronic locations, but also function as an ESS when located in exons [19]. This is commonly known as "context dependence" of SREs. Same effects also apply to splicing factors: the same splicing factor may either activate or inhibit splicing by binding to its cognate SREs in different pre-mRNA regions [18, 20, 21]. More details of splicing regulation can be found in some recent reviews [17, 22-25].

1.3 RNA binding proteins and its role in splicing regulation

Many RNA-binding proteins (RBPs) facilitate splicesome assembly on pre-mRNA during splicing process. In addition to RNA splicing, specific interactions between RBPs and RNAs also play essential roles in many other mRNA metabolism, including RNA editing, translocation, and degradation [26]. Altering the expression of RBPs can have dramatic impact on in various RNA-related cellular functions, and aberrant RBP function can also lead to a wide range of human diseases including cancer [27] and neurodegenerative diseases [28, 29].

The sequence-specific interaction of RBPs and RNAs are mediated through various RNA binding domains that contains 60-100 amino acids with α-helix and β-strand topology, including RNA Recognition Motif (RRM), K homology (KH) domain, Piwi/Argonaute/Zwille (PAZ) domain, and etc. With recent advances in RNA biology, many proteins have been identified that interacts with RNAs and at least more than 40 RNA biding domains (RBDs) have been categorized [30]. Among them, RRM is the most frequent domains presented in >50% of RBPs [31]. In addition, many RBPs contain more than one RBD, and it is unclear how each RBD contributes to the binding specificity of the RBPs with multiple RBD and whether all RBDs are

required for target binding. Previous study has shown that multicellular organisms have more RBPs than single cell organisms and the number of RBD within a RBP has expanded significantly in mammal as well [32]. However, the mechanism by which the number of RBPs and RBDs increase during evolution and the functional implication of such expansion remain unclear.

One of the most important cellular process RBPs regulate is AS which is the main subject of this study. Some RBPs that regulate splicing have been identified, however, many of them have different functions when binding to different sequence targets [18, 23], which makes the splicing regulation more complicated. For example, the CUGBP2 splicing factor can promote the inclusion of cTNT exon 5 via biding to its downstream intron region, while in the brain, it can silence the N1 exon of the NMDA R1 receptor through binding to its upstream intron region [33, 34]. Some of the splicing regulation proteins can compete with each other for the binding sites. For example, in HIV tat exon 3, both hnRNPA1 and SF2 can bind to the exon to inhibit or to enhance the splicing respectively [35]. A1 binds to an ESS to repress the splicing but the SF2 can also bind to the ESEs located in the same exon to promote the splicing. When SF2 presents, it blocks the A1 repression and allow the exon inclusion. Another example is the regulation of exon 11 of the insulin receptor gene, where hnRNP F and SRSF1 compete with the hnRNP A1 for the binding site to promote or inhibit exon 11 inclusion [36]. Some of the splicing regulation proteins are dosage depend. For example, the relative ratio of A1 to ASF/SF2 can regulate splicing patterns differently [37] [38].

Although it's believed that many more proteins can regulate RNA splicing, the four major groups of known splicing regulators are: the heterogeneous nuclear ribonucleoprotein (hnRNP), serine-argine (SR) proteins, embryonic lethal, abnormal vision (ELAV)-like proteins

and CUG-BP and ETR-like factor (CELF) proteins [39]. Review of these proteins can be found in [40-43].

1.4 Splicing in cancers

The mis-regulation of splicing is a common cause of various human diseases including cancer. Hundreds of genes are mis-spliced in a typical cancer cell, and many cancer-specific splicing isoforms play key roles in pathogenesis and growth of tumors. For example, in glioblastoma, a tumor-specific α-exon skipping isoform, FDFR1β is overexpressed in the tumor cells and the overexpression is regulated by the increase in expression of a splicing inhibitory, PTB (hnRNPI) [44]. As one of the molecular hallmarks of human cancer, the splicing mis-regulation is thought to be controlled by the changes of expression levels and/or activity of certain splicing factors. Several splicing factors, including SRSF1 and hnRNP A2, were found to act as a proto-oncogene to induce malignant transformation of normal cells [45-47], while other splicing factors, such as RBM5 [48] and RBM4 [49], can serve as tumor suppressor genes to inhibit cancer growth. Therefore the relationship between splicing factors and the cancer-specific splicing profile is a very important research subject.

Another mis-regulation of cancer splicing pathway is through the *cis*-acting SREs. Since AS regulation is through the sequence-specific interaction of splicing factors and its binding targets, mutations of the *cis*-elements can disrupt such specific interaction and lead to splicing mis-regulation. For example a missense mutation in BRCA1 gene increase the binding affinity of splicing repressors, hnRNP A1 and hnRNP H/F, which can increase exon skipping in breast cancer [50].

It has been shown that the changes of alternative splicing can be used as a powerful biomarker to diagnose cancer or to predict the response to therapy [51, 52]. In the past, several studies were designed to perturb each splicing factor and further determine how an individual splicing factor affected splicing of specific genes or entire transcriptome. However the TCGA consortium presented a unique dataset that mimic the perturbance of various splicing factors at same time in a large number of clinical samples. Therefore we can study the splicing regulation in cancer by measuring the correlation of splicing changes with levels of all known and putative splicing factors across all samples. Through investigating the correlation between splicing and gene expression, we aim to identify key splicing factors that responsible for the splicing mis-regulation in cancers. Furthermore, these findings can provide novel anti-cancer therapeutic targets based on the cancer specific RNA-RBP interactions.

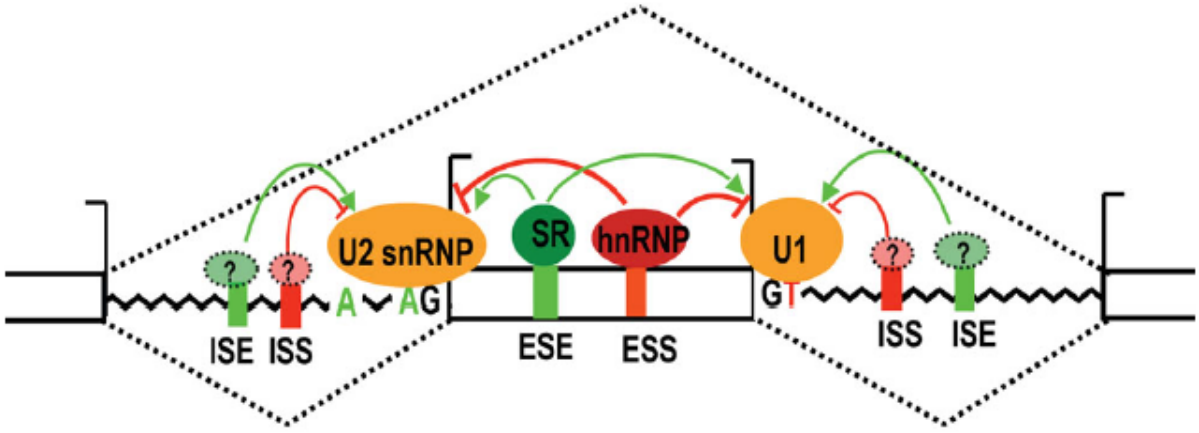

Figure 1.1 Schematic diagram of splicing regulation.

Open boxes represent exons and jagged lines are introns. The brackets represent splice sites. The adenosine in branch point is also indicated. Splicing is regulated by *trans*-acting splicing factors (SR protein, hnRNP or unknown factors) that recognize *cis*-elements (classified as ESE, ESS, ISS and ISE). Figure adapted from [18].

CHAPTER 2

IDENTIFICATION OF PREVALENT RNA RECOGNITION MOTIF DUPLICATION IN THE HUMAN GENOME.[1]

2.1 Overview

The sequence-specific recognition of RNA by proteins is mediated through various RNA binding domains, with the RNA recognition motif (RRM) being the most frequent and present in >50% of RNA-binding proteins (RBPs). Many RBPs contain multiple RRMs, and it is unclear how each RRM contributes to the binding specificity of the entire protein. We found that RRMs within the same RBP (i.e., sibling RRMs) tend to have significantly higher similarity than expected by chance. Sibling RRM pairs from RBPs shared by multiple species tend to have lower similarity than those found only in a single species, suggesting that multiple RRMs within the same protein might arise from domain duplication followed by divergence through random mutations. This finding is exemplified by a recent RRM domain duplication in DAZ proteins and an ancient duplication in PABP proteins. Additionally, we found that different similarities between sibling RRMs are associated with distinct functions of an RBP and that the RBPs tend to contain repetitive sequences with low complexity. Taken together, this study suggests that the

[1] This chapter previously appeared as an article in the Journal of *RNA*. The original citation is as

number of RBPs with multiple RRMs has expanded in mammals and that the multiple sibling RRMs may recognize similar target motifs in a cooperative manner.

2.2 Introduction

Specific interactions between RNAs and proteins play an essential role in regulating mRNA processing, including RNA splicing, polyadenylation, translocation, and degradation [26]. Altering the level or activity of RNA-binding proteins (RBPs) has a dramatic impact on various RNA-related cellular functions, with aberrant RBP function leading to human diseases [27]. For example, many RBPs specifically recognize regulatory cis-elements in pre-mRNA and thereby inhibit or promote use of nearby splicing sites [17, 18]. The binding between these splicing factors and their RNA target is crucial to many cellular processes, as most human genes undergo alternative splicing to produce multiple isoforms with distinct functions. Therefore, examining the interactions between different RBPs and their RNA targets is an important component in understanding various gene regulation pathways.

The sequence-specific interaction between RBPs and single-stranded RNAs is usually mediated through various RNA binding domains (RBDs) including the RNA recognition motif (RRM), the pentatricopeptide repeat (PPR), the K homology (KH), the zinc-finger, the Pumilio/FBF (PUF), and the cold-shock (CSD) domains [53]. Although protein sequence elements outside of the RBD may contact RNA and affect RNA binding [54, 55], the RBD is the key determinant of RNA binding specificity [56]. Among them, the RRM is the most abundant and present in over 50% of RBPs in humans [31]. A typical RRM contains 80–90 aa that fold into a β1α1β2β3α2β4 topology, where the four anti-parallel β-sheets and the two additional α helices create ample surface that interacts with RNA [31, 57]. The most conserved region of the

RRM consists of two short sites (6–8 aa) in β1 and β3 (named RNP-2 and RNP-1, respectively) that are crucial for RNA interaction [57-59]. However, recent structures of various RRMs bound by their cognate RNA show that RRMs may interact with RNA through diverse mechanisms [60-62]. For example, hnRNP I (poly-pyrimidine tract binding protein or PTB) has four RRMs with similar specificities. The β3 of each RRM contributes only weakly to RNA binding, whereas the hydrophobic side chains in β2 are responsible for binding to RNA bases through hydrophobic interactions [60]. In other cases, like hnRNP F, interactions between the RNA target and the RRM were found mainly in the loop region rather than in the β-sheet of the RRM [63, 64].

RRMs usually recognize a short RNA element of 2–5 nt (nucleotide), and some RBPs contain multiple RRMs. The tandem RRMs in the same RBP can either bind to similar RNA sequences and function cooperatively [62, 64] or have very different RNA binding activities/specificities [65], or only one/some of the RRMs are functional while the others do not exhibit RNA binding [66]. Therefore, for RBPs with multiple RRMs, the general rules for how each RRM contributes to binding specificity are largely unclear.

We conducted a detailed sequence analysis of the RRM-containing RBPs in humans and other organisms. Surprisingly, we found a strong trend indicating that RRMs within the same protein (hereafter referred to as “sibling RRMs”) have higher sequence similarity to each other than the RRM pairs from different proteins. In addition, sibling RRMs within the RBPs specific to a single species have higher similarity than those shared by multiple species. Together, these findings suggest that prevalent domain duplications of RRMs have occurred within many RBPs during evolution. This result is further illustrated by cases of both a recent and an ancient RRM duplication. In addition, we found that the RBPs with similar sibling RRMs are more likely to bind to the 3′ UTR than those proteins having more divergent sibling RRMs and that the RBP

sequence regions outside RRMs have a strong bias for low complexity and/or repetitive sequences. Altogether, these analyses reveal important implications regarding RBP evolution.

2.3 Results

2.3.1 Increased numbers of RBPs in mammals.

The number of proteins with canonical RBDs has expanded significantly in mammals. In Table 2.1, we list the common RBDs and the number of proteins containing common RNA-binding domains from five different organisms whose proteomes are thoroughly annotated. Humans have the most RBPs among all species examined, and there is a large expansion in the number of RBPs in mammals with the exception of PAZ domain-containing proteins. In addition, we found that the number of RRMs within a single RBP has increased in mammals compared to other low-complexity organisms when examining the RRM-containing RBPs across different species (Supplemental Figure S2.1). These observations lead to intriguing fundamental questions such as why do humans need so many RBPs and why is it that many RBPs contain multiple RBDs? One possible explanation could be that multiple RBDs allow RBPs to bind RNA with higher sequence specificity and/or affinity than those with a single binding domain. Another possible reason is that multiple domains may help RBPs to bind to longer RNA sequences. On average, a single RBD binds to 4–6 nt; thus, multiple RBDs may have provided some selective advantage for increased binding specificity and affinity and/or also facilitate binding to longer RNA targets.

2.3.2 Sibling RRMs are more similar to each other

To study these questions, we analyzed RRM-containing RBPs at a proteome-wide scale across multiple species. We applied a series of filters to obtain unique human RBPs that have well-defined RRM domains and extracted the sequence of each RRM using the consensus annotation from three domain annotation databases (Figure 2.1A). This process was repeated for three other species (*Mus musculus, Drosophila melanogaster, Caenorhabditis elegans*), and both the species-specific and conserved RBPs were extracted by the same set of filters for further analyses (see Materials and Methods). After filtering for gene duplication and database redundancy, we extracted 453 unique RRMs from human RBPs.

We aligned each of the 453 unique human RRMs to all others to calculate sequence similarity scores and plotted the mean score ± standard deviation (1 × SD) as vertical gray bars (Figure 2.1B), obtaining an average similarity score of ~23. However, similarity scores between sibling RRMs appear to be skewed toward higher similarity (denoted by red circles in Figure 2.1B), indicating that the sibling RRMs within the same RBPs have significantly higher sequence similarity to each other than what is expected by chance ($P = 2.4 \times 10^{-20}$ by Kolmogorov-Smirnov test, or $P = 2.4 \times 10^{-17}$ by t-test if assuming normal distribution) (Figure 2.1B). In particular, among the 1186 sibling RRM pairs, 467 pairs (39.4%) had similarity scores higher than the mean plus 1 × SD, whereas 38 pairs (3.2%) scored below mean −1 × SD. This skewed distribution was not dependent on the score system that we used in measuring similarity, as we observed similar results using additional score methods and matrices (Supplemental Figure S2.2). Further analysis suggested that the increased similarity between sibling RRMs was unrelated to the length of the peptide between these domains, as we did not find any correlation between the RRM similarities and their distances (Supplemental Figure S2.3). This increased

similarity is not limited to a single species, as the same results were obtained when we analyzed the sibling RRMs in the *D. melanogaster* genome (Supplemental Figure S2.4A). In addition to RRM, we also analyzed KH and C2H2 zinc finger domains, both of which are commonly found in human RBPs. While comparing the similarity scores of sibling domain pairs to those of all other pairs (i.e., nonsibling pairs), we again found a higher sequence similarity in sibling pairs in both sibling KH and zinc finger domains (Supplemental Figure S2.4B), suggesting the increased similarity between sibling RNA binding domains is a common feature for different types of RBPs.

There is a possibility that some proteins are under a global selection to preserve certain sequence bias, resulting in the increased sequence similarity between sibling RRM pairs within a single protein compared to random pairs. To measure the potential sequence bias, we calculated the average frequency of each amino acid in RRMs for different RBP groups with 2, 3, 4, and 5 RRM domains (Supplemental Figure S2.5A). These groups include 112 proteins with two RRMs (112 sibling pairs), 44 proteins with three RRMs (132 sibling pairs), 11 proteins with four RRMs (66 sibling pairs), and one protein with five RRMs (10 sibling pairs). Overall, we found that the RRMs from different groups or within the same group have similar sequence composition. Five out of 20 aa have significant differences in mean of frequency between groups as judged by the ANOVA F-statistic (P-value < 0.01). Nevertheless, to better control the subtle sequence bias, we generated control groups of RRM pairs with matched composition distance to the real sibling RRM pairs for the rest of our analyses (see Materials and Methods section for details and Supplemental Figure S2.5B).

We further analyzed the RBPs containing multiple RRMs and compared the cumulative distributions of RRM similarity scores in proteins with different numbers of RRMs. When

compared to a control set of 1000 randomly selected RRM pairings, we found that the different sets of RBPs all have higher similarity between their sibling RRMs than the randomly chosen RRM controls (except the 5-domain RBP set that contains a single member), as judged by the right shifts of plots ($P = 6.3 \times 10{-}13$, $4 \times 10{-}12$, $2.2 \times 10{-}14$, and 0.8 for control vs. 2-, 3-, 4-, and 5-domain RBPs, respectively, by the Kolmogorov–Smirnov test) (Figure 2.1C). This result suggests that the higher similarity between sibling RRMs (Figure 2.1B) is a common property for all RBPs with different numbers of RRMs.

A potential explanation of these observations is that the sibling RRMs resulted from domain duplication during evolution [67]. However, there is an alternative explanation that all the RRMs in proteins with multiple RRMs might be more conserved (i.e., similar to each other) regardless of whether they coexist in the same protein. To address this possibility, we selected the set of RBPs with two or three RRMs and shuffled the sibling relationship of these RRMs within each set. This shuffling of sibling relationships was conducted by randomly selecting two or three RRMs to form a simulated RBP (112 proteins with two RRMs and 44 proteins with three RRMs were generated in each shuffle), and this simulation was repeated 1000 times. We found that the mean similarity scores for shuffled RRM pairs were significantly less than the real sibling pairs ($P = 0.001$ by a rank test) (Figure 2.1D,E), suggesting that the higher similarity observed is, indeed, due to a sibling (duplication) relationship rather than the natural sequence bias between the sets of the "singleton RRMs" and the RRMs with siblings. Consistently, the similarity scores of random pairs of RRMs with siblings (mean = 22 for RBPs with two or three RRMs) (Figure 2.1D,E) are similar to those of random pairs of all RRMs (mean = 23) (Figure 2.1B).

2.3.3 Sibling RRM pairs in species-specific RBPs are more similar to each other

We further examined the sequence conservation of sibling RRM pairs from different species whose proteomes are well annotated. For each of four species (human, mouse, fruit fly, and worm), we selected the RBPs shared among all species and the RBPs found only in one species (see Materials and Methods) and compared the similarity between sibling RRMs within the same protein. We found that, in all species except worms, the similarity between sibling RRMs is significantly higher in the species-specific RBPs as compared to that of sibling RRMs in RBPs shared across all four species. Generally, genes conserved across multiple species are more ancient, as they appeared before speciation, whereas genes unique to certain species are more recently evolved. According to this simple assumption, our finding suggests that the RRM sibling pairs in "younger" (i.e., species-specific) proteins have higher sequence similarities than those in "older" proteins (i.e., conserved across distant species). A simple explanation is again that most sibling RRMs arose from domain duplication during evolution, which was then followed by sequence drift in each species through random mutations. The sibling RRMs in older proteins resulted from more ancient duplication and, therefore, would be expected to have higher sequence divergence. In particular, such increased similarity was more obvious between the sibling RRMs specific to human and mouse (Figure 2.2A,B), suggesting an extensive RRM duplication in mammals. We are aware that our explanation is based on a usual assumption in gene evolution; however, there is an alternative but less likely scenario that the unique genes could have existed in the common ancestor but were subsequently lost in all species except one.

2.3.4 Recent RRM duplication in DAZ proteins.

We observed an outlier with similarity score of 100 between RRMs of human DAZ

proteins (i.e., completely identical). The DAZ proteins have four paralogs on the Y chromosome: DAZ1 (3 RRMs), DAZ2 (1 RRM), DAZ3 (1 RRM), and DAZ4 (2 RRMs); one paralog on chromosome 3: DAZL (DAZ like) (1 RRM), and one on chromosome 2: BOLL (1 RRM) (Figure 2.3A). Among these six proteins, the RRMs in four DAZ proteins and DAZL are completely identical, whereas the RRM in BOLL has 53% identity with the others. Previous sequence analyses suggests that at least two gene duplication events were required to generate this protein family: The first duplication gave rise to DAZL and BOLL, which was followed by a second duplication of DAZL to generate Y chromosome-specific DAZ proteins [68-70]. The second duplication could either be a single duplication that generated four DAZ proteins, or alternatively, several sequential duplications that happened within a short time window so as to produce four proteins.

Among the six proteins within the DAZ family, only human DAZ1 and DAZ4 have multiple RRMs. To improve the annotation of this family, the sequences of the six human DAZ family proteins were compared against the genomes of chimpanzee, macaque, gorilla, chicken, frog, and zebra fish (Figure 2.3A). Such reannotation is necessary, since the nomenclature does not necessarily reflect the real evolutionary route of these proteins in some species (e.g., Dazl in worm is the ortholog of human Boll). The single RRM proteins DAZL and BOLL can be found in all species tested, whereas DAZ proteins with multiple RRMs can only be identified in certain primates (human, chimpanzee, and macaque, but not in gorilla) (Figure 2.3A). This result suggests that there was an RRM domain duplication following the second gene duplication on the Y chromosome, generating new DAZ family members with multiple RRMs. This domain duplication appears to be a recent event that happened only in a subgroup of primates including humans. It is also possible that such domain duplication happens in multiple steps, as the DAZ

proteins with multiple RRMs were detected in macaque but not gorilla. Alternatively, assembly errors in this repetitive region of the Y chromosome could also prevent the detection of DAZ proteins with multiple RRMs in gorilla.

To examine their evolution over a more recent time frame, we further determined the SNP density within the DAZ protein family (Figure 2.3B). We calculated the SNP density (number of SNPs/gene length) for each DAZ gene, as well as the average SNP density of 100 genes randomly selected from the same chromosome (gray bars). The SNP density of BOLL is similar to that of other genes randomly selected from chromosome 2, while the SNP density of DAZL is slightly lower than that of the randomly selected genes on chromosome 3. However, the SNP densities of the four DAZ genes are two orders of magnitude less than the densities of other randomly selected genes on the Y chromosome. Since the majority of gene variation observed in a population is due to random drift of neutral (or nearly neutral) mutations, as proposed by the neutral theory of molecular evolution [71], the SNP density is correlated with the functional importance and evolution time of the gene [72]. Our observation of SNP densities is consistent with the hypothesis that there has been at least one very recent RRM domain duplication event that generated DAZ1 with multiple RRMs.

2.3.5 Ancient RRM duplications in PABPs

In addition to recent domain duplication, we also found a case of ancient duplication of RRMs in the human genome. Human polyadenylate-binding proteins (PABPs) belong to a conserved protein family that binds to the poly(A) tail of mRNA through RRMs [73]. Six PABP paralogs in humans (PABP1, PABP3, PABP4, PABP5, PAP1L, and PAP4L) contain four RRM domains, with some members containing an additional C-terminal domain called PABC. In addition, the human PABP2 and EPAB2 (embryonic PABP2) contain a single RRM, and

PAP1M contains two RRMs (Figure 2.4A). The family of PABP proteins in other species (*M. musculus, D. melanogaster, C. elegans, and Schizosaccharomyces pombe*) contains members with one RRM or four RRMs, with the exception of a yeast protein (PABX) that contains three RRMs. Through multiple sequence alignments of all 21 RRMs in different species, we clustered these RRMs according to similarity and found that these RRMs clustered predominantly by the relative locations in a protein rather than by the species (Figure 2.4B). For example, the first of the four RRMs in all PABPs across five species has higher similarity to each other than to its sibling RRMs, thus forming a monophyletic clade. The same observation is also valid for the second, third, and fourth RRM in different proteins across all species. This relationship was clearly demonstrated in Figure 2.4B, where we color-coded the RRMs by different positions and observed that the RRMs of the same color were mostly clustered together (forming a monophyletic clade) in the phylogenetic tree (Figure 2.4B). The proteins with single RRMs are also clustered with each other across different species, and this clade is more similar to the first of the four RRMs in other proteins. This conservation pattern suggests that the domain duplication generating four sibling RRMs had most likely happened in the common ancestor of all these species (additional duplications might also occur in human and nematode), producing a larger family of PABPs. We speculate that there may be additional ancient domain duplications similar to PABPs, but such events are difficult to identify due to the lack of reliable measurement to distinguish ancient duplication vs. nonduplicated RRMs. For the future work, we may be able to compare the ages of all genes vs. all potentially duplicated RRM domains (with a correct background model for age of the individual domain and entire protein) and thus to determine if there is a correlation between the similarity score of sibling RRMs and the approximate age of the duplication.

These two specific examples in DAZ and PABP families represent both a recent and an ancient RRM duplication, strongly supporting our finding in analyzing all sibling RRMs (Figure 2.1B). Taken together, our results suggest a model wherein RRM duplication has happened frequently during evolution, followed by random evolutionary drift that introduces additional sequence variation. This simple model is consistent with the finding that the number of proteins with multiple RRMs has expanded in humans and other mammals (Supplemental Figure S2.1).

2.3.6 Similarity between sibling RRMs is associated with RBP functions

In addition to the time since duplication, other features might also contribute to the similarities between sibling RRMs. For example, evolutionary constraints can also affect how fast the sequence drifts through random mutations after domain duplication. To study if the similarities between sibling RRMs are associated with certain functional preferences of RBPs, we conducted a survey of functional differences in the RBPs with multiple RRMs. We observed a general trend that the proteins that bind to polyadenylated RNA in the 3′ UTR tend to have more similar sibling RRM pairs, whereas the proteins that bind to the 5′ UTR tend to have dissimilar sibling RRMs (Figure 2.5A), suggesting there may be some association between the similarity of sibling RRM and the RBP function.

To further study this potential relationship, we conducted a gene ontology analysis on all human RBPs having multiple RRMs. According to the similarity scores between each RRM pair, we divided all pairs into six groups, each containing ~100 RRM pairs. The corresponding proteins in each group were subjected to functional enrichment analysis by he DAVID annotation tool (http://david.abcc.ncifcrf.gov/) [74], and the results were compared across all groups (Figure 2.5B). As expected, the function of “singlestranded RNA binding” and “mRNA binding” are significantly enriched across all groups (Figure 2.5B, bottom), serving as a positive

control. Consistent with the earlier observation, we also found a significant enrichment of "mRNA 5′-UTR binding" ($P = 1.8 \times 10{-}5$, fold enrichment = 406) in proteins with dissimilar sibling RRMs (group 1: similarity score = 1–20). In contrast, enrichment of "polyadenylated RNA binding" ($P = 5.1 \times 10{-}5$, fold enrichment = 256) occurred in proteins having sibling RRM pairs with the highest similarity (group 6: similarity score = 42–100). In addition, the RBPs with similar sibling RRMs were also found to be enriched in poly(U) RNA binding, poly-pyrimidine track binding, and poly-purine track binding, suggesting that these RRMs are more likely to bind repetitive RNA elements (groups 4–6) (Figure 2.5B). This finding is consistent with the notion that the requirement of binding to repetitive targets may impose additional selective pressure on these RBPs after RRM duplication. Individual RRMs are known to specifically recognize short sequences (usually 2–5 nt), and thus, RBPs with similar sibling RRMs could be expected to facilitate the binding to longer RNA targets containing repetitive elements.

Compared to other regions of mRNA, the 5′-UTR region usually contains binding sites for factors that affect the translation efficiency of mRNA [75]. On the other hand, the 3′ UTR usually contains more repetitive sequences used to control RNA stability (e.g., AU-rich elements) [76, 77]. As expected, the RBPs with dissimilar sibling RRMs (group 1) are enriched only in 5′-UTR binding and translation regulation (Figure 2.5B). Conversely, proteins with similar sibling RRMs have a small bias toward binding to the 3′ UTR. Recently, a comprehensive identification of the binding motifs for RBPs suggested that the RBDs with higher protein similarity are more likely to bind to similar RNA motifs [53]. Our data raise an interesting prediction that mRNA metabolism is controlled by more diverse elements in the 5′ UTR but by more repetitive elements in the 3′ UTR. This hypothesis seems to be true for translation control and RNA degradation through AU-rich elements, but its generality remains to

be examined.

2.3.7 RRM-containing RBPs are enriched with repetitive motifs

In addition to the RRMs, other sequence motifs may also contribute to RRM-containing RBP function or even RNA binding affinity/specificity [78]. Thus, we analyzed the non-RRM fragments of RBPs to determine their characteristics that may contribute to function. We removed RRM sequences from the RBPs and calculated the frequency of each amino acid in the remaining fragments. To estimate the enrichment of each amino acid, we computed logarithm value for the ratio of amino acid frequency in these fragments vs. that in all human proteins and found that amino acids A, G, P, Q, R, S, and Y were highly enriched (Figure 2.6A). We further searched for the frequent words flanking these enriched amino acids (five residues up- and down- stream) (Figure 2.6B). As expected, we found that RS di-peptides were highly enriched in this data set because the Ser/Arg-rich proteins (SR proteins) are a major class of splicing factors that recognize RNA targets through RRMs. In addition, we found a high frequency of GY di-peptides as well as many other low-complexity poly-G and poly-P sequences. These repetitive motifs were represented by a word cloud plot (Figure 2.6C), where the occurrences of all possible di-, tri- (with arbitrary second amino acid), and tetra-peptides (the second and third amino acids could be any amino acid) were computed after removing the RRM from the RBP sequences. We found that the Gly-rich, Pro-rich, Ser-rich, and Ala-rich sequences often co-occurred with the RRMs (Figure 2.6C); some of these repetitive motifs were also reported in an unbiased identification of all mammalian RBPs [32]. To determine whether these repetitive sequences are specific to RRM-containing proteins, we analyzed sequences of RBPs containing the KH or zinc finger C2H2 domain (Supplemental Figs. S2.6, S2.7). All RBPs with RRM, KH, and zinc finger C2H2 domains have low complexity poly-G and poly-P motifs. Furthermore, we

found the RS di-peptide repeats were only found in RRM-containing proteins, whereas the poly-S was found to be enriched in RBPs with the KH and zinc finger C2H2 domain (cf. Figure 2.6B and Supplemental Figures S2.6B, S2.7B).

2.4 Discussion

Proteins that specifically bind to single-stranded RNA play critical roles in regulating various RNA processing pathways; thus a detailed sequence analysis of these RBPs will provide key insights into gene regulation at the RNA level. This study suggests extensive domain duplications of RRM. Such duplications are probably followed by random evolutionary drift that introduces additional sequence variation, leading to the observed higher degree of sequence divergence in old proteins with ancient RRM duplications (Figures 2.2, 2.4). This domain duplication may play a significant role in the function of RBPs. One possible consequence could be that multiple RRMs allow a protein to bind RNA with higher sequence specificity and/or affinity than those RBPs with a single binding domain. Another consequence could be that multiple RRM domains may help RBPs to bind to longer RNA sequences. Typically a single RRM recognizes 2–5 nt; thus tandem RRMs may provide some selective advantage to increase binding specificity and bind to longer RNA targets. Consistent with this notion, the sibling RRMs in several RBPs, for example, PTB [60], were found to recognize similar RNA motifs. The domain duplication of RRMs might provide a possible explanation of why the RRM-containing proteins are so abundant in the human genome.

The extensive RRM duplication during evolution raises some fundamental questions in RNA biology. The human genome (or mammals, in general) has the highest number of RBPs with RRM duplications, and this RRM expansion probably contributed to the increased

complexity of RNA processing pathways in mammals. For example, the majority of human genes undergo alternative splicing, and a predominant fraction of splicing factors are RBPs with multiple RRM domains. In fact, we observed that the RBPs with different similarities in their sibling RRMs are functionally separated from each other (Figure 2.5). The proteins with very similar sibling RRMs tend to bind the 3′ end of mRNA and might function in RNA polyadenylation, whereas the RBPs with more divergent sibling RRMs tend to bind the 5′ UTR of mRNA and might affect the RNA translation. We speculate that RRM duplication, together with their diverging RNA binding targets in the transcriptome, allows the mutual selection in RNA–protein interaction and eventually leads to the functional divergence of RBPs.

We also found that, compared to all other human proteins, the RRM-containing RBPs are more likely to have repetitive sequences in the regions outside the RRMs. These repetitive sequences frequently mediate protein–protein interactions, as RBPs with low-complexity domains tend to aggregate to form protein fibers [79]. The association of RRM-containing proteins with repetitive sequences (encoded by low complexity DNAs) raises an interesting possibility that these sequences may provide a mechanism for domain duplication, as the repetitive DNA sequences are less stable during replication and tend to cause local DNA duplication/expansions [80, 81]. Alternatively, such repetitive sequences could be a result of RRM duplication that is caused by local DNA duplication; however, the RBPs with a single RRM also contain low-complexity sequences. Nevertheless, the mechanism of domain duplication is an interesting question emerging from our study.

We described a systematic analysis of RBPs, focusing on the proteins with the RRM as their RNA-recognition domain. Surprisingly, we found an increase in the number of RBPs containing multiple RRMs in mammals (Supplemental Figure S2.1) and that the sibling RRMs

within these proteins are more similar to each other than what would be predicted by controls (Figure 2.1). In addition, the sibling RRM pairs are more similar to each other in the species-specific RBPs when compared to the ancient RBPs shared by multiple species, suggesting a general RRM duplication in many genes of the mammalian genome. Such domain duplication is further supported by two extreme examples: In the case of the DAZ protein family, a very recent RRM duplication appears to have happened in humans and several primate species, generating multiple RBPs containing identical sibling RRMs (Figure 2.3). In another case, the RRM duplications within the PABP proteins probably happened in the common ancestor of all eukaryotes, as similar duplication was found from yeast to human (Figure 2.4). Taken together, these results suggested a new and simple model wherein RRM duplication happened frequently during evolution, resulting in increased numbers of RBPs with multiple RRMs.

2.5 Materials and Methods

2.5.1 RNA binding proteins and RNA recognition motifs

We extracted the RRM sequences according to the scheme in Figure 2.1A. First, the RRM sequences were downloaded from InterPro Biomart (http://www.ebi.ac.uk/interpro/biomart /martview/) with the following configuration: DATABASE: "InterPro BioMart," DATASET: "InterPro Entry Annotation," Filters: "InterPro," Entry ID: "IPR000504," and Source Signature Database: "Pfam, SMART, and Prosite." We selected Pfam annotation if there was inconsistency between Source Signature Databases. Unless specified, both SwissProt and TrEMBL proteins were included, but only unique sequences were used. As a result, 453 unique RRMs with peptide sequence length ≥45 amino acids were included (Supplemental Table S2.1). Data of three other species, *M. musculus, D. melanogaster, and C. elegans* were also downloaded for ortholog

analysis (Supplemental Tables S2.2, S2.3). Other protein attributes, such as Gene Ontology (GO), Gene Orthologs, and Gene IDs, used in other databases, were downloaded from Ensembl Biomart (www.ensembl.org/biomart/martview). Because protein IDs are not standardized between or even within some databases, we performed a protein ID conversion as well as manual curation to combine our data sources.

2.5.2 Sequence similarity score calculation

ClustalW2 was used to compute all the pairwise alignment scores for every RRM pair. The similarity score was calculated by calibrating the number of identities between the two sequences with the length of alignment, and it is represented as a percentage, i.e., 0–100. The default protein weight matrix (Gonnet 250) was used for all the pairwise alignments in the main text. However, we also compared the similarity scores generated by using Gonnet 250 with BLOSUM30 (Supplemental Figure S2.2A) and PAM350 (Supplemental Figure S2.2B). We also repeated Figure 2.1B by using BLOSUM 30 as the weight matrix and obtained similar results (Supplemental Figure S2.2C).

2.5.3 Sequence composition and composition distance

We calculated the sequence composition as the frequency of the 20 amino acids in each RRM sequence. Therefore, a sequence composition for an RRM is a vector with 20 dimensions. To measure the similarity of sequence compositions between two RRMs, we used the city block distance between two vectors (i.e., sum of the frequency difference of each amino acid). We named such measurement the "composition distance," which ranges from 0 to 2.

To control for the sequence bias when choosing random RRM pairs, we use those with the composition distance matched to the real sibling pairs. For example, in Figure 2.1C, the 1000

control RRM pairs were randomly picked from all RRM pairs with composition distances within the 0.37–0.60 range (i.e., mean ± 1 SD of the composition distance from real sibling pairs) (see Supplemental Figure S2.5B). In Figure 2.1, D and E, all control RRM-pairs have a composition distance within 0.41–0.62 and 0.34–0.57, respectively.

2.5.4 RBP orthologs

H. sapiens, M. musculus, D. melanogaster, and C. elegans ortholog data were downloaded from the inparanoid database (http://inparanoid.sbc.su.se/download/current/sqltables/) [82]. We downloaded six files, each containing orthologs between two species. We combined all the files and gathered more than 3000 proteins with orthologs found in all four species and thousands of species-specific proteins. These protein sequences were submitted to Pfam for domain analysis with an E-value cutoff of 0.1 (http://pfam.sanger.ac.uk/search#tabview=tab1). Only proteins with more than one predicted RRM were used to calculate the sequence similarity scores. Among the >3000 orthologs between the four species, 80 are RNA-binding proteins, among which 41 human RBPs, 41 mouse RBPs, 34 fly RBPs and 33 worm RBPs contain more than one RRM. We then extracted RBPs that are unique to the individual species and obtained 9, 12, 19, and 12 species-specific RBPs for human, mouse, fly, and worm, respectively. For the sequence similarity score calculation, the sequence pair of RRM from the same RBP were aligned to each other using ClustaW2. All proteins used are listed in Supplementary Tables S2.2 and S2.3. The location of the RRM sequence and the sequence similarity score were also included in the table.

2.5.5 Analyses of DAZ and PABP protein family

To obtain orthologs of the human DAZ protein family, we used inparanoid version 8.0

(http://inparanoid.sbc.su.se/download/8.0_current/Orthologs/). Six species were examined: chimpanzee, macaque, gorilla, chicken, frog, and zebrafish. When no ortholog was annotated in the inparanoid database for selected species, we manually searched the protein sequence database by blast to identify potential orthologs. If there are one DAZ domain and multiple DAZ-like repeats, we classified it as an ortholog of either DAZ3 or DAZ2, since orthologs of these two proteins are hard to distinguish. When the occurrence of the RRM is two, we consider it as a DAZ4 ortholog. If the occurrence of RRM is three, we count it as a DAZ1 ortholog. All the orthologs identified by both inparanoid and manual searches are listed in Supplemental Table S2.4 with their scores and bootstrap probabilities.

The protein sequences of polyadenylate-binding proteins of five species (*H. sapiens, M. musculus, D. melanogaster, C. elegans, and S. pombe*) were downloaded from uniprot, and their RRM sequences were extracted. We used ClustaW2 to build the phylogenetic tree according to multiple sequence alignments (default parameters were used, i.e., Protein Weight Matrix: gonnet, Clustering type: Neighbor-joining). The ClustaW2-generated guide tree file was then visualized via theTreeView program.

Table 2.1 Number of proteins containing different RBDs in five species as reported from Ensembl biomart 0n 07/09/13

Domain name (Interpro ID)	Number of proteins in different species				
	H. sapiens	***M. musculus***	***D. melanogaster***	***C. elegans***	***S. cerevisiae***
RNA recognition motif (IPR000504)	242	248	139	105	54
K Homology domain (IPR004087)	39	39	29	28	9
C2H2 Zinc finger (IPR007087)	805	693	291	176	48
CCCH Zinc finger (IPR000571)	63	50	30	37	10
S1 RNA-binding domain (IPR022967)	9	9	11	6	7
PAZ domain (IPR003100)	10	9	7	29	NULL
Pumilio RNA-binding repeat (IPR001313)	4	4	3	12	7
Total (without any filter)	63,253	38,561	15,628	46,589	7,126

Data set used: *Homo sapiens* (GRCh37.p11), *Mus musculus* (GRCm38.p1), *Drosophila melanogaster* (BDGP5), *Caenorhabditis elegans* (WBcel235), and *Saccharomyces cerevisiae* (EF4). The two Zn finger domains can bind to both RNA and DNA.

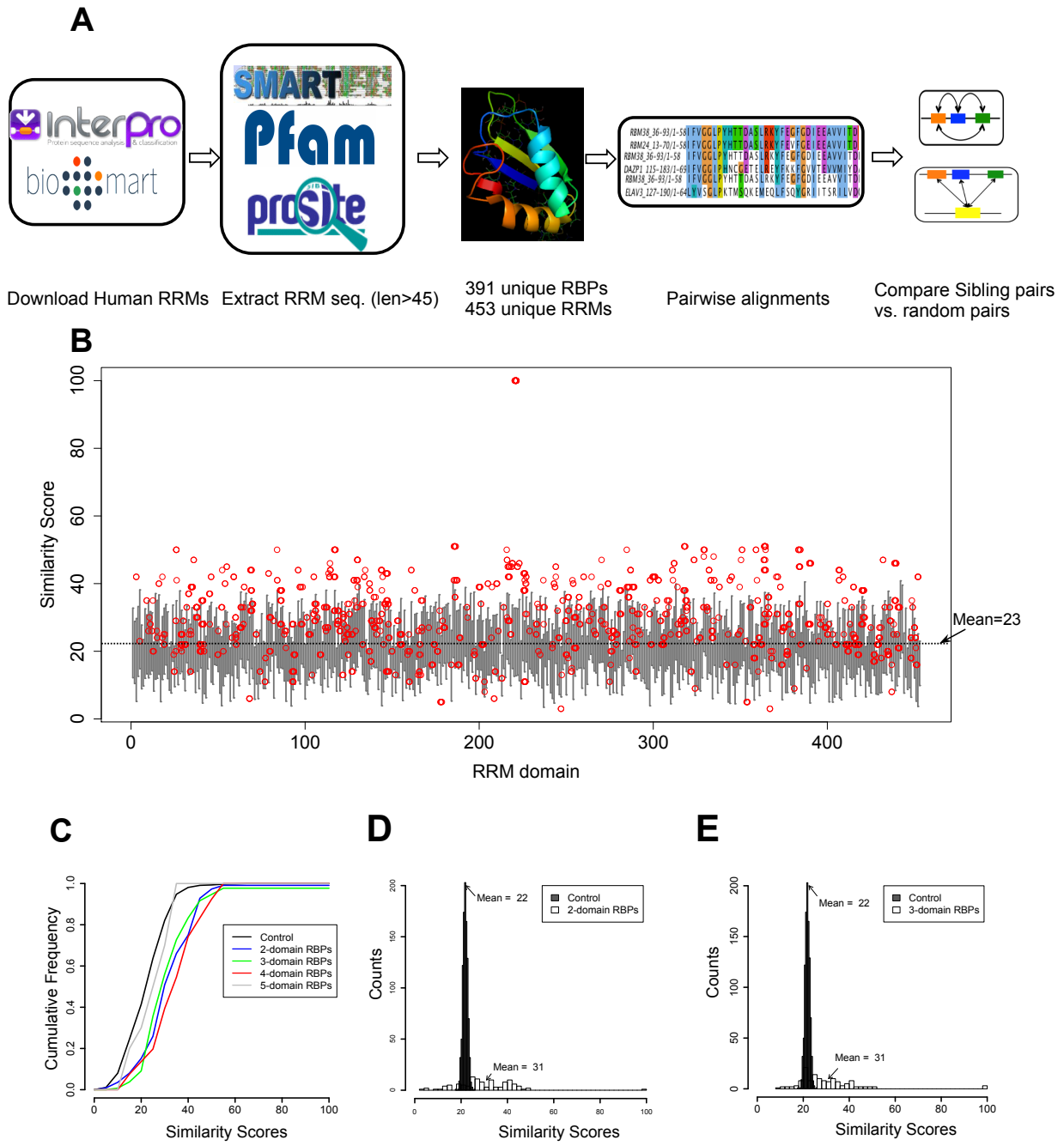

Figure 2.1 Elevated sequence similarity between sibling RRMs in human RBPs.

(*A*) Workflowoftheanalyses.ThehumanproteinscontainingRRMs were obtained from the InterPro database, and the RRM sequences were extracted according to the consensus annotations from three different da- tabases. After filtering out the duplicated sequence, 453 RRMs from 391 unique RBPs were analyzed through sequence comparison. (*B*) Similarity scores between all RRM pairs in human RBPs. Each RRM was aligned with all other 452 RRMs, where the distribution of similarity score is represented by a gray vertical line spanning the mean ± 1× standard derivation. The similarity score between sibling RRMs was represented as a red circle.

The order of RRMs along the x-axis is arbitrary. (*C*) The cumulative frequency of similarity scores between sibling RRM pairs in proteins with 2, 3, 4, or 5 RRMs. As a control, we randomly selected 1000 RRM pairs and computed the cumulative frequency of their similarity scores. (*D*) Sibling RRMs are more conserved than the shuffled pairs. The histograms of similarity scores between sibling RRM pairs from 112 RBPs that contain two RRMs were plotted (open boxes). As controls, we shuffled the order of these RRMs to generate a simulated set of 112 RBPs with matched sequence composition. The shuffle was repeated 1000 times with replacement, and the mean similarity scores of RRM pairs were plotted as filled boxes. (*E*) Same as panel D, except 44 RBPs with three RRMs were analyzed.

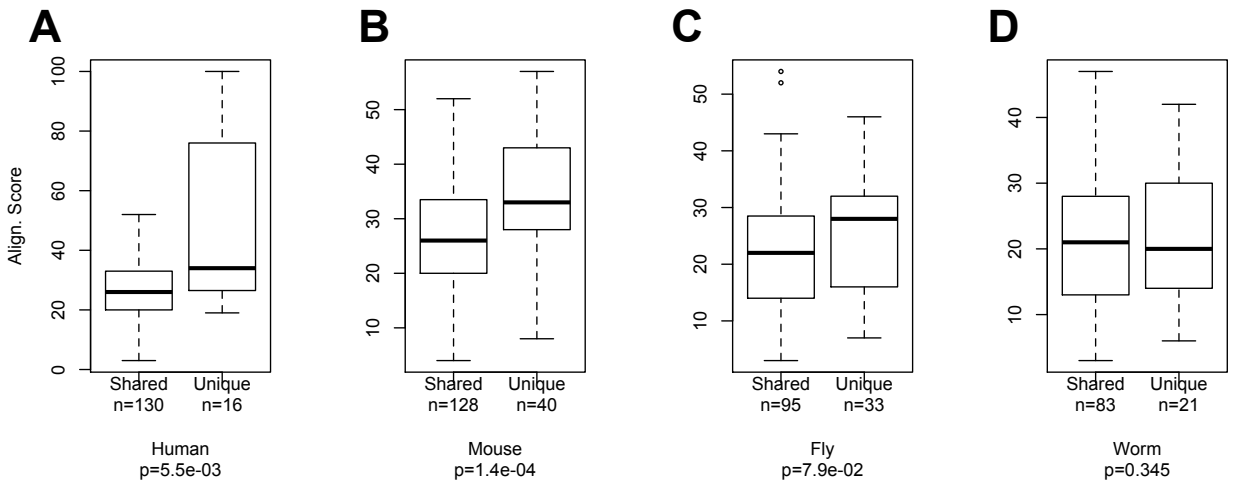

Figure 2.2 Sibling RRM pairs in species-specific proteins are more conserved.

(*A*) HumanRBPs with multiple RRMs were divided into two classes: the proteins shared among four different species (H. sapiens, M. musculus, D. melanogaster, and C. elegans) and the proteins found only in human. The similarity scores between sibling RRMs were calculated for each class and represented as a box plot. The score distributions were compared by t-test with P value indicated. The same analyses were also carried out using RBPs from M. musculus (*B*), D. melanogaster (*C*), and C. elegans (*D*).

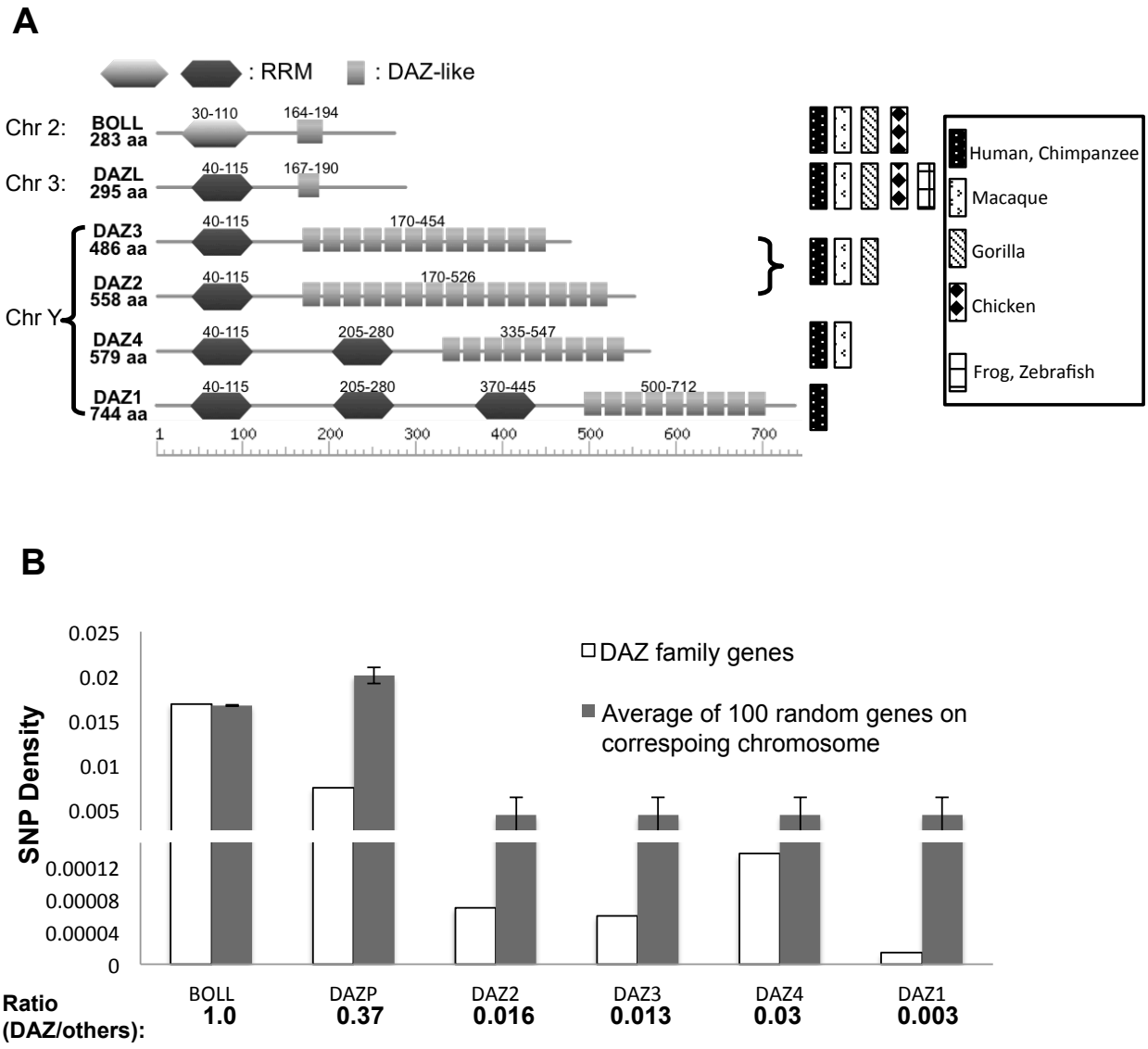

Figure 2.3 An RBP family with recent RRM duplications.

(*A*) ThemembersinthehumanDAZ protein family contain one or more RRMs and DAZ-like domains. All RRMs in DAZ1, DAZ2, DAZ3, DAZ4, and DAZL are identical, whereas the RRM in BOLL has 53% sequence identity with the other RRM. The DAZ1 to DAZ4 are in the Y chromosome, while BOLL and DAZL are in chromosomes 2 and 3. The ortholog genes in other species were identified by a combination of inparanoid annotation and blast search, and species that contain various DAZ proteins were represented with different boxes. The DAZ proteins with multiple RRMs were only found in cer- tain primates. (*B*) The SNP density of each human DAZ protein was compared with the average density of other genes in the same chromosome. The SNP density ratios between DAZ genes relative to other genes in the same chromosome are indicated. The genes in the Y chromo- some encoding DAZ proteins have lower SNP density, suggesting that they are more recently di- verged genes.

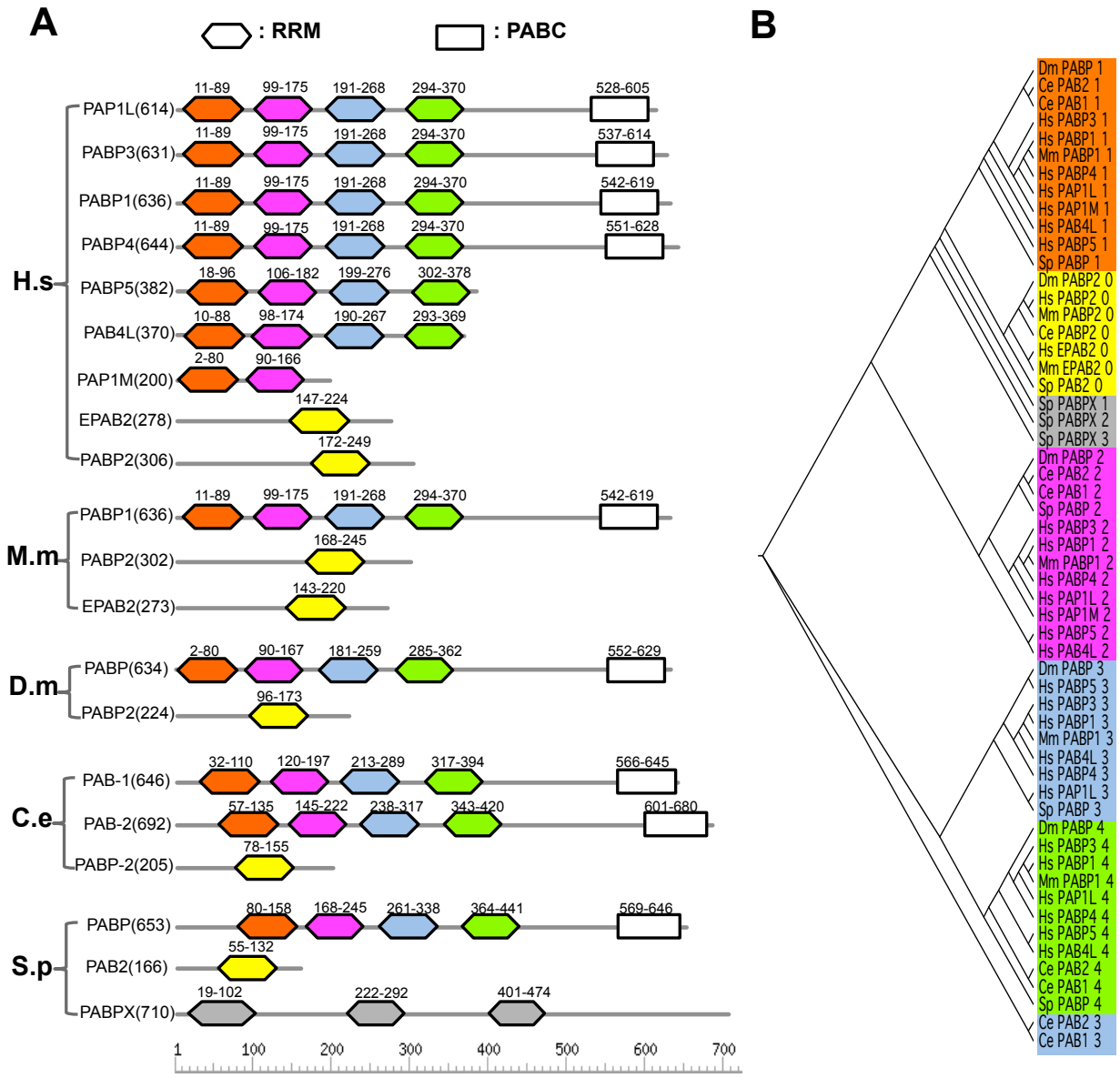

Figure 2.4 An RBP family with ancient RRM duplications.

(*A*) The diagram of PABP proteins from five species (H. sapiens, M. musculus, D. melanogaster, C. elegans, and S. pombe). The members in this protein family contain one to four RRMs, and some also contain a C-terminal PABC domain. Each RRM is colored according to their relative positions within the protein. (*B*) The phylogenetic tree of RRMs in the PABP family was visualized via TreeView. The RRMs are colored in the same scheme as in panel A, and the RRMs in the same position are more similar to each other across all species.

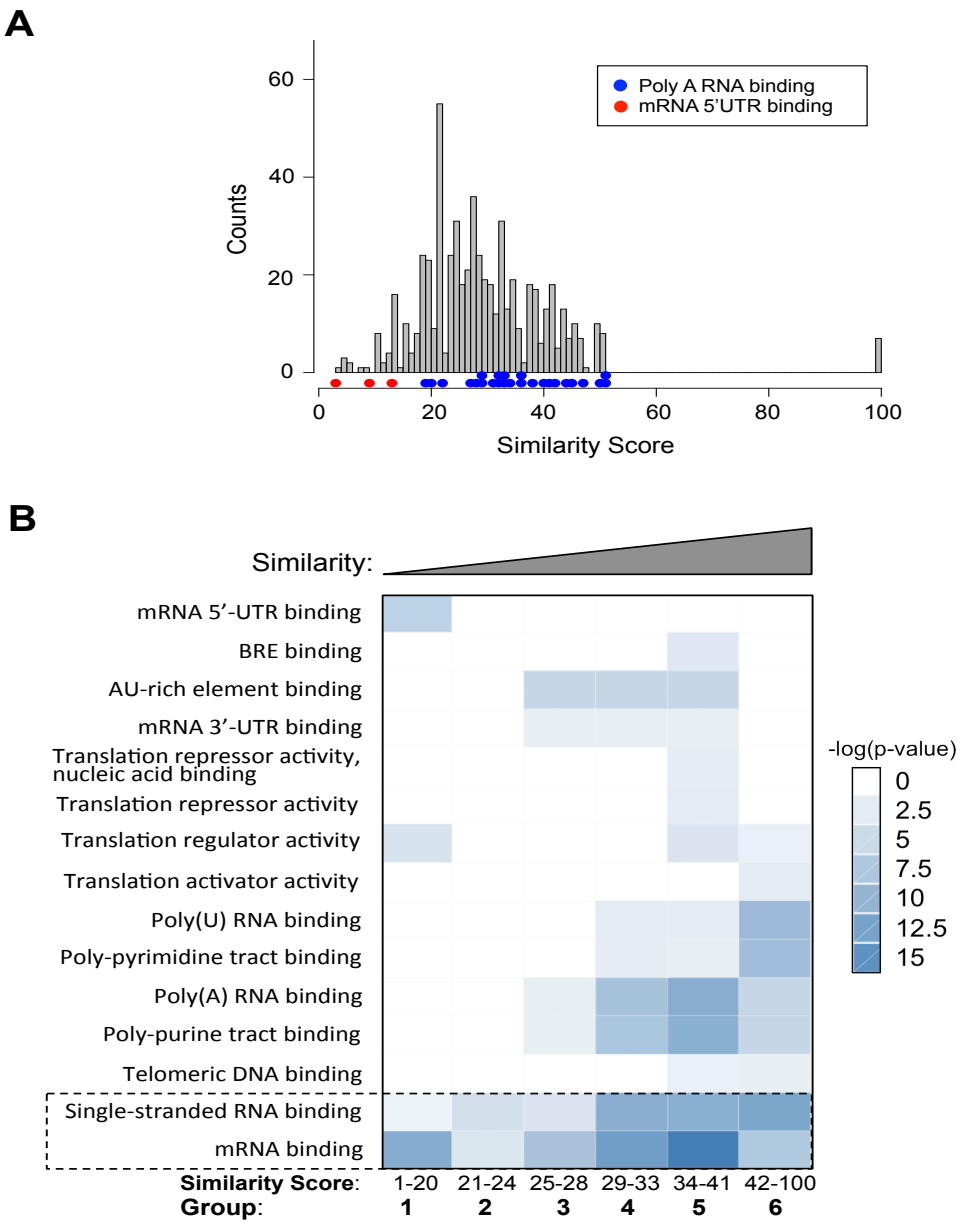

Figure 2.5 Gene Ontology analysis of human RBPs with multiple RRMs.

(*A*) Sibling RRMs with different similarities tend to bind distinct regions of mRNA. The similarities between sibling RRM pairs are rep- resented with a histogram (gray), with the colored dots indicating the gene ontology (GO) terms enriched in the genes from different bins of the histogram. (*B*) According to the domain similarity score between sibling RRMs, all RBPs were divided into six groups as equally as possi- ble: 1–20 (108 pairs), 21–24 (92 pairs), 25–28 (106 pairs), 29–33 (104 pairs), 34–41 (97 pairs), and 42–100 (93 pairs). The GO analyses were carried out, and the enriched functional terms in each bin are represented with a heat map to indicate the significance of enrichment. The func- tions common to all groups are marked.

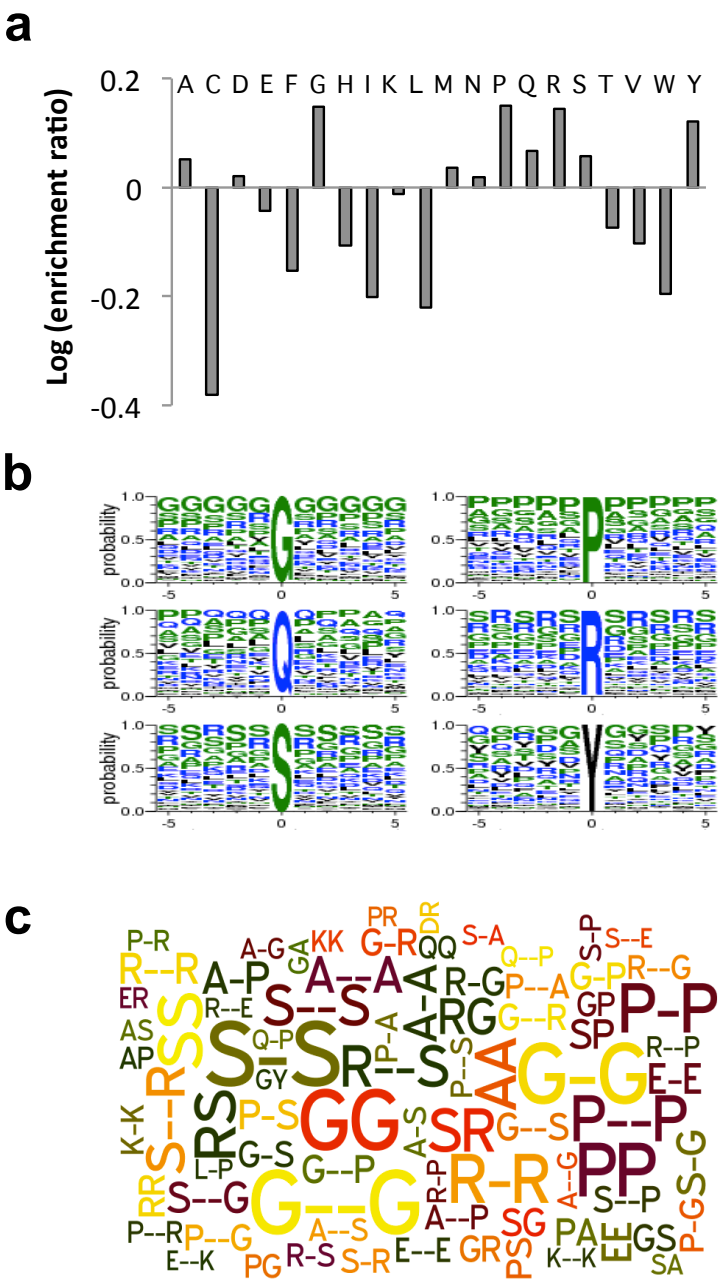

Figure 2.6 Sequence motifs enriched in the RRM-containing RBPs.

(*A*) We removed the RRM sequence from the RBPs and analyzed the re- maining sequence for amino acid propensities. For all 20 amino acids, their frequencies within non-RRM regions were compared to other pro- teins in the human proteome and the relative ratio is plotted. (*B*) Sequence logos around the most enriched amino acid residues in RBPs. The height of each single-letter amino acid code corresponds to the probability of occurrence at each position. (*C*) Repetitive sequence patterns that significantly co-occur with RRM in all human proteins. The size of each pattern corresponds to the number of occurrence. The word cloud was generated with the Wordle online tool. The top 80 motifs are shown.

2.6 Supplementary Material

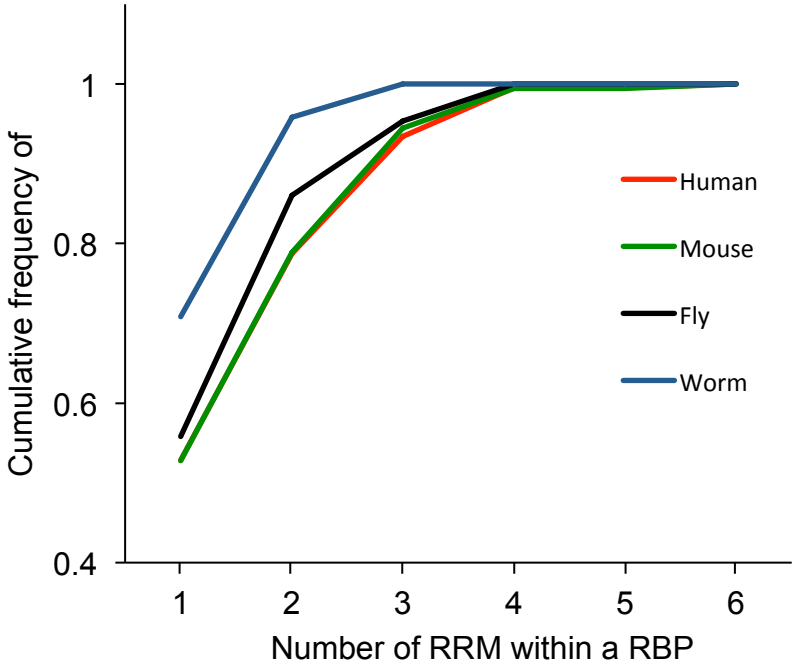

Figure S2.1 Increased numbers of RRMs within a single RBP in mammals.

Cumulative frequency of RBP with different numbers of RRM was plotted in four species (*H. sapien*, *M. musculus*, *D. melanogaster* and *C. elegans*).

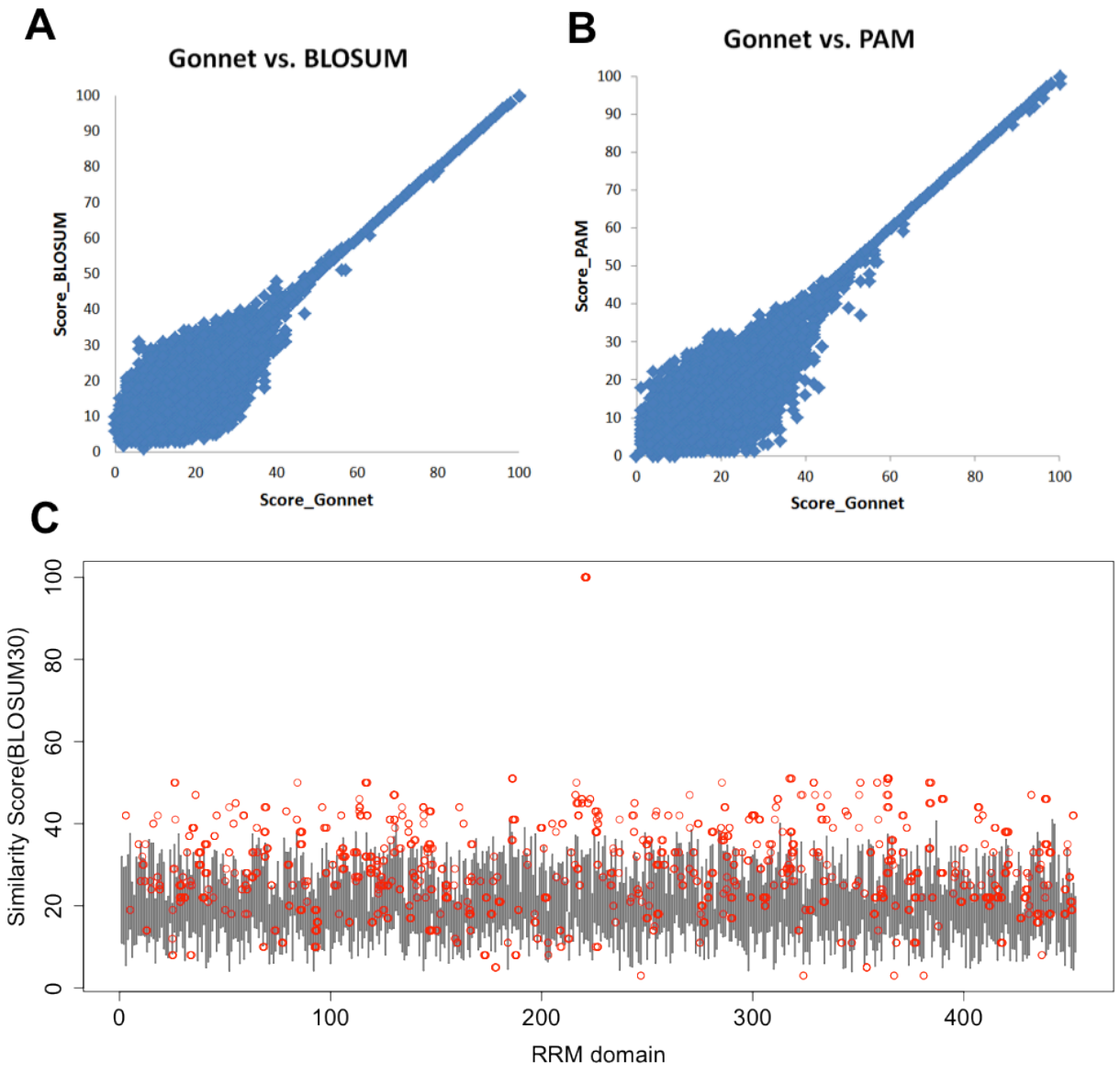

Figure S2.2 The similarity scores of RRMs measured with different scoring matrices.

To control for possible variations in our scoring schemes, we used different scoring matrices (Gonnet250, BLOSUM30 and PAM350) to measure similarity scores between all 453 RRM pairs in RBPs. (A) Scatter plot of scores measured by Gonnet250 vs. BLOSUM30. (B) Scatter plot of scores measured by Gonnet250 vs. PAM350. (C) Similarity scores between all RRM pairs in human RBPs. The data is plotted like Figure 1B, except the BLOSUM30 was used as the scoring matrix instead of Gonnet250.

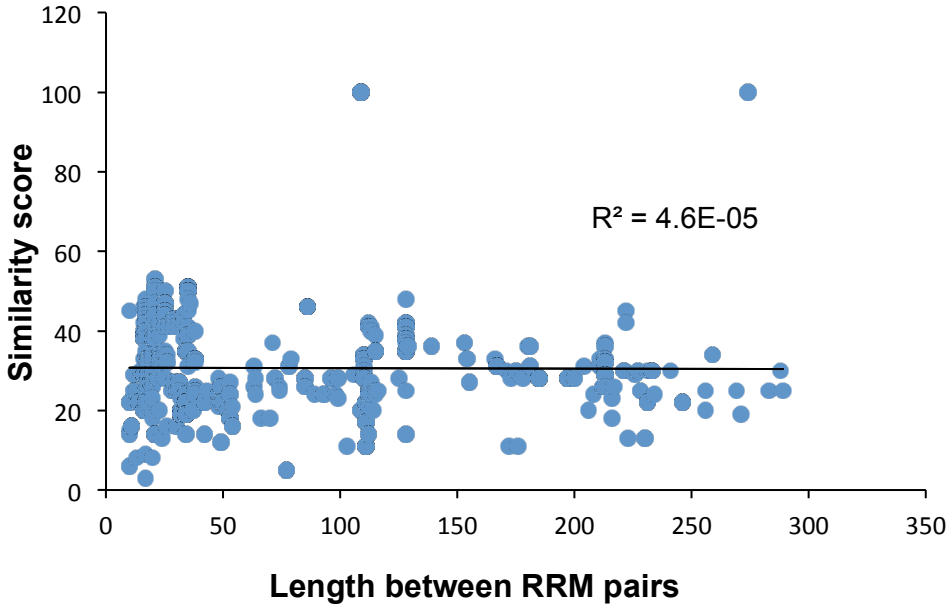

Figure S2.3 The lengths between sibling RRMs do not affect the similarity.

The similarity scores between sibling RRMs is plotted against the length (i.e. number of amino acid) between sibling RRMs. No correlation is found between the length and similarity scores.

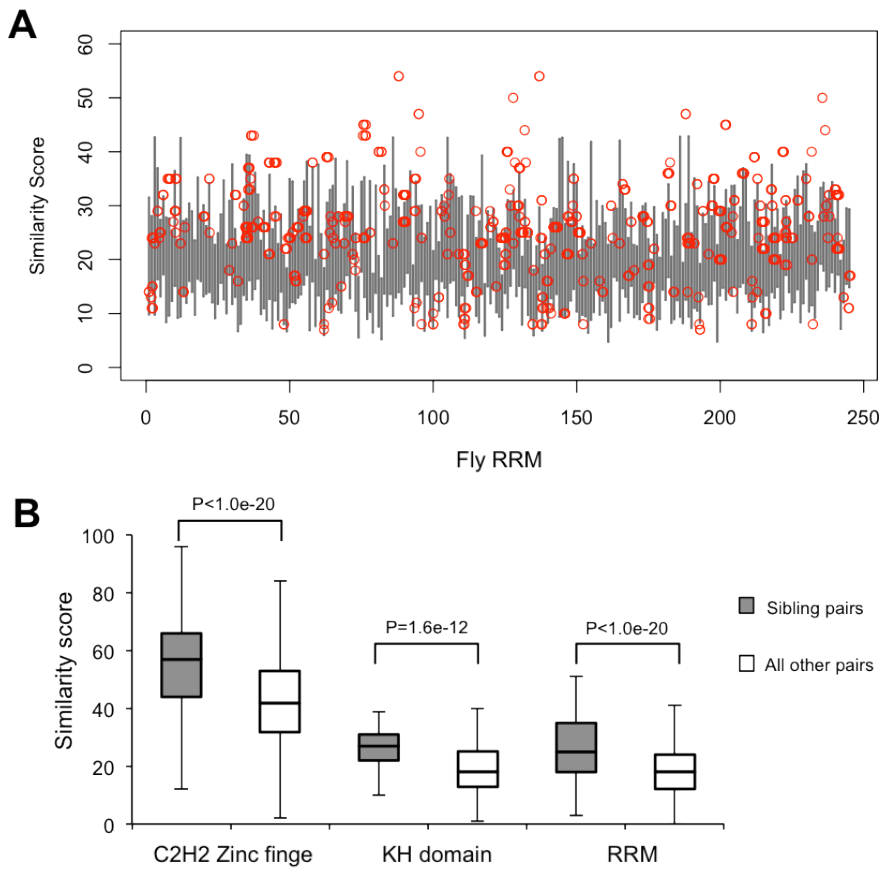

Figure S2.4 Similarity scores between RRM pairs in *D. melanogaster* RBPs and pairs of KH domain in human RBPs.

(A) Each RRM was aligned with all other RRMs in *D. melanogaster*, where the distribution of similarity score is represented by a grey vertical line spanning the mean ± 1× standard derivation. For proteins with multiple RRMs, the similarity score between the sibling RRMs was represented as a red circle. The order of RRM along the x-axis is arbitrary. (B) Boxplots of sequence similarities for both sibling domain pairs (gray box) and all other non-sibling domain pairs (white box) were shown in the three types of RNA binding domains. The sequences of C2H2 Zinc finger, KH domains and RRMs were extracted from human proteins.

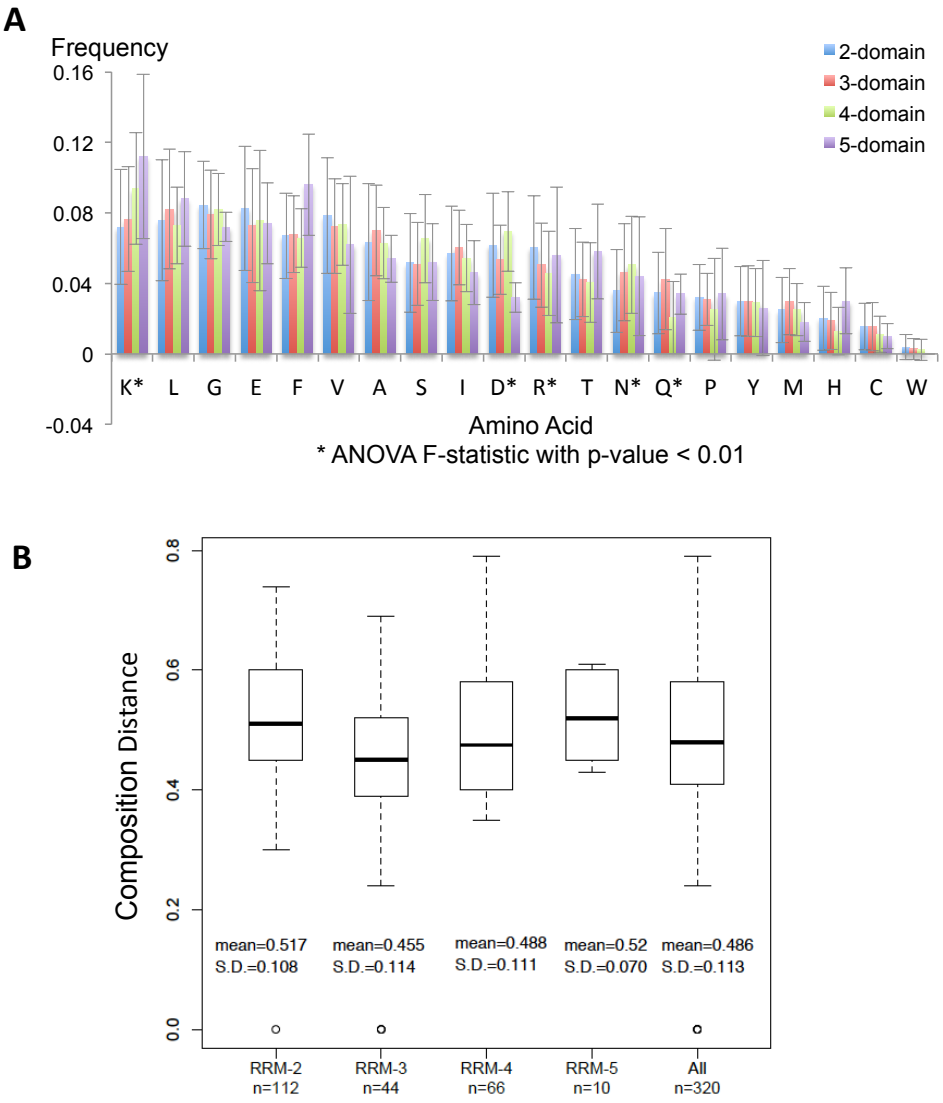

Figure S2.5 Amino acid composition frequency and composition distance in real RRM-pairs.

(A) Amino acid frequencies of each RRM were calculated. We plotted the mean ±1 S.D. for RRMs in each group and compare the difference between groups using ANOVA test. Compositions that are significantly different between groups are denoted with *. (B) Composition distances between real RRM-pairs were calculated in each group and box plot of the distribution were plotted. We also listed the mean and standard deviation (S.D.) of the distribution.

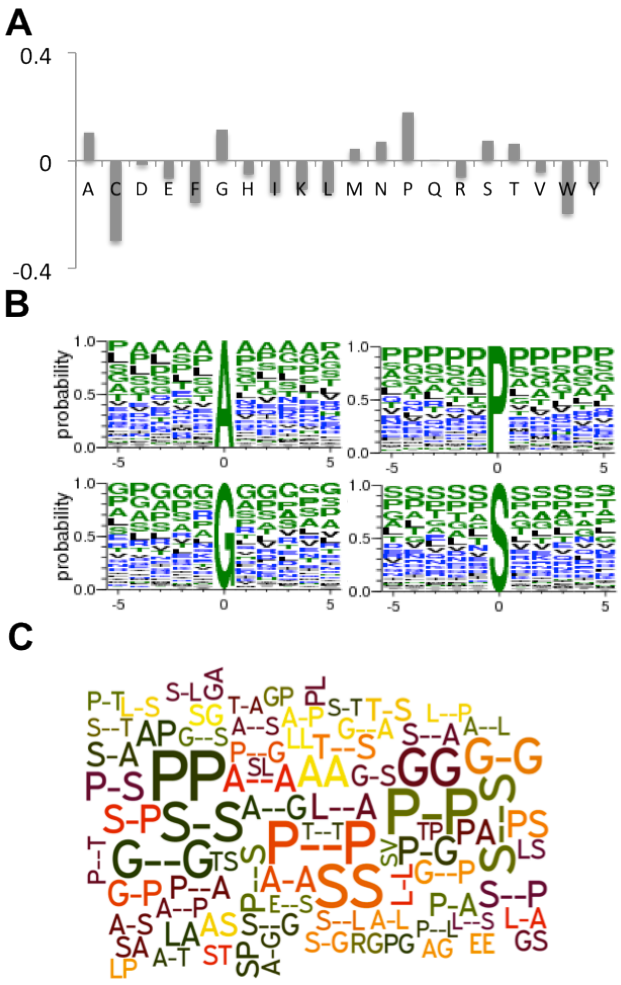

Figure S 2.6 Sequence motifs enriched in human RBPs containing KH domain(s).

(A) We removed the KH sequences from the RBPs and analyzed the remaining sequence for amino acid propensities. For all 20 amino acids, their frequencies in non-KH regions were compared to other proteins in the human proteome and the relative ratio plotted. (B) Sequence logos around the most enriched amino acid residues in RBPs. The height of each single-letter amino acid code corresponds to the probability of occurrence at each position. (C) Repetitive sequence patterns that significantly co-occur with KH in all human proteins. The size of each pattern corresponds to the number of occurrence. The top 80 motifs are shown.

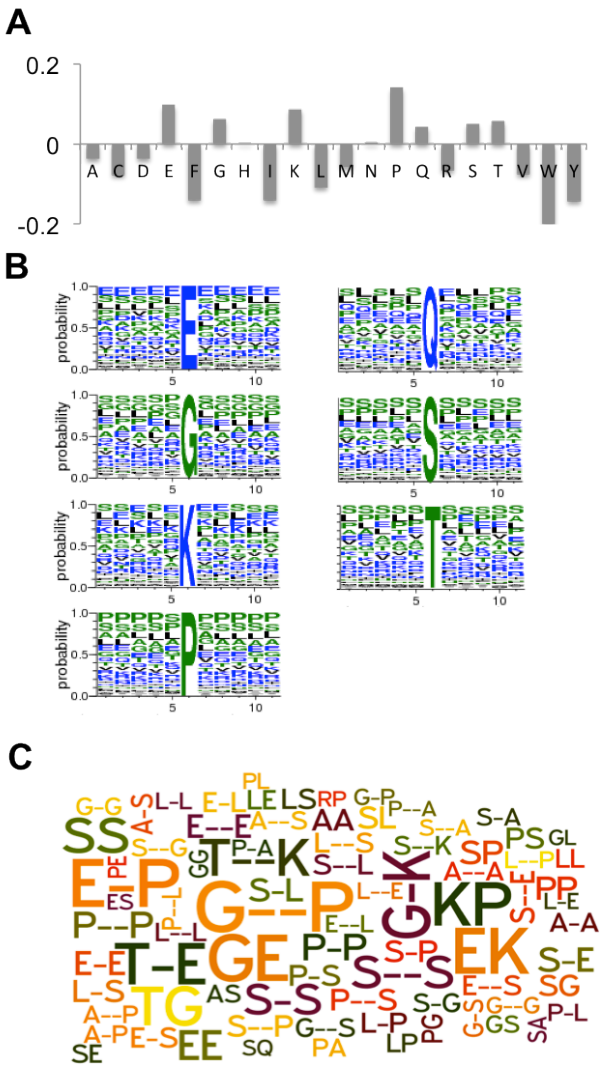

Figure S2.7 Sequence motifs enriched in human RBPs containing Zinc finger C2H2 domain(s).

(A) We removed the Zinc finger C2H2 sequences from the RBPs and analyzed the remaining sequence for amino acid propensities. For all 20 amino acids, their frequencies within non-Zinc finger C2H2 regions were compared to other proteins in the human proteome and the relative ratio plotted. (B) Sequence logos around the most enriched amino acid residues in RBPs. The height of each single-letter amino acid code corresponds to the probability of occurrence at each position. (C) Repetitive sequence patterns that significantly co-occur with Zinc finger C2H2 in all human proteins. The size of each pattern corresponds to the number of occurrence. The top 80 motifs are shown.

Table S 2.1 Detailed information of RRM-containing RBPs in human

UniProt/ SwissProt ID	Associated Gene Name	No. of RRM	RRM position	RRM align score(s)	PFAM ID
SRSF9_HUMAN	SRSF9	2	113-174, 16-83	16	PF00076
RBM46_HUMAN	RBM46	3	63-126, 143-204, 238-301	25 14 22	PF00076
PM14_HUMAN	AC008073.5	1	21-87	NULL	PF00076
PPIL4_HUMAN	PPIL4	1	242-312	NULL	PF00160 PF00076
RAVR1_HUMAN	RAVER1	2	61-123, 135-203	25	PF00076
U2AF2_HUMAN	U2AF2	3	400-459, 151-224, 261-330	11 18 20	PF00076
RBM4B_HUMAN	RBM4B	2	4-64, 81-141	44	PF00076 PF00098
RBM23_HUMAN	RBM23	2	168-233, 265-334	19	PF00076
DAZ3_HUMAN	DAZ2	1	42-98	NULL	PF00076
CPEB2_HUMAN	CPEB2	1	334-395	NULL	PF00076
RBM11_HUMAN	RBM11	1	12-80	NULL	PF00076
RBY1D_HUMAN	RBMY1D	1	10-79	NULL	PF08081 PF00076
RBM47_HUMAN	RBM47	3	73-137, 153-214, 248-311	29 17 19	PF00076
RBM18_HUMAN	RBM18	1	27-98	NULL	PF00076
FUS_HUMAN	FUS	1	287-365	NULL	PF00641 PF00076
MK67I_HUMAN	MKI67IP	1	47-116	NULL	PF12196 PF00076
IF2B1_HUMAN	IGF2BP1	2	85-150, 4-68	3	PF00013 PF00076
RBY1A_HUMAN	RBMY1A1	1	10-79	NULL	PF08081 PF00076
IF4H_HUMAN	EIF4H	1	45-112	NULL	PF00076
HNRH1_HUMAN	HNRNPH1	3	15-82, 293-355, 115-181	31 44 41	PF00076 PF08080
RBMS3_HUMAN	RBMS3	2	63-123, 142-206	27	PF00076
RBMXL_HUMAN	RBMXL1	1	10-80	NULL	PF00076 PF08081
PTBP2_HUMAN	PTBP2	3	459-518, 183-247, 340-405	26 21 24	PF00076
PRGC1_HUMAN	PPARGC1A	1	679-739	NULL	PF00076
PABP5_HUMAN	PABPC5	4	20-88, 108-175, 304-372, 201-269	23 23 24 36 32 39	PF00076
SART3_HUMAN	SART3	2	706-774, 803-871	27	PF00076 PF05391
RBM14_HUMAN	RBM14	2	81-142, 3-60	50	PF00076
REXON_HUMAN	AC004381.6	1	510-571	NULL	PF00929 PF00076
SR140_HUMAN	U2SURP	1	276-348	NULL	PF01805 PF00076
RBM5_HUMAN	RBM5	1	100-163	NULL	PF01585 PF00076 PF00641
HNRPR_HUMAN	HNRNPR	3	167-231, 343-404, 248-309	25 22 22	PF00076
SRS12_HUMAN	SRSF12	1	12-81	NULL	PF00076
ENOX1_HUMAN	ENOX1	1	144-199	NULL	PF00076
SREK1_HUMAN	SREK1	1	68-136	NULL	PF00076
RB12B_HUMAN	RBM12B	2	402-470, 288-353	12	PF00076
EPAB2_HUMAN	PABPN1L	1	149-218	NULL	PF00076
SLIRP_HUMAN	SLIRP	1	22-84	NULL	PF00076
EIF3G_HUMAN	EIF3G	1	241-303	NULL	PF12353 PF00076
RBM45_HUMAN	RBM45	3	34-92, 406-457, 127-183	26 35 13	PF00076

TDR10_HUMAN	TDRD10	1	36-100	NULL	PF00567 PF00076
RFOX1_HUMAN	RBFOX1	1	120-187	NULL	PF12414 PF00076
RBM34_HUMAN	RBM34	1	289-358	NULL	PF00076
DAZ2_HUMAN	DAZ2	1	42-98	NULL	PF00076
ROA1_HUMAN	HNRNPA1	2	107-175, 16-84	33	PF00076
PPRC1_HUMAN	PPRC1	1	1545-1604	NULL	PF00076
PABP1_HUMAN	PABPC1	4	101-168, 296-363, 193-261, 13-83	41 47 33 51 36 31	PF00076 PF00658
CELF4_HUMAN	CELF4	2	154-216, 56-123	41	PF00076
SET1A_HUMAN	SETD1A	1	101-165	NULL	PF11764 PF00856 PF00076
DAZP1_HUMAN	DAZAP1	2	115-183, 12-80	31	PF00076
TSAP1_HUMAN	TRNAU1AP	2	5-73, 98-161	28	PF00076
RBP56_HUMAN	TAF15	1	236-314	NULL	PF00076 PF00641
CELF5_HUMAN	CELF5	3	136-198, 402-471, 47-113	31 42 25	PF00076
PAP1L_HUMAN	PABPC1L	4	101-168, 193-261, 296-363, 13-83	50 41 29 48 33 29	PF00076 PF00658
CPSF7_HUMAN	CPSF7	1	84-155	NULL	PF00076
CPEB4_HUMAN	CPEB4	1	474-535	NULL	PF00076
RBY1B_HUMAN	RBMY1B	1	10-79	NULL	PF08081 PF00076
IF2B3_HUMAN	IGF2BP3	2	85-150, 4-68	9	PF00013 PF00076
HNRDL_HUMAN	HNRPDL	2	151-218, 235-296	40	PF00076
U1SBP_HUMAN	SNRNP35	1	53-122	NULL	PF00076
G3BP2_HUMAN	G3BP2	1	333-390	NULL	PF02136 PF00076
LARP7_HUMAN	LARP7	1	127-188	NULL	PF08777 PF05383 PF00076
HNRPC_HUMAN	HNRNPC	1	18-80	NULL	PF00076
SRS11_HUMAN	SRSF11	1	37-101	NULL	PF00076
CELF1_HUMAN	CELF1	3	403-472, 19-84, 110-175	22 25 34	PF00076
PSPC1_HUMAN	PSPC1	2	84-148, 158-216	22	PF00076 PF08075
HTSF1_HUMAN	HTATSF1	2	289-343, 135-212	5	PF00076
MYEF2_HUMAN	MYEF2	3	235-303, 525-592, 102-171	45 28 39	PF00076
SRSF1_HUMAN	SRSF1	2	18-85, 123-184	25	PF00076
HNRPQ_HUMAN	SYNCRIP	3	164-229, 245-306, 340-401	20 22 19	PF00076
RBMS2_HUMAN	RBMS2	2	137-201, 58-118	29	PF00076
RBM38_HUMAN	RBM38	1	36-93	NULL	PF00076
CNOT4_HUMAN	CNOT4	1	130-187	NULL	PF00076
PPIE_HUMAN	PPIE	1	8-78	NULL	PF00160 PF00076
RBM25_HUMAN	RBM25	1	89-157	NULL	PF01480 PF00076
RDM1_HUMAN	RDM1	1	32-88	NULL	PF00076 PF04098
RBMS1_HUMAN	RBMS1	2	64-124, 143-207	27	PF00076
CELF2_HUMAN	CELF2	3	134-198, 425-494, 43-111	30 33 21	PF00076
MTHSD_HUMAN	MTHFSD	1	308-372	NULL	PF01812 PF00076
RBY1E_HUMAN	RBMY1E	1	10-79	NULL	PF08081 PF00076
DAZL_HUMAN	DAZL	1	42-108	NULL	PF00076
SRSF6_HUMAN	SRSF6	2	4-64, 112-177	24	PF00076 PF08777
RA1L2_HUMAN	HNRNPA1L2	2	107-175, 16-84	33	PF00076 PF11627
SRSF7_HUMAN	SRSF7	1	13-77	NULL	PF00076

NCB2L_HUMAN	NCBP2L	1	43-104	NULL	PF00076
ROA0_HUMAN	HNRNPA0	2	9-75, 100-168	35	PF00076
SRS10_HUMAN	BX511012.1 SRSF10	1	12-81	NULL	PF00076
ROA3_HUMAN	HNRNPA3	2	37-105, 128-196	34	PF00076
SRSF2_HUMAN	SRSF2	1	18-86	NULL	PF00076
IF4B_HUMAN	EIF4B	1	99-167	NULL	PF00076
RBM41_HUMAN	RBM41	1	311-379	NULL	PF00076
SET1B_HUMAN	SETD1B	1	110-174	NULL	PF11764 PF00856 PF00076
NCBP2_HUMAN	NCBP2	1	42-112	NULL	PF00076
HNRH2_HUMAN	HNRNPH2	3	15-82, 293-355, 115-181	33 44 41	PF00076 PF08080
CPEB3_HUMAN	CPEB3	1	443-504	NULL	PF00076
TADBP_HUMAN	TARDBP	2	193-241, 106-171	28	PF00076
RAVR2_HUMAN	RAVER2	3	71-132, 235-302, 145-209	11 8 16	PF00076
HNRGT_HUMAN	RBMXL2	1	10-80	NULL	PF00076 PF08081
RBM44_HUMAN	RBM44	1	833-896	NULL	PF00076
ZCRB1_HUMAN	ZCRB1	1	12-82	NULL	PF00076 PF00098
RBY1F_HUMAN	RBMY1F RBMY1J	1	10-79	NULL	PF08081 PF00076
IF2B2_HUMAN	IGF2BP2	2	5-62, 86-148	13	PF00013 PF00076
SCAF8_HUMAN	SCAF8	1	479-544	NULL	PF04818 PF00076
MSI2H_HUMAN	MSI2	2	24-91, 112-178	44	PF00076
ENOX2_HUMAN	ENOX2	1	130-186	NULL	PF00076
RBM3_HUMAN	RBM3	1	8-78	NULL	PF00076
A1CF_HUMAN	A1CF	3	233-296, 138-199, 58-124	24 20 22	PF00076
ZN638_HUMAN	ZNF638	1	679-730	NULL	PF00076
U2AF4_HUMAN	U2AF1L4	1	88-140	NULL	PF00642 PF00076
RBM8A_HUMAN	RBM8A	1	75-144	NULL	PF00076 PF08777
SNRPA_HUMAN	SNRPA	2	210-272, 12-82	14	PF00076
HNRPL_HUMAN	HNRNPL	2	106-155, 204-257	26	PF00076
CPSF6_HUMAN	CPSF6	1	83-152	NULL	PF00076
RBM22_HUMAN	RBM22	1	234-298	NULL	PF00076
RU2B_HUMAN	SNRPB2	2	153-212, 9-79	26	PF00076
PABP3_HUMAN	PABPC3	4	101-168, 13-83, 296-363, 193-261	32 38 44 29 28 51	PF00076 PF00658
ROA2_HUMAN	HNRNPA2B1	2	114-181, 23-91	39	PF00076 PF11627
EIF3B_HUMAN	EIF3B	1	207-257	NULL	PF08662 PF00076
TRA2B_HUMAN	TRA2B	1	122-190	NULL	PF00076
RFOX2_HUMAN	RBFOX2	1	124-191	NULL	PF12414 PF00076
ROAA_HUMAN	HNRNPAB	2	72-139, 155-224	39	PF00076 PF08143
PUF60_HUMAN	PUF60	2	131-201, 228-297	41	PF00076
THOC4_HUMAN	THOC4	1	108-176	NULL	PF00076
ELAV3_HUMAN	ELAVL3	3	127-190, 41-111, 286-355	39 28 30	PF00076
U2AFL_HUMAN	ZRSR1	1	246-302	NULL	PF00642 PF00076
SFR15_HUMAN	SCAF4	1	510-573	NULL	PF04818 PF00076
RBM15_HUMAN	RBM15	1	376-444	NULL	PF07744 PF00076
DND1_HUMAN	DND1	1	60-123	NULL	PF00076

PTBP1_HUMAN	PTBP1	3	339-405, 186-250, 458-518	24 24 24	PF00076
SAFB1_HUMAN	SAFB	1	409-477	NULL	PF00076 PF02037
CIRBP_HUMAN	CIRBP	1	8-78	NULL	PF00076
SFPQ_HUMAN	SFPQ	2	374-433, 299-363	16	PF00076 PF08075
DJC17_HUMAN	DNAJC17	1	189-233	NULL	PF00226 PF00076
G3BP1_HUMAN	G3BP1	1	342-396	NULL	PF02136 PF00076
RU17_HUMAN	SNRNP70	1	105-174	NULL	PF12220 PF00076
RNPS1_HUMAN	RNPS1	1	165-234	NULL	PF00076
SRSF5_HUMAN	SRSF5	2	6-67, 110-175	25	PF00076
MSI1H_HUMAN	MSI1	2	111-177, 23-90	47	PF00076
RALYL_HUMAN	RALYL	1	23-85	NULL	PF00076
HNRPF_HUMAN	HNRNPF	2	115-181, 293-356	42	PF00076 PF08080
DAZ4_HUMAN	DAZ1	2	207-263, 42-98	100	PF00076
GRSF1_HUMAN	GRSF1	1	254-319	NULL	PF00076
DAZ1_HUMAN	DAZ1	3	207-263, 372-428, 42-98	100 100 100	PF00076
RBM12_HUMAN	RBM12	3	432-498, 859-928, 548-613	22 22 22	PF00076
NELFE_HUMAN	RDBP	1	269-323	NULL	PF00076
U2AFM_HUMAN	ZRSR2	1	241-297	NULL	PF00642 PF00076
PABP4_HUMAN	PABPC4	4	13-83, 296-363, 101-168, 193-261	32 29 34 42 50 45	PF00076 PF00658
TIA1_HUMAN	TIA1	3	108-178, 9-77, 216-280	27 33 36	PF00076
ROD1_HUMAN	ROD1	3	479-539, 184-248, 360-426	22 18 24	PF00076
HNRPG_HUMAN	RBMX	1	10-80	NULL	PF00076 PF08081
RBM39_HUMAN	RBM39	2	155-220, 252-322	19	PF00076
SRSF4_HUMAN	SRSF4	2	106-171, 4-64	22	PF00076
RBM19_HUMAN	RBM19	5	834-904, 4-72, 732-804, 404-473, 302-362	30 25 24 22 31 28 13 34 19 14	PF00076
CSTF2_HUMAN	CSTF2	1	18-88	NULL	PF00076
EWS_HUMAN	EWSR1	1	363-441	NULL	PF00641 PF00076
HNRPD_HUMAN	HNRNPD	2	100-167, 184-243	46	PF00076 PF08143
HNRPM_HUMAN	HNRNPM	3	74-143, 655-722, 206-274	42 31 42	PF00076 PF11532
RBM40_HUMAN	RNPC3	2	422-496, 29-95	25	PF00076
NUCL_HUMAN	NCL	4	574-640, 309-377, 398-459, 488-554	28 35 38 14 11 43	PF00076
CELF6_HUMAN	CELF6	3	136-198, 398-467, 48-115	28 39 25	PF00076
U2AF1_HUMAN	U2AF1	1	91-141	NULL	PF00642 PF00076
RBM28_HUMAN	RBM28	3	337-395, 116-184, 6-68	37 25 28	PF00076
TRA2A_HUMAN	TRA2A	1	123-191	NULL	PF00076
BOLL_HUMAN	BOLL	1	35-103	NULL	PF00076 PF08777
LA_HUMAN	SSB	1	113-178	NULL	PF08777 PF05383 PF00076
PAP1M_HUMAN	PABPC1L2B PABPC1L2A	2	92-159, 4-74	26	PF00076
HNRH3_HUMAN	HNRNPH3	2	199-262, 20-86	40	PF00076
CELF3_HUMAN	CELF3	3	382-451, 9-76, 97-161	20 30 38	PF00076
PDIP3_HUMAN	POLDIP3	1	284-344	NULL	PF00076
RBM7_HUMAN	RBM7	1	12-81	NULL	PF00076
NOL8_HUMAN	NOL8	1	10-79	NULL	PF00076
PRGC2_HUMAN	PPARGC1B	1	904-955	NULL	PF00076

TIAR_HUMAN	TIAL1	3	99-169, 11-79, 207-271	27 33 35	PF00076
RBM42_HUMAN	RBM42	1	383-453	NULL	PF00076
MINT_HUMAN	SPEN	4	339-406, 440-507, 8-68, 520-583	14 19 20 26 25 19	PF00076 PF07744
PABP2_HUMAN	PABPN1	1	174-243	NULL	PF00076
RBM24_HUMAN	RBM24	1	13-70	NULL	PF00076
ELAV2_HUMAN	ELAVL2	3	41-111, 127-193, 278-347	40 31 26	PF00076
STPAP_HUMAN	TUT1	1	58-121	NULL	PF03828 PF00076
SAFB2_HUMAN	SAFB2	1	409-478	NULL	PF00076 PF02037
RBM4_HUMAN	RBM4	2	4-64, 81-141	42	PF00076 PF00098
RFOX3_HUMAN	RBFOX3	1	102-169	NULL	PF12414 PF00076
RBMX2_HUMAN	RBMX2	1	38-108	NULL	PF00076
RB15B_HUMAN	RBM15B	2	420-480, 339-408	29	PF00076 PF07744
ELAV4_HUMAN	ELAVL4	3	48-118, 299-368, 134-200	31 38 28	PF00076
ELAV1_HUMAN	ELAVL1	3	22-92, 246-314, 108-174	33 29 28	PF00076
RALY_HUMAN	RALY	1	23-85	NULL	PF00076
SRSF3_HUMAN	SRSF3	1	12-77	NULL	PF00076
RBM10_HUMAN	RBM10	1	132-194	NULL	PF01585 PF00076 PF00641
HNRLL_HUMAN	HNRPLL	2	80-128, 180-230	20	PF00076
SF3B4_HUMAN	SF3B4	2	102-173, 15-85	35	PF00076
SLTM_HUMAN	SLTM	1	386-455	NULL	PF00076 PF02037
CSTFT_HUMAN	CSTF2T	1	18-88	NULL	PF00076
RBPS2_HUMAN	RBPMS2	1	33-94	NULL	PF00076
RBPMS_HUMAN	RBPMS	1	26-87	NULL	PF00076
UHMK1_HUMAN	UHMK1	1	345-398	NULL	PF00069 PF07714 PF00076
HNRCL_HUMAN	HNRNPCL1	1	18-79	NULL	PF00076
RMXL3_HUMAN	RBMXL3	1	10-79	NULL	PF00076
NONO_HUMAN	NONO	2	150-210, 76-140	14	PF00076 PF08075
E7EX17_HUMAN	E7EX17	1	99-167	NULL	PF00076
B4E312_HUMAN	B4E312	1	39-117	NULL	PF00076 PF00641
B4DMJ1_HUMAN	B4DMJ1	1	18-80	NULL	PF00076
Q4W5M7_HUMAN	Q4W5M7	1	679-739	NULL	PF00076
Q6IA98_HUMAN	Q6IA98	2	152-217, 249-318	19	PF00076
B4DDC7_HUMAN	B4DDC7	1	131-193	NULL	PF00076
Q5T760_HUMAN	Q5T760	1	37-101	NULL	PF00076
C9JAB2_HUMAN	C9JAB2	1	13-77	NULL	PF00076
Q05BU6_HUMAN	Q05BU6	1	37-102	NULL	PF00076
Q49AS9_HUMAN	Q49AS9	2	2-46, 84-148	26	PF00076
Q5QP71_HUMAN	Q5QP71	1	429-490	NULL	PF00076
B4DQI6_HUMAN	B4DQI6	1	22-90	NULL	PF00076
B7ZLP6_HUMAN	B7ZLP6	1	409-477	NULL	PF00076 PF02037
B4E2U5_HUMAN	B4E2U5	3	385-454, 2-67, 92-157	22 25 34	PF00076
D6RF44_HUMAN	D6RF44	1	2-69	NULL	PF00076
F5H1L1_HUMAN	F5H1L1	1	304-372	NULL	PF00076 PF02037
Q53FG6_HUMAN	Q53FG6	2	15-85, 102-173	33	PF00076
F5H656_HUMAN	F5H656	2	24-88, 98-156	22	PF00076 PF08075

Q5SZ64_HUMAN	Q5SZ64	1	59-130	NULL	PF00076
F5H6R1_HUMAN	F5H6R1	1	38-108	NULL	PF00076
F5H101_HUMAN	F5H101	1	10-79	NULL	PF00076
Q6ZP53_HUMAN	Q6ZP53	1	183-237	NULL	PF00076 PF02136
Q59ES8_HUMAN	Q59ES8	2	540-607, 106-174	42	PF00076
E7EWT5_HUMAN	E7EWT5	1	8-70	NULL	PF00076
A8MY68_HUMAN	A8MY68	1	32-88	NULL	PF00076 PF04098
E9PCZ6_HUMAN	E9PCZ6	1	146-211	NULL	PF00076
B7Z1U7_HUMAN	B7Z1U7	1	163-230	NULL	PF00076
B4DTA2_HUMAN	B4DTA2	2	2-69, 86-147	40	PF00076
Q5JRI3_HUMAN	Q5JRI3	1	12-81	NULL	PF00076
Q53TM1_HUMAN	Q53TM1	1	21-87	NULL	PF00076
B4DWI8_HUMAN	B4DWI8	2	24-88, 98-156	22	PF08075 PF00076
B4DUN1_HUMAN	B4DUN1	1	239-293	NULL	PF00076
B5BUD8_HUMAN	B5BUD8	1	6-67	NULL	PF00076
B4DSS8_HUMAN	B4DSS8	3	194-258, 475-535, 356-421	26 24 16	PF00076 PF00642 PF01480
B3KY61_HUMAN	B3KY61	1	561-612	NULL	PF00076
A4D2F7_HUMAN	A4D2F7	1	18-86	NULL	PF00076
B4DSJ1_HUMAN	B4DSJ1	1	23-85	NULL	PF00076
B4DEH8_HUMAN	B4DEH8	1	46-115	NULL	PF00076
Q5T9T9_HUMAN	Q5T9T9	1	10-80	NULL	PF00076
Q5CZ65_HUMAN	Q5CZ65	1	9-73	NULL	PF00076
B5BU08_HUMAN	B5BU08	1	91-141	NULL	PF00642 PF00076
E9PFS2_HUMAN	E9PFS2	1	47-116	NULL	PF00076
B7ZW38_HUMAN	B7ZW38	1	18-79	NULL	PF00076
Q6AI50_HUMAN	Q6AI50	1	10-79	NULL	PF00076
B4DSU5_HUMAN	B4DSU5	3	191-255, 467-527, 348-413	26 24 16	PF00076
B1ALY6_HUMAN	B1ALY6	3	384-444, 89-153, 265-331	22 18 24	PF00076
Q32P45_HUMAN	Q32P45	1	342-396	NULL	PF02136 PF00076
F2Z2U1_HUMAN	F2Z2U1	1	10-73	NULL	PF00076
E7ENA6_HUMAN	E7ENA6	1	42-98	NULL	PF00076
B2RA86_HUMAN	B2RA86	3	129-193, 426-495, 38-106	30 33 21	PF00076
D6REZ6_HUMAN	D6REZ6	1	73-137	NULL	PF00076
B4DT28_HUMAN	B4DT28	3	28-92, 204-265, 109-170	25 22 22	PF00076
B4DFT9_HUMAN	B4DFT9	1	16-83	NULL	PF00076
B7Z4C2_HUMAN	B7Z4C2	2	2-51, 61-121	6	PF08075 PF00076
D6RBS9_HUMAN	D6RBS9	2	73-137, 153-214	29	PF00076
C9J2Z9_HUMAN	C9J2Z9	1	101-165	NULL	PF00076
E9PC62_HUMAN	E9PC62	3	141-205, 438-507, 50-118	30 33 21	PF00076
Q6P392_HUMAN	Q6P392	1	242-312	NULL	PF00160 PF00076
D6RIT2_HUMAN	D6RIT2	1	15-82	NULL	PF08080 PF00076
Q6NTA2_HUMAN	Q6NTA2	2	89-138, 187-240	26	PF00076
B4DT00_HUMAN	B4DT00	3	110-174, 401-470, 19-87	30 33 21	PF00076
B7Z5E0_HUMAN	B7Z5E0	3	14-84, 251-320, 100-166	31 38 28	PF00076
C9JGD3_HUMAN	C9JGD3	3	216-279, 121-182, 41-107	24 20 22	PF00076

B7Z2F6_HUMAN	B7Z2F6	1	252-320	NULL	PF00076
C9JBI6_HUMAN	C9JBI6	1	679-730	NULL	PF00076
B4DDB6_HUMAN	B4DDB6	2	15-83, 106-174	34	PF00076
B0QYK1_HUMAN	B0QYK1	1	307-385	NULL	PF00641 PF00076
A8K6V7_HUMAN	A8K6V7	1	300-357	NULL	PF02136 PF00076
B1ANR1_HUMAN	B1ANR1	1	13-83	NULL	PF00076
Q2TSD2_HUMAN	Q2TSD2	3	99-169, 11-79, 207-271	27 33 35	PF00076
E5RFU5_HUMAN	E5RFU5	2	79-139, 158-222	27	PF00076
C9IZZ3_HUMAN	C9IZZ3	1	368-437	NULL	PF00076 PF02037
C9JTN7_HUMAN	C9JTN7	2	97-167, 9-77	27	PF00076
A4D2F6_HUMAN	A4D2F6	1	274-342	NULL	PF00076
Q5QP20_HUMAN	Q5QP20	1	27-92	NULL	PF00076
B4DZH7_HUMAN	B4DZH7	1	259-310	NULL	PF00076
E9PQU5_HUMAN	E9PQU5	1	89-157	NULL	PF00076
B7Z5X3_HUMAN	B7Z5X3	1	316-384	NULL	PF00076
A8K8A6_HUMAN	A8K8A6	1	10-80	NULL	PF00076 PF08081
B4DZ74_HUMAN	B4DZ74	1	10-65	NULL	PF00076 PF04098
B3KME7_HUMAN	B3KME7	3	384-444, 89-153, 265-331	22 18 24	PF00076
Q53GL6_HUMAN	Q53GL6	1	23-85	NULL	PF00076
A6NJK7_HUMAN	A6NJK7	1	10-79	NULL	PF00076
E9PHU9_HUMAN	E9PHU9	1	269-338	NULL	PF00076
C9JGE3_HUMAN	C9JGE3	1	290-368	NULL	PF00641 PF00076
B4DK81_HUMAN	B4DK81	1	276-348	NULL	PF00076
B3KP14_HUMAN	B3KP14	3	88-152, 337-399, 186-251	28 35 46	PF00076
B4E3T4_HUMAN	B4E3T4	1	26-87	NULL	PF00076
B4DJ90_HUMAN	B4DJ90	2	284-342, 63-131	37	PF00076
Q6IPF2_HUMAN	Q6IPF2	2	107-175, 16-84	33	PF00076 PF11627
E5RGH4_HUMAN	E5RGH4	1	38-95	NULL	PF00076
F5GZU3_HUMAN	F5GZU3	1	239-307	NULL	PF00076
B7Z7Q9_HUMAN	B7Z7Q9	1	42-98	NULL	PF00076
B4DZW4_HUMAN	B4DZW4	3	251-318, 56-123, 148-216	41 50 47	PF00658 PF00076
B4DRW3_HUMAN	B4DRW3	2	109-157, 22-87	30	PF00076
B4DVF8_HUMAN	B4DVF8	1	71-124	NULL	PF00076
B7Z6E0_HUMAN	B7Z6E0	1	5-73	NULL	PF00076
A6NLN1_HUMAN	A6NLN1	2	13-71, 124-184	27	PF00076
B4DSC7_HUMAN	B4DSC7	2	97-143, 9-76	40	PF00076
E7ETU5_HUMAN	E7ETU5	2	31-91, 110-174	27	PF00076
B3KWE6_HUMAN	B3KWE6	2	7-59, 259-328	28	PF00076
E9PM61_HUMAN	E9PM61	2	4-64, 81-138	46	PF00076
B6RF28_HUMAN	B6RF28	1	149-218	NULL	PF00076
Q9C059_HUMAN	Q9C059	1	37-99	NULL	PF00076
E7EU98_HUMAN	E7EU98	1	42-98	NULL	PF00076
B4DTC3_HUMAN	B4DTC3	2	48-115, 132-191	46	PF00076 PF08143
B7Z4L7_HUMAN	B7Z4L7	2	6-63, 95-165	20	PF00076
A8KAI7_HUMAN	A8KAI7	1	13-70	NULL	PF00076
B4DMB1_HUMAN	B4DMB1	3	129-193, 305-366, 210-271	25 22 22	PF00076

E7EN82_HUMAN	E7EN82	1	42-98	NULL	PF00076
E7ETR3_HUMAN	E7ETR3	1	42-98	NULL	PF00076
E1P603_HUMAN	E1P603	1	13-70	NULL	PF00076
Q13344_HUMAN	Q13344	1	290-368	NULL	PF00641 PF00076
E7EWR4_HUMAN	E7EWR4	1	18-88	NULL	PF00076
E7EQV3_HUMAN	E7EQV3	3	56-123, 251-318, 148-216	41 47 51	PF00658 PF00076
B4DJM2_HUMAN	B4DJM2	1	42-112	NULL	PF00076
F5GYZ3_HUMAN	F5GYZ3	2	2-51, 61-121	6	PF08075 PF00076
B4DMD1_HUMAN	B4DMD1	2	88-149, 183-244	22	PF00076
D6RDL0_HUMAN	D6RDL0	1	14-82	NULL	PF00076
Q53XX5_HUMAN	Q53XX5	1	8-78	NULL	PF00076
B4DHA8_HUMAN	B4DHA8	2	143-205, 56-123	41	PF00076
F5H669_HUMAN	F5H669	1	84-155	NULL	PF00076
B2RDQ3_HUMAN	B2RDQ3	1	122-190	NULL	PF00076
D6RAM1_HUMAN	D6RAM1	1	15-82	NULL	PF00076
F5GXV1_HUMAN	F5GXV1	1	10-57	NULL	PF00076
B4DRS4_HUMAN	B4DRS4	2	289-343, 135-212	5	PF00076
B4DI28_HUMAN	B4DI28	1	152-216	NULL	PF00076
Q32Q83_HUMAN	Q32Q83	1	242-312	NULL	PF00160 PF00076
E9PAU2_HUMAN	E9PAU2	2	78-140, 152-220	25	PF00076
D6RJ04_HUMAN	D6RJ04	1	15-82	NULL	PF00076
Q8TER1_HUMAN	Q8TER1	1	73-136	NULL	PF00076
E7EVG6_HUMAN	E7EVG6	1	1281-1340	NULL	PF00076
E7ETC0_HUMAN	E7ETC0	1	11-79	NULL	PF00076
D3DU92_HUMAN	D3DU92	1	165-234	NULL	PF00076
B4DJP9_HUMAN	B4DJP9	1	12-81	NULL	PF00076
Q96G38_HUMAN	Q96G38	1	16-66	NULL	PF08662 PF00076
Q5HYB4_HUMAN	Q5HYB4	1	62-128	NULL	PF00076
A8K644_HUMAN	A8K644	2	106-171, 4-64	22	PF00076
D6R9T0_HUMAN	D6R9T0	1	15-82	NULL	PF00076
Q9Y6G0_HUMAN	Q9Y6G0	1	38-108	NULL	PF00076
E9PBY2_HUMAN	E9PBY2	1	42-98	NULL	PF00076
B4DN17_HUMAN	B4DN17	1	307-371	NULL	PF01812 PF00076
B4DHS3_HUMAN	B4DHS3	2	2-46, 84-148	24	PF00076
E5RG67_HUMAN	E5RG67	1	9-77	NULL	PF00076
Q59EQ8_HUMAN	Q59EQ8	1	18-87	NULL	PF00076
B4DMY3_HUMAN	B4DMY3	2	64-131, 147-215	39	PF00076 PF08143
E5RFD8_HUMAN	E5RFD8	1	13-60	NULL	PF00076
Q75MU1_HUMAN	Q75MU1	1	45-112	NULL	PF00076
B4DPK8_HUMAN	B4DPK8	2	34-83, 132-185	26	PF00076
B3KXC1_HUMAN	B3KXC1	1	135-196	NULL	PF00076
Q68DG4_HUMAN	Q68DG4	1	115-177	NULL	PF08080 PF00076
F5H532_HUMAN	F5H532	1	10-80	NULL	PF00076
B2RXH8_HUMAN	B2RXH8	1	18-77	NULL	PF00076
B7Z2Z1_HUMAN	B7Z2Z1	1	304-372	NULL	PF00076 PF02037
B4DVB8_HUMAN	B4DVB8	3	49-119, 273-341, 135-201	33 29 28	PF00076

A8K3M9_HUMAN	A8K3M9	2	16-83, 113-174	16	PF00076
B4DHY1_HUMAN	B4DHY1	1	91-154	NULL	PF00076
C9JQN2_HUMAN	C9JQN2	1	50-105	NULL	PF09004 PF00076
Q9BQ02_HUMAN	Q9BQ02	2	346-412, 260-326	38	PF00076
D6RFF0_HUMAN	D6RFF0	1	127-188	NULL	PF05383 PF00076
B4DR70_HUMAN	B4DR70	1	216-294	NULL	PF00641 PF00076
D6RF41_HUMAN	D6RF41	1	63-126	NULL	PF00076
E9PSH0_HUMAN	E9PSH0	1	19-84	NULL	PF00076
C9JPX0_HUMAN	C9JPX0	1	305-374	NULL	PF00076 PF02037
D6RGD8_HUMAN	D6RGD8	1	3-60	NULL	PF00076
C9JJZ7_HUMAN	C9JJZ7	1	19-67	NULL	PF00076
B3KUJ0_HUMAN	B3KUJ0	2	102-173, 15-85	33	PF00076
B4E241_HUMAN	B4E241	1	12-77	NULL	PF00076
Q59G45_HUMAN	Q59G45	2	358-427, 73-137	30	PF00076
F5H6R6_HUMAN	F5H6R6	1	30-90	NULL	PF00076
B7Z6T4_HUMAN	B7Z6T4	1	269-320	NULL	PF00076
F5GWQ7_HUMAN	F5GWQ7	1	39-117	NULL	PF00076 PF00641
F5GXV8_HUMAN	F5GXV8	1	12-81	NULL	PF00076
B4E1E3_HUMAN	B4E1E3	3	226-289, 131-192, 51-117	24 20 22	PF00076 PF08777 PF05383
E9PFH8_HUMAN	E9PFH8	1	113-178	NULL	PF00076
E7EPC9_HUMAN	E7EPC9	1	99-167	NULL	PF00076
B4DHE8_HUMAN	B4DHE8	2	2-69, 90-156	44	PF00076
Q53F64_HUMAN	Q53F64	2	73-140, 156-224	39	PF08143 PF00076
E5RI26_HUMAN	E5RI26	1	13-73	NULL	PF00076
E7ERQ6_HUMAN	E7ERQ6	1	42-98	NULL	PF00076
E7EN40_HUMAN	E7EN40	1	15-82	NULL	PF00076
C9JIJ9_HUMAN	C9JIJ9	2	62-122, 141-205	27	PF00076
B4DM91_HUMAN	B4DM91	1	10-79	NULL	PF00076
D6RFJ8_HUMAN	D6RFJ8	1	275-332	NULL	PF02136 PF00076
B1APY8_HUMAN	B1APY8	3	53-123, 304-373, 139-205	31 38 28	PF00076
A8K5C4_HUMAN	A8K5C4	3	11-79, 99-169, 207-271	26 35 32	PF00076
F5H659_HUMAN	F5H659	1	771-832	NULL	PF00076
Q6ZS99_HUMAN	Q6ZS99	4	467-533, 202-270, 381-447, 291-352	28 38 35 11 14 41	PF00076
O95652_HUMAN	O95652	1	43-87	NULL	PF00076
B2R603_HUMAN	B2R603	1	18-80	NULL	PF00076
F5H357_HUMAN	F5H357	1	26-87	NULL	PF00076
B4DM75_HUMAN	B4DM75	4	76-143, 272-340, 169-237, 2-56	36 32 18 39 18 27	PF00076
Q59FM6_HUMAN	Q59FM6	1	222-291	NULL	PF00076
E9PQ56_HUMAN	E9PQ56	2	113-183, 210-279	41	PF00076
B1ALY5_HUMAN	B1ALY5	3	451-511, 156-220, 332-398	22 18 24	PF00076
B3V096_HUMAN	B3V096	1	18-88	NULL	PF00076
Q7Z3W9_HUMAN	Q7Z3W9	1	61-111	NULL	PF00076
B7ZMC0_HUMAN	B7ZMC0	1	10-79	NULL	PF08081 PF00076
A6NEW5_HUMAN	A6NEW5	1	101-168	NULL	PF00076
C9JLZ0_HUMAN	C9JLZ0	1	495-558	NULL	PF04818 PF00076

B4DZ07_HUMAN	B4DZ07	1	2-49	NULL	PF00076
F5H0M7_HUMAN	F5H0M7	1	145-209	NULL	PF00076
B3KTP9_HUMAN	B3KTP9	4	497-563, 232-300, 321-382, 411-477	28 35 38 14 11 43	PF00076
B1APY9_HUMAN	B1APY9	3	53-123, 290-359, 139-205	31 38 28	PF00076
E7EU39_HUMAN	E7EU39	1	42-98	NULL	PF00076
B2RUT9_HUMAN	B2RUT9	1	10-79	NULL	PF08081 PF00076
C9J286_HUMAN	C9J286	1	84-143	NULL	PF00076
D6RIU0_HUMAN	D6RIU0	2	15-82, 115-168	38	PF00076
Q0VGM7_HUMAN	Q0VGM7	1	38-108	NULL	PF00076
A8K1C9_HUMAN	A8K1C9	2	421-495, 29-95	25	PF00076
D6RFL5_HUMAN	D6RFL5	1	73-136	NULL	PF00076
B4DXN6_HUMAN	B4DXN6	1	54-104	NULL	PF08662 PF00076
E9PFP2_HUMAN	E9PFP2	2	289-343, 135-212	5	PF00076
B9ZVT1_HUMAN	B9ZVT1	2	402-470, 288-353	12	PF00076
Q6PJB9_HUMAN	Q6PJB9	1	37-101	NULL	PF00076
B4DFK9_HUMAN	B4DFK9	2	15-82, 248-310	33	PF08080 PF00076
B4DSI2_HUMAN	B4DSI2	2	131-191, 12-77	16	PF00076
D6RFM3_HUMAN	D6RFM3	2	115-163, 15-82	42	PF00076
A8K583_HUMAN	A8K583	1	679-730	NULL	PF00076
Q5QP21_HUMAN	Q5QP21	1	133-198	NULL	PF00076
B7Z1C7_HUMAN	B7Z1C7	1	208-275	NULL	PF00076
B4DM66_HUMAN	B4DM66	1	231-300	NULL	PF00076
B7Z5B6_HUMAN	B7Z5B6	1	340-408	NULL	PF00076 PF02037
Q701P4_HUMAN	Q701P4	1	91-141	NULL	PF00642 PF00076
B4DSZ2_HUMAN	B4DSZ2	1	19-87	NULL	PF00076
B4E2A3_HUMAN	B4E2A3	1	8-70	NULL	PF00076
B3KSX3_HUMAN	B3KSX3	1	23-85	NULL	PF00076
Q68DD9_HUMAN	Q68DD9	2	138-203, 235-305	19	PF00076
Q5VVE3_HUMAN	Q5VVE3	1	298-365	NULL	PF00076
Q53HH4_HUMAN	Q53HH4	1	342-396	NULL	PF02136 PF00076
B3KVY2_HUMAN	B3KVY2	1	18-86	NULL	PF00076
B0QYK0_HUMAN	B0QYK0	1	325-403	NULL	PF00641 PF00076
B4E187_HUMAN	B4E187	1	258-309	NULL	PF00076
D6R9K7_HUMAN	D6R9K7	2	81-138, 4-64	43	PF00076
B4DU52_HUMAN	B4DU52	1	196-254	NULL	PF00076
B4DRM3_HUMAN	B4DRM3	1	99-167	NULL	PF00076
A8K5K5_HUMAN	A8K5K5	1	241-303	NULL	PF12353 PF00076
B4DDE7_HUMAN	B4DDE7	3	118-182, 413-482, 27-95	30 33 21	PF00076
D6RIH9_HUMAN	D6RIH9	1	15-82	NULL	PF00076
Q59HD3_HUMAN	Q59HD3	1	161-228	NULL	PF00076
C9JE21_HUMAN	C9JE21	1	150-204	NULL	PF00076
Q05DU0_HUMAN	Q05DU0	1	38-108	NULL	PF00076
Q53GD7_HUMAN	Q53GD7	1	12-81	NULL	PF00076
B7ZKM0_HUMAN	B7ZKM0	2	670-738, 767-835	27	PF00076 PF05391
E7EU28_HUMAN	E7EU28	2	145-210, 242-311	19	PF00076

A4D210_HUMAN	A4D210	1	168-218	NULL	PF08662 PF00076
Q59H49_HUMAN	Q59H49	1	203-267	NULL	PF00076
F5H0D8_HUMAN	F5H0D8	1	46-102	NULL	PF00076
Q0VAC0_HUMAN	Q0VAC0	2	107-175, 16-84	33	PF11627 PF00076
Q6IBQ5_HUMAN	Q6IBQ5	1	287-365	NULL	PF00641 PF00076
E7ERC4_HUMAN	E7ERC4	1	113-178	NULL	PF05383 PF00076
D3DPB2_HUMAN	D3DPB2	2	110-174, 31-91	27	PF00076
Q5H919_HUMAN	Q5H919	1	135-212	NULL	PF00076
B4E2S3_HUMAN	B4E2S3	2	62-122, 141-205	27	PF00076
F5GWN9_HUMAN	F5GWN9	1	10-79	NULL	PF00076
C9J9B2_HUMAN	C9J9B2	2	63-123, 142-206	27	PF00076
B7Z213_HUMAN	B7Z213	1	34-95	NULL	PF00076
B7Z974_HUMAN	B7Z974	1	12-76	NULL	PF00076
Q8NAK9_HUMAN	Q8NAK9	1	18-86	NULL	PF00076
Q2VIP3_HUMAN	Q2VIP3	4	101-168, 13-83, 296-363, 193-261	32 38 44 29 28 51	PF00658 PF00076
Q59GY3_HUMAN	Q59GY3	2	38-98, 146-211	24	PF00076
A6NIT8_HUMAN	A6NIT8	1	71-124	NULL	PF00076
F2Z2G3_HUMAN	F2Z2G3	1	35-103	NULL	PF00076
B7WPG3_HUMAN	B7WPG3	1	80-128	NULL	PF00076
F5H718_HUMAN	F5H718	1	101-157	NULL	PF00076
A0PJ47_HUMAN	A0PJ47	1	409-478	NULL	PF00076 PF02037
Q68DZ9_HUMAN	Q68DZ9	1	110-178	NULL	PF00076
B7ZM40_HUMAN	B7ZM40	1	865-916	NULL	PF00076
Q05CK9_HUMAN	Q05CK9	3	200-265, 281-342, 376-437	20 22 19	PF00076
Q6PKC9_HUMAN	Q6PKC9	1	37-101	NULL	PF00076
E5RGV0_HUMAN	E5RGV0	1	63-129	NULL	PF00076
C9J6C5_HUMAN	C9J6C5	1	28-84	NULL	PF00076
B7Z959_HUMAN	B7Z959	1	159-227	NULL	PF00076 PF02037
F5GZT4_HUMAN	F5GZT4	1	23-85	NULL	PF08080 PF00076
Q6IQ42_HUMAN	Q6IQ42	1	12-58	NULL	PF00076
B4DIB6_HUMAN	B4DIB6	2	23-87, 320-389	30	PF00076
A8K4H1_HUMAN	A8K4H1	1	286-364	NULL	PF00641 PF00076
Q9BUQ0_HUMAN	Q9BUQ0	3	365-431, 186-250, 484-544	24 24 24	PF00076
B3KT61_HUMAN	B3KT61	1	23-85	NULL	PF00076
C9J9G0_HUMAN	C9J9G0	1	75-123	NULL	PF00076
B4E3U4_HUMAN	B4E3U4	1	10-74	NULL	PF00076
B5BUB5_HUMAN	B5BUB5	1	113-178	NULL	PF08777 PF05383 PF00076
Q86UM1_HUMAN	Q86UM1	1	207-257	NULL	PF00076
B7Z5Y7_HUMAN	B7Z5Y7	1	58-118	NULL	PF00076
Q5SZQ6_HUMAN	Q5SZQ6	3	382-451, 15-77, 97-161	20 30 33	PF00076
Q5QP23_HUMAN	Q5QP23	1	154-219	NULL	PF00076
D6RDU3_HUMAN	D6RDU3	1	15-82	NULL	PF00076
Q4VC03_HUMAN	Q4VC03	4	13-83, 296-363, 101-168, 193-261	32 29 34 42 50 45	PF00658 PF00076
Q86VW9_HUMAN	Q86VW9	1	366-434	NULL	PF07744 PF00076
B4DEP6_HUMAN	B4DEP6	1	99-167	NULL	PF00076

A8K5X9_HUMAN	A8K5X9	5	834-904, 4-72, 732-804, 404-473, 302-362	30 25 24 22 31 28 13 34 19 14	PF00076
B0LM41_HUMAN	B0LM41	1	3-60	NULL	PF00076 PF00098
E7EV40_HUMAN	E7EV40	1	80-146	NULL	PF00076
B4DJ45_HUMAN	B4DJ45	1	77-125	NULL	PF00076
Q6ZR81_HUMAN	Q6ZR81	1	8-70	NULL	PF00076
F5GY08_HUMAN	F5GY08	1	12-81	NULL	PF00076
Q569H2_HUMAN	Q569H2	1	26-87	NULL	PF00076
E7EQD1_HUMAN	E7EQD1	1	86-154	NULL	PF00076
B2RDP1_HUMAN	B2RDP1	2	152-217, 249-318	19	PF00076
E7EX81_HUMAN	E7EX81	4	466-532, 201-269, 290-351, 380-446	28 35 38 14 11 43	PF00076
C9JAA9_HUMAN	C9JAA9	1	6-68	NULL	PF00076
B3KWX7_HUMAN	B3KWX7	2	131-196, 228-298	19	PF00076
Q65ZQ3_HUMAN	Q65ZQ3	2	128-196, 37-105	30	PF00076
B5BU25_HUMAN	B5BU25	3	396-455, 151-224, 261-330	11 18 20	PF00076
E7EQJ0_HUMAN	E7EQJ0	1	15-82	NULL	PF00076
E7ET38_HUMAN	E7ET38	1	130-187	NULL	PF00076
B5BTZ8_HUMAN	B5BTZ8	2	153-212, 9-79	25	PF00076
B7Z8K4_HUMAN	B7Z8K4	1	160-214	NULL	PF00076
B9EIP3_HUMAN	B9EIP3	1	10-79	NULL	PF08081 PF00076
A8K4T9_HUMAN	A8K4T9	1	23-85	NULL	PF00076
Q59G98_HUMAN	Q59G98	3	146-216, 47-115, 295-359	27 33 36	PF00076
A8KAQ5_HUMAN	A8KAQ5	1	105-174	NULL	PF12220 PF00076
F5H160_HUMAN	F5H160	1	779-840	NULL	PF00076
B4DLU6_HUMAN	B4DLU6	1	13-77	NULL	PF00076
A8K329_HUMAN	A8K329	1	409-477	NULL	PF00076 PF02037
B4DEM8_HUMAN	B4DEM8	1	4-64	NULL	PF00076
B3KQH4_HUMAN	B3KQH4	3	68-135, 495-564, 184-249	22 22 22	PF00076
B1AM48_HUMAN	B1AM48	2	41-111, 127-189	42	PF00076
B4DWT1_HUMAN	B4DWT1	1	37-102	NULL	PF00076
Q2VIN3_HUMAN	Q2VIN3	1	10-80	NULL	PF08081 PF00076
B7Z2D8_HUMAN	B7Z2D8	3	38-102, 287-349, 136-201	28 35 46	PF00076
A2RRD3_HUMAN	A2RRD3	2	133-198, 230-300	19	PF00076
E7ERJ4_HUMAN	E7ERJ4	2	15-83, 106-174	34	PF00076
Q59G49_HUMAN	Q59G49	2	4-48, 86-150	26	PF00076
B1AM49_HUMAN	B1AM49	3	69-139, 155-221, 306-375	40 31 26	PF00076
E7EPM3_HUMAN	E7EPM3	1	752-813	NULL	PF00076
B4DKS8_HUMAN	B4DKS8	2	38-104, 216-279	42	PF08080 PF00076
B7Z4G7_HUMAN	B7Z4G7	3	51-121, 288-357, 137-203	31 38 28	PF00076
Q53FN0_HUMAN	Q53FN0	1	18-86	NULL	PF00076
Q6MZY7_HUMAN	Q6MZY7	2	6-63, 95-165	20	PF00076
D6R9P3_HUMAN	D6R9P3	2	73-140, 156-224	39	PF00076 PF08143
D6RBM0_HUMAN	D6RBM0	2	15-82, 115-181	44	PF00076
B4DZ01_HUMAN	B4DZ01	1	197-266	NULL	PF00076
A2ABK1_HUMAN	A2ABK1	1	264-318	NULL	PF00076
C9J1W7_HUMAN	C9J1W7	1	510-573	NULL	PF04818 PF00076

A8K525_HUMAN	A8K525	2	150-210, 76-140	14	PF08075 PF00076
Q6FGE0_HUMAN	Q6FGE0	1	6-67	NULL	PF00076
C9JYS8_HUMAN	C9JYS8	1	76-140	NULL	PF00076 PF08075
E5RHG7_HUMAN	E5RHG7	1	13-73	NULL	PF00076
B4E3E6_HUMAN	B4E3E6	2	106-174, 15-83	33	PF00076
Q5H918_HUMAN	Q5H918	1	135-212	NULL	PF00076
Q2L7G6_HUMAN	Q2L7G6	3	129-193, 305-366, 210-271	25 22 22	PF00076
C9JZG1_HUMAN	C9JZG1	1	168-218	NULL	PF00076
Q59GZ7_HUMAN	Q59GZ7	1	4-61	NULL	PF00658 PF00076
Q6PJY9_HUMAN	Q6PJY9	1	37-102	NULL	PF00076
Q7Z3L0_HUMAN	Q7Z3L0	2	6-63, 95-165	20	PF00076
B4E0L0_HUMAN	B4E0L0	1	301-361	NULL	PF00076
B4DG28_HUMAN	B4DG28	2	21-83, 261-330	28	PF00076
B4DJB6_HUMAN	B4DJB6	3	136-198, 371-440, 48-115	28 39 25	PF00076
E5RFV3_HUMAN	E5RFV3	1	23-86	NULL	PF00076
Q14730_HUMAN	Q14730	1	1-66	NULL	PF08777 PF00076
A4QPE1_HUMAN	A4QPE1	1	158-214	NULL	PF00076
Q8N220_HUMAN	Q8N220	1	18-78	NULL	PF00076
C9IZL7_HUMAN	C9IZL7	2	76-140, 150-206	15	PF00076
B7Z6I4_HUMAN	B7Z6I4	1	130-187	NULL	PF00076
Q59H57_HUMAN	Q59H57	1	61-139	NULL	PF00641 PF00076
E9PCL7_HUMAN	E9PCL7	1	264-318	NULL	PF00076
E9PEG6_HUMAN	E9PEG6	1	2-48	NULL	PF00076
A8MXT5_HUMAN	A8MXT5	1	409-478	NULL	PF00076 PF02037
Q3S611_HUMAN	Q3S611	1	8-78	NULL	PF00160 PF00076
Q0P607_HUMAN	Q0P607	1	510-573	NULL	PF04818 PF00076
B3KUF7_HUMAN	B3KUF7	1	18-73	NULL	PF00076
A8K4L9_HUMAN	A8K4L9	3	11-96, 116-186, 224-288	26 35 33	PF00076
F5GYA8_HUMAN	F5GYA8	1	2-49	NULL	PF00076
Q71RF1_HUMAN	Q71RF1	1	18-68	NULL	PF00076 PF00642
B4E2X2_HUMAN	B4E2X2	1	8-78	NULL	PF00076
Q86X94_HUMAN	Q86X94	1	236-314	NULL	PF00076 PF00641
E9PAM1_HUMAN	E9PAM1	1	308-372	NULL	PF01812 PF00076
F2Z2W2_HUMAN	F2Z2W2	1	3-60	NULL	PF00076
E7EQS3_HUMAN	E7EQS3	2	4-64, 81-139	44	PF00076
E9PDD9_HUMAN	E9PDD9	1	196-254	NULL	PF00076
Q59EF5_HUMAN	Q59EF5	1	31-96	NULL	PF00076
Q6MZS5_HUMAN	Q6MZS5	3	144-208, 225-289, 323-384	20 25 22	PF00076
Q53F48_HUMAN	Q53F48	2	20-86, 199-263	41	PF00076
F5H047_HUMAN	F5H047	1	84-155	NULL	PF00076
B4E1U0_HUMAN	B4E1U0	2	81-139, 4-64	42	PF00076
A8K588_HUMAN	A8K588	2	4-64, 112-177	24	PF00076
B2R802_HUMAN	B2R802	2	210-272, 12-82	14	PF00076
F5H0R1_HUMAN	F5H0R1	1	96-159	NULL	PF03828 PF00076
Q8IWE6_HUMAN	Q8IWE6	1	37-101	NULL	PF00076
B4DZ27_HUMAN	B4DZ27	3	63-126, 143-204, 238-301	25 14 22	PF00076

B4DRP9_HUMAN	B4DRP9	3	231-298, 658-727, 347-412	22 22 22	PF00076
D6W592_HUMAN	D6W592	1	80-128	NULL	PF00076
E7EPF2_HUMAN	E7EPF2	1	31-91	NULL	PF00076
E9PGX9_HUMAN	E9PGX9	1	62-118	NULL	PF00076
E7EMJ8_HUMAN	E7EMJ8	2	106-171, 4-64	22	PF00076
Q5VZZ6_HUMAN	Q5VZZ6	3	110-174, 415-484, 19-87	30 33 21	PF00076
D6W5H1_HUMAN	D6W5H1	1	679-730	NULL	PF00076
B3KUY1_HUMAN	B3KUY1	1	6-74	NULL	PF00076
B2R6F3_HUMAN	B2R6F3	1	12-77	NULL	PF00076
SRSF8_HUMAN	Q9BRL6	1	18-86	NULL	PF00076
E9PLB0_HUMAN	E9PLB0	2	81-141, 4-64	44	PF00076
B3KM80_HUMAN	B3KM80	4	400-466, 135-203, 224-285, 314-380	28 35 38 14 11 43	PF00076
D3DQM9_HUMAN	D3DQM9	1	486-547	NULL	PF00076
E5RG71_HUMAN	E5RG71	1	23-84	NULL	PF00076
B7Z3R7_HUMAN	B7Z3R7	3	124-188, 373-435, 222-287	28 35 46	PF00076
E5RGH3_HUMAN	E5RGH3	1	56-108	NULL	PF00076
Q3B867_HUMAN	Q3B867	1	46-113	NULL	PF00076
Q6IQ30_HUMAN	Q6IQ30	4	13-83, 296-363, 101-168, 193-261	32 29 34 42 50 45	PF00658 PF00076
D6R9M7_HUMAN	D6R9M7	1	73-137	NULL	PF00076
E7EUX0_HUMAN	E7EUX0	1	285-363	NULL	PF00641 PF00076
B7ZLC2_HUMAN	B7ZLC2	1	308-372	NULL	PF01812 PF00076
Q5QPM2_HUMAN	Q5QPM2	1	23-85	NULL	PF00076
B4DEG4_HUMAN	B4DEG4	2	74-143, 167-235	31	PF11532 PF00076
B4DQL3_HUMAN	B4DQL3	1	177-246	NULL	PF00076
Q1RMF9_HUMAN	Q1RMF9	3	207-263, 372-428, 42-98	100 100 100	PF00076
B4DJK0_HUMAN	B4DJK0	1	6-67	NULL	PF00076
B3KRR4_HUMAN	B3KRR4	1	242-312	NULL	PF00160 PF00076
D3DQF6_HUMAN	D3DQF6	1	304-355	NULL	PF00076
B2R7W4_HUMAN	B2R7W4	3	167-231, 343-404, 248-309	25 22 22	PF00076
B7ZLP7_HUMAN	B7ZLP7	3	73-137, 153-214, 248-311	29 17 19	PF00076
E9PN18_HUMAN	E9PN18	1	151-221	NULL	PF00076
C9JZB7_HUMAN	C9JZB7	1	110-177	NULL	PF00076
Q2PYN1_HUMAN	Q2PYN1	1	3-60	NULL	PF00076
C9JT33_HUMAN	C9JT33	1	4-68	NULL	PF00013 PF00076
E9PB61_HUMAN	E9PB61	1	115-183	NULL	PF00076
E9PCY7_HUMAN	E9PCY7	3	15-82, 293-355, 115-181	31 44 41	PF08080 PF00076
B4DMV6_HUMAN	B4DMV6	1	45-112	NULL	PF00076
D3DSV0_HUMAN	D3DSV0	1	26-87	NULL	PF00076
F5H606_HUMAN	F5H606	1	12-81	NULL	PF00076
A8K9U0_HUMAN	A8K9U0	2	81-141, 4-64	42	PF00076 PF00098
Q96DI9_HUMAN	Q96DI9	1	280-340	NULL	PF00076
B0QYV1_HUMAN	B0QYV1	1	95-162	NULL	PF00076
F5GXS4_HUMAN	F5GXS4	1	160-229	NULL	PF00076 PF02037
B4DSM4_HUMAN	B4DSM4	1	47-116	NULL	PF00076
B2R959_HUMAN	B2R959	2	75-124, 173-226	26	PF00076

E9PMU7_HUMAN	E9PMU7	1	151-221	NULL	PF00076
D6R9D6_HUMAN	D6R9D6	3	73-137, 153-214, 248-311	29 17 19	PF00076
B1ANR7_HUMAN	B1ANR7	1	75-142	NULL	PF00658 PF00076
B7Z5F9_HUMAN	B7Z5F9	3	69-133, 318-380, 167-232	28 35 46	PF00076
E7EWI9_HUMAN	E7EWI9	2	15-83, 106-174	34	PF00076
B4DMM2_HUMAN	B4DMM2	1	128-185	NULL	PF00076
D6RD18_HUMAN	D6RD18	2	73-140, 156-224	39	PF00076 PF08143
Q0VGD6_HUMAN	Q0VGD6	3	141-205, 317-378, 222-283	25 22 22	PF00076
B4DUD5_HUMAN	B4DUD5	1	18-88	NULL	PF00076
B7Z888_HUMAN	B7Z888	1	545-610	NULL	PF04818 PF00076
F5H0I5_HUMAN	F5H0I5	1	10-80	NULL	PF00076
Q96MX4_HUMAN	Q96MX4	1	368-446	NULL	PF00641 PF00076
Q5JRI1_HUMAN	Q5JRI1	1	12-81	NULL	PF00076
B3KW50_HUMAN	B3KW50	3	135-199, 384-446, 233-298	28 35 46	PF00076
B4DGF8_HUMAN	B4DGF8	1	84-155	NULL	PF00076
Q6PIX2_HUMAN	Q6PIX2	2	374-433, 299-363	16	PF08075 PF00076
C9JB16_HUMAN	C9JB16	1	1-58	NULL	PF00076
Q5QPM1_HUMAN	Q5QPM1	1	23-85	NULL	PF00076
F5H330_HUMAN	F5H330	1	785-836	NULL	PF00076
Q75MU2_HUMAN	Q75MU2	1	45-112	NULL	PF00076
A8K894_HUMAN	A8K894	2	75-123, 175-225	20	PF00076
B4E1M7_HUMAN	B4E1M7	2	146-211, 243-313	19	PF00076
E5RJM0_HUMAN	E5RJM0	1	449-510	NULL	PF00076
Q6FI03_HUMAN	Q6FI03	1	342-396	NULL	PF02136 PF00076
B7ZMD9_HUMAN	B7ZMD9	1	10-79	NULL	PF08081 PF00076
Q6PKH5_HUMAN	Q6PKH5	1	4-66	NULL	PF00076 PF00641
B7Z645_HUMAN	B7Z645	3	66-131, 147-208, 242-303	20 22 19	PF00076
Q96MN4_HUMAN	Q96MN4	1	307-385	NULL	PF00641 PF00076
B4DRA0_HUMAN	B4DRA0	2	133-198, 230-300	19	PF00076
Q5VX58_HUMAN	Q5VX58	4	101-168, 13-83, 296-363, 193-261	32 38 44 29 28 51	PF00658 PF00076
B2R5W2_HUMAN	B2R5W2	1	18-80	NULL	PF00076
B4DYX9_HUMAN	B4DYX9	1	276-330	NULL	PF00076
Q9BSM5_HUMAN	Q9BSM5	2	2-62, 85-153	32	PF00076 PF11627
A8K1L8_HUMAN	A8K1L8	2	18-85, 123-184	25	PF00076
Q5VVE2_HUMAN	Q5VVE2	2	78-145, 179-246	14	PF00076
B4DF29_HUMAN	B4DF29	1	102-169	NULL	PF00076
Q7KYM9_HUMAN	Q7KYM9	2	495-562, 46-114	42	PF00076
Q5TGA3_HUMAN	Q5TGA3	1	8-78	NULL	PF00160 PF00076
E7ETM7_HUMAN	E7ETM7	2	183-244, 88-149	22	PF00076
F5H4Y5_HUMAN	F5H4Y5	1	19-79	NULL	PF00076
E5RGV5_HUMAN	E5RGV5	1	9-77	NULL	PF00076
E7ENA5_HUMAN	E7ENA5	1	42-98	NULL	PF00076
F5H5I6_HUMAN	F5H5I6	3	38-102, 287-349, 136-201	28 35 46	PF00076
C9IYN3_HUMAN	C9IYN3	1	80-128	NULL	PF00076
B3KPW0_HUMAN	B3KPW0	2	243-305, 92-157	46	PF00076

Q59EK7_HUMAN	Q59EK7	2	176-229, 9-70	29	PF00076
B4DN89_HUMAN	B4DN89	1	18-86	NULL	PF00076
B0QYY4_HUMAN	B0QYY4	1	94-161	NULL	PF00076
B3KT93_HUMAN	B3KT93	4	101-168, 193-261, 296-363, 13-83	47 39 33 50 30 36	PF00658 PF00076
A4QPG7_HUMAN	A4QPG7	2	93-153, 12-81	29	PF07744 PF00076
B4DF54_HUMAN	B4DF54	1	785-836	NULL	PF00076
C9JGC2_HUMAN	C9JGC2	1	83-152	NULL	PF00076
B7Z8Z7_HUMAN	B7Z8Z7	3	35-99, 115-176, 210-273	29 17 19	PF00076
A1A693_HUMAN	A1A693	1	332-400	NULL	PF07744 PF00076
Q6IRX3_HUMAN	Q6IRX3	1	12-81	NULL	PF00076
Q59FA2_HUMAN	Q59FA2	2	50-117, 155-217	25	PF00076
F5H540_HUMAN	F5H540	1	10-80	NULL	PF00076
A0AV56_HUMAN	A0AV56	1	409-477	NULL	PF00076 PF02037
B8ZZ74_HUMAN	B8ZZ74	1	3-60	NULL	PF00076
B4DLM0_HUMAN	B4DLM0	2	128-193, 225-295	19	PF00076
Q3ZCS4_HUMAN	Q3ZCS4	4	101-168, 13-83, 296-363, 193-261	32 38 44 29 28 51	PF00658 PF00076
Q32ND0_HUMAN	Q32ND0	1	101-157	NULL	PF00076
O15187_HUMAN	O15187	2	60-130, 168-232	33	PF00076
B4DUA4_HUMAN	B4DUA4	1	6-67	NULL	PF00076
F5GWK3_HUMAN	F5GWK3	1	185-249	NULL	PF00076
Q6IAM0_HUMAN	Q6IAM0	1	241-303	NULL	PF12353 PF00076
D6RA49_HUMAN	D6RA49	1	73-123	NULL	PF00076
Q53F45_HUMAN	Q53F45	2	106-171, 4-64	22	PF00076
Q8NB80_HUMAN	Q8NB80	1	13-77	NULL	PF00076 PF00098
Q5CZ66_HUMAN	Q5CZ66	1	9-73	NULL	PF00076
Q5SZQ7_HUMAN	Q5SZQ7	3	332-401, 9-76, 97-161	20 30 38	PF00076
Q8TCM0_HUMAN	Q8TCM0	1	120-187	NULL	PF00076
B4DSU6_HUMAN	B4DSU6	1	18-80	NULL	PF00076
Q6PKD2_HUMAN	Q6PKD2	1	18-79	NULL	PF00076
B4DP51_HUMAN	B4DP51	1	23-85	NULL	PF08080 PF00076
B3KUB0_HUMAN	B3KUB0	1	145-209	NULL	PF00076
E5RFP2_HUMAN	E5RFP2	1	84-145	NULL	PF00076
Q5T0W7_HUMAN	Q5T0W7	2	138-199, 58-124	22	PF00076
Q5JB52_HUMAN	Q5JB52	1	80-128	NULL	PF00076
B7Z3A4_HUMAN	B7Z3A4	1	557-622	NULL	PF04818 PF00076
B4DM51_HUMAN	B4DM51	1	51-117	NULL	PF00076
Q147Y3_HUMAN	Q147Y3	1	46-113	NULL	PF00076
PAB4L_HUMAN	P0CB38	4	12-82, 100-167, 192-259, 295-362	29 29 26 47 39 47	PF00076
Q7Z780_HUMAN	Q7Z780	1	19-68	NULL	PF00642 PF00076
B3KX96_HUMAN	B3KX96	1	18-80	NULL	PF00076
Q8N1H4_HUMAN	Q8N1H4	1	86-154	NULL	PF00076
B7ZLC0_HUMAN	B7ZLC0	1	307-371	NULL	PF01812 PF00076
Q86VG2_HUMAN	Q86VG2	2	374-433, 299-363	16	PF08075 PF00076
Q7Z2U9_HUMAN	Q7Z2U9	1	58-121	NULL	PF00076
Q8TBR3_HUMAN	Q8TBR3	1	287-365	NULL	PF00641 PF00076

E1P5S2_HUMAN	E1P5S2	2	6-63, 95-165	20	PF00076
Q8TAL0_HUMAN	Q8TAL0	1	590-641	NULL	PF00076
B4DEK2_HUMAN	B4DEK2	1	13-77	NULL	PF00076
E9PQK4_HUMAN	E9PQK4	1	19-84	NULL	PF00076
Q6N037_HUMAN	Q6N037	2	6-63, 95-165	20	PF00076
A8K9S4_HUMAN	A8K9S4	2	63-123, 142-206	27	PF00076
D6W5Y5_HUMAN	D6W5Y5	1	8-78	NULL	PF00076
F5H3W3_HUMAN	F5H3W3	1	42-98	NULL	PF00076
Q05DR1_HUMAN	Q05DR1	1	12-82	NULL	PF00076 PF00098
Q53GL4_HUMAN	Q53GL4	1	7-74	NULL	PF00658 PF00076
B3KWU8_HUMAN	B3KWU8	3	63-126, 143-204, 238-301	25 14 22	PF00076
F5H4D6_HUMAN	F5H4D6	1	160-214	NULL	PF00076
F5H7Z1_HUMAN	F5H7Z1	1	212-279	NULL	PF00076
B7Z570_HUMAN	B7Z570	1	18-79	NULL	PF00076
Q9BSV4_HUMAN	Q9BSV4	2	301-360, 226-290	16	PF08075 PF00076
Q86YK2_HUMAN	Q86YK2	1	9-79	NULL	PF00076
C9J323_HUMAN	C9J323	1	84-143	NULL	PF00076
Q549U1_HUMAN	Q549U1	1	123-191	NULL	PF00076
Q3ZB86_HUMAN	Q3ZB86	1	376-444	NULL	PF07744 PF00076
Q4VY17_HUMAN	Q4VY17	3	101-168, 193-261, 13-83	50 29 33	PF00076
D6RBZ0_HUMAN	D6RBZ0	2	73-140, 156-224	39	PF00076 PF08143
Q8N8Y7_HUMAN	Q8N8Y7	1	10-67	NULL	PF08081 PF00076
B4DSS0_HUMAN	B4DSS0	1	1281-1340	NULL	PF00076
E7EW00_HUMAN	E7EW00	1	276-348	NULL	PF00076
E9PL19_HUMAN	E9PL19	1	168-228	NULL	PF00076
Q59GA1_HUMAN	Q59GA1	1	110-178	NULL	PF00076
B4DFI3_HUMAN	B4DFI3	1	22-84	NULL	PF00076
E7ERJ7_HUMAN	E7ERJ7	3	264-331, 161-229, 13-83	51 36 31	PF00658 PF00076
F5GYB8_HUMAN	F5GYB8	1	18-78	NULL	PF00076
F5H6M0_HUMAN	F5H6M0	1	84-155	NULL	PF00076
B2R8Z8_HUMAN	B2R8Z8	3	164-229, 245-306, 340-401	20 22 19	PF00076
E9PKU1_HUMAN	E9PKU1	1	46-111	NULL	PF00076
B7ZLH9_HUMAN	B7ZLH9	1	140-207	NULL	PF00076
B3KRJ9_HUMAN	B3KRJ9	2	23-85, 184-252	23	PF00076
Q5IRN2_HUMAN	Q5IRN2	2	115-159, 12-80	35	PF00076
B1AKF7_HUMAN	B1AKF7	1	101-157	NULL	PF00076
B4DV79_HUMAN	B4DV79	1	131-181	NULL	PF08662 PF00076
A8KAM9_HUMAN	A8KAM9	1	8-78	NULL	PF00160 PF00076
D6RAF8_HUMAN	D6RAF8	1	100-167	NULL	PF00076 PF08143
F5H8B8_HUMAN	F5H8B8	1	10-80	NULL	PF00076
Q96J71_HUMAN	Q96J71	3	78-141, 1-62, 237-306	33 28 30	PF00076
B3KVW0_HUMAN	B3KVW0	1	840-891	NULL	PF00076
B3KWQ8_HUMAN	B3KWQ8	2	265-310, 168-233	23	PF00076
Q69YJ7_HUMAN	Q69YJ7	3	454-520, 881-950, 570-635	22 22 22	PF00076
B3KQ99_HUMAN	B3KQ99	1	130-187	NULL	PF00076
E9PB51_HUMAN	E9PB51	2	4-64, 81-141	42	PF00076

F5H0H3_HUMAN	F5H0H3	1	340-408	NULL	PF00076 PF02037
A6NNE8_HUMAN	A6NNE8	1	13-77	NULL	PF00076
C9JD39_HUMAN	C9JD39	1	510-573	NULL	PF04818 PF00076
E5RH24_HUMAN	E5RH24	1	13-83	NULL	PF00076
D3DPI2_HUMAN	D3DPI2	1	18-80	NULL	PF00076
E9PK21_HUMAN	E9PK21	1	8-78	NULL	PF00076
Q6P2D7_HUMAN	Q6P2D7	1	210-277	NULL	PF00076
Q69YM5_HUMAN	Q69YM5	1	68-136	NULL	PF00076
B0QYY7_HUMAN	B0QYY7	1	137-200	NULL	PF00076
B5BU15_HUMAN	B5BU15	1	4-64	NULL	PF00076
D6RBQ9_HUMAN	D6RBQ9	1	81-148	NULL	PF00076 PF08143
B4DY08_HUMAN	B4DY08	1	18-80	NULL	PF00076
E2PSN0_HUMAN	E2PSN0	1	88-139	NULL	PF00076
B4DUA9_HUMAN	B4DUA9	1	22-90	NULL	PF00076
B1AKP7_HUMAN	B1AKP7	2	193-241, 106-171	28	PF00076
Q9BTF3_HUMAN	Q9BTF3	2	88-153, 6-68	34	PF00076
Q5JPA7_HUMAN	Q5JPA7	1	317-370	NULL	PF00069 PF00076
B4DP35_HUMAN	B4DP35	1	16-84	NULL	PF00076 PF11627
B4DN88_HUMAN	B4DN88	2	64-124, 143-207	27	PF00076
B4DYA2_HUMAN	B4DYA2	2	285-339, 367-435	25	PF00076 PF05391
Q5QPL9_HUMAN	Q5QPL9	1	23-85	NULL	PF00076
B4E0W4_HUMAN	B4E0W4	1	30-90	NULL	PF00076
B7Z8M4_HUMAN	B7Z8M4	2	6-63, 95-164	20	PF00076
Q7Z3D7_HUMAN	Q7Z3D7	1	197-259	NULL	PF01585 PF00076 PF00641
Q5U0Q1_HUMAN	Q5U0Q1	1	342-396	NULL	PF02136 PF00076
A8K9A4_HUMAN	A8K9A4	1	18-80	NULL	PF00076
E7EU30_HUMAN	E7EU30	1	42-98	NULL	PF00076
B4E0B5_HUMAN	B4E0B5	1	16-84	NULL	PF00076 PF11627
E7ET15_HUMAN	E7ET15	1	275-347	NULL	PF01805 PF00076 PF08312
B4DS31_HUMAN	B4DS31	3	134-198, 431-500, 43-111	30 33 21	PF00076
B4DQX0_HUMAN	B4DQX0	3	264-331, 161-229, 13-83	51 36 31	PF00658 PF00076
E7EU33_HUMAN	E7EU33	1	42-98	NULL	PF00076
Q96B58_HUMAN	Q96B58	2	108-178, 9-77	27	PF00076
E9PID8_HUMAN	E9PID8	1	18-88	NULL	PF00076
Q15164_HUMAN	Q15164	2	107-174, 4-72	50	PF00076
B4DTC1_HUMAN	B4DTC1	1	37-102	NULL	PF00076
E7ERE4_HUMAN	E7ERE4	3	167-231, 347-408, 248-313	25 21 22	PF00076
B7ZLP5_HUMAN	B7ZLP5	1	409-477	NULL	PF00076 PF02037
Q9Y655_HUMAN	Q9Y655	3	182-250, 49-118, 472-539	24 42 39	PF00076
B1ANR0_HUMAN	B1ANR0	4	13-83, 296-363, 101-168, 193-261	32 29 34 42 50 45	PF00658 PF00076
B4DRG9_HUMAN	B4DRG9	1	592-639	NULL	PF00076
B7ZLQ8_HUMAN	B7ZLQ8	1	449-510	NULL	PF00076
E7ETJ9_HUMAN	E7ETJ9	1	60-130	NULL	PF00076
E7EUF4_HUMAN	E7EUF4	1	42-98	NULL	PF00076
B4DP00_HUMAN	B4DP00	2	199-261, 15-74	28	PF08080 PF00076

B7ZW41_HUMAN	B7ZW41	1	18-77	NULL	PF00076
B7Z876_HUMAN	B7Z876	1	524-589	NULL	PF04818 PF00076
Q6NXQ0_HUMAN	Q6NXQ0	1	18-78	NULL	PF00076
F5H0B8_HUMAN	F5H0B8	2	286-340, 368-436	25	PF00076 PF05391
Q5JQF3_HUMAN	Q5JQF3	2	140-208, 37-105	39	PF00076
A4D198_HUMAN	A4D198	1	111-178	NULL	PF00076
E1CJT3_HUMAN	E1CJT3	1	40-101	NULL	PF00076
Q86VN0_HUMAN	Q86VN0	1	241-297	NULL	PF00642 PF00076
A2A2V2_HUMAN	A2A2V2	1	267-336	NULL	PF00076
E5RJB9_HUMAN	E5RJB9	1	13-83	NULL	PF00076
Q12771_HUMAN	Q12771	2	79-146, 163-223	45	PF08143 PF00076
Q59GL1_HUMAN	Q59GL1	3	136-201, 217-278, 312-373	20 22 19	PF00076
Q96FE8_HUMAN	Q96FE8	1	362-440	NULL	PF00641 PF00076
B4E0E5_HUMAN	B4E0E5	1	9-77	NULL	PF00076

Table S 2.2 RRM-containing RBPs unique to four species and their multiRRM scores

HumUnique	Hum RRM1	Hum RRM2	Hum_Scores
ENSP00000352440	33-80	117-184	25
ENSP00000383344	42-98	207-263	100
ENSP00000341285	16-84	107-175	34
ENSP00000383298	16-84	107-175	34
ENSP00000362618	4-74	92-159	26
ENSP00000281589	13-83	101-168	32
ENSP00000281589	13-83	193-261	28
ENSP00000281589	13-83	296-363	27
ENSP00000281589	101-168	193-261	44
ENSP00000281589	101-168	296-363	38
ENSP00000281589	193-261	296-363	52
ENSP00000352956	168-233	265-334	19
ENSP00000302745	52-108	217-273	100
ENSP00000302745	52-108	382-438	100
ENSP00000302745	217-273	382-438	100
ENSP00000362621	4-74	92-159	26

MouUnique	Mou RRM1	Mou RRM2	Mou_Scores
ENSMUSP00000066639	13-83	101-168	32
ENSMUSP00000066639	13-83	193-261	31
ENSMUSP00000066639	13-83	296-363	32
ENSMUSP00000066639	101-168	193-261	47
ENSMUSP00000066639	101-168	296-363	39
ENSMUSP00000066639	193-261	296-363	57
ENSMUSP00000111809	10-78	101-169	33
ENSMUSP00000072775	23-92	114-182	34
ENSMUSP00000071646	37-106	128-196	36
ENSMUSP00000098682	2-69	94-162	41
ENSMUSP00000098682	2-69	197-264	42
ENSMUSP00000098682	94-162	197-264	47
ENSMUSP00000050792	13-83	101-168	35
ENSMUSP00000050792	13-83	193-261	33
ENSMUSP00000050792	13-83	306-373	33
ENSMUSP00000050792	101-168	193-261	48
ENSMUSP00000050792	101-168	306-373	44
ENSMUSP00000050792	193-261	306-373	55
ENSMUSP00000006628	4-64	81-141	44
ENSMUSP00000058811	93-162	226-294	28
ENSMUSP00000058811	93-162	459-484	26
ENSMUSP00000058811	226-294	459-484	53
ENSMUSP00000093293	12-61	202-262	10
ENSMUSP00000079967	13-83	101-168	29
ENSMUSP00000079967	13-83	193-261	34
ENSMUSP00000079967	13-83	296-363	30
ENSMUSP00000079967	101-168	193-261	44
ENSMUSP00000079967	101-168	296-363	42
ENSMUSP00000079967	193-261	296-363	48
ENSMUSP00000053555	5-69	156-223	30
ENSMUSP00000053555	5-69	285-351	16
ENSMUSP00000053555	5-69	403-458	23
ENSMUSP00000053555	5-69	763-831	24
ENSMUSP00000053555	156-223	285-351	23
ENSMUSP00000053555	156-223	403-458	28
ENSMUSP00000053555	156-223	763-831	25
ENSMUSP00000053555	285-351	403-458	19
ENSMUSP00000053555	285-351	763-831	8
ENSMUSP00000053555	403-458	763-831	30
ENSMUSP00000049830	36-103	120-178	30

FlyUnique	Fly RRM1 vs RRM2		Fly_Scores
FBpp0084255	205-273	293-361	42
FBpp0112339	32-99	120-186	46
FBpp0082602	44-113	129-195	29
FBpp0085233	177-243	260-324	7
FBpp0085233	177-243	342-408	8
FBpp0085233	177-243	422-482	13
FBpp0085233	260-324	342-408	21
FBpp0085233	260-324	422-482	9
FBpp0085233	342-408	422-482	16
FBpp0072253	115-183	205-272	29
FBpp0111476	98-168	216-271	35
FBpp0075370	109-163	194-237	34
FBpp0070086	152-234	250-316	29
FBpp0070086	152-234	404-473	31
FBpp0070086	250-316	404-473	32
FBpp0082318	34-96	138-195	39
FBpp0076141	132-155	220-276	16
FBpp0099744	34-102	125-186	35
FBpp0070207	95-165	181-244	31
FBpp0111494	463-533	567-633	28
FBpp0082785	267-331	340-403	10
FBpp0082959	453-518	706-730	24
FBpp0099578	119-189	205-267	36
FBpp0071961	113-192	244-313	22
FBpp0071961	113-192	376-431	8
FBpp0071961	244-313	376-431	17
FBpp0073543	28-62	80-116	11
FBpp0073543	28-62	132-198	25
FBpp0073543	28-62	275-344	25
FBpp0073543	80-116	132-198	29
FBpp0073543	80-116	275-344	32
FBpp0073543	132-198	275-344	29
FBpp0084669	33-101	124-190	26

WormUnique	Worm RRM1 vs RRM2		Worm_Scores
CE14916	19-34	42-91	18
CE26612	11-73	98-160	9
CE21988	42-108	134-204	29
CE21988	42-108	258-319	12
CE21988	134-204	258-319	30
CE18030	579-620	640-705	14
CE34717	90-114	289-346	24
CE02193	59-128	147-215	24
CE02193	59-128	240-308	20
CE02193	59-128	345-413	30
CE02193	147-215	240-308	37
CE02193	147-215	345-413	37
CE02193	240-308	345-413	42
CE43237	227-296	329-392	32
CE03762	4-64	114-180	14
CE40238	70-136	160-226	17
CE42017	5-35	99-127	6
CE42017	5-35	157-209	19
CE42017	99-127	157-209	17
CE30079	72-140	183-236	22
CE42199	821-853	891-913	13

Table S 2.3 RRM-containing RBP orthologs in four species and their multiRRM scores

HumShared	Hum RRM1 vs RRM2		Hum_Scores	MouShared	Mou RRM1 vs RRM2		Mou_Scores
ENSP00000284073	24-91	112-178	43	ENSMUSP00000096222	86-140	177-234	27
ENSP00000373277	63-123	142-206	26	ENSMUSP00000096222	86-140	403-457	12
ENSP00000352612	69-139	155-221	38	ENSMUSP00000096222	177-234	403-457	18
ENSP00000352612	69-139	306-375	31	ENSMUSP00000104945	23-85	187-255	23
ENSP00000352612	155-221	306-375	26	ENSMUSP00000042658	16-84	107-175	34
ENSP00000334538	23-85	184-252	25	ENSMUSP00000089958	12-80	115-183	31
ENSP00000261741	4-72	298-363	15	ENSMUSP00000005041	151-225	261-330	18
ENSP00000261741	4-72	404-473	27	ENSMUSP00000005041	151-225	400-459	11
ENSP00000261741	4-72	589-642	20	ENSMUSP00000005041	261-330	400-459	18
ENSP00000261741	4-72	732-804	30	ENSMUSP00000001809	13-83	101-168	33
ENSP00000261741	4-72	834-904	30	ENSMUSP00000001809	13-83	193-261	33
ENSP00000261741	298-363	404-473	15	ENSMUSP00000001809	13-83	296-363	35
ENSP00000261741	298-363	589-642	9	ENSMUSP00000001809	101-168	193-261	45
ENSP00000261741	298-363	732-804	18	ENSMUSP00000001809	101-168	296-363	41
ENSP00000261741	298-363	834-904	27	ENSMUSP00000001809	193-261	296-363	52
ENSP00000261741	404-473	589-642	29	ENSMUSP00000066312	93-162	226-294	28
ENSP00000261741	404-473	732-804	32	ENSMUSP00000066312	93-162	516-583	39
ENSP00000261741	404-473	834-904	24	ENSMUSP00000066312	226-294	516-583	47
ENSP00000261741	589-642	732-804	29	ENSMUSP00000008477	27-84	153-212	17
ENSP00000261741	589-642	834-904	35	ENSMUSP00000066311	5-69	156-223	30
ENSP00000261741	732-804	834-904	25	ENSMUSP00000066311	5-69	285-351	16
ENSP00000271628	15-85	102-173	36	ENSMUSP00000066311	5-69	403-458	23
ENSP00000364912	8-68	339-406	19	ENSMUSP00000066311	5-69	760-829	26
ENSP00000364912	8-68	440-507	24	ENSMUSP00000066311	156-223	285-351	23
ENSP00000364912	8-68	535-587	24	ENSMUSP00000066311	156-223	403-458	28
ENSP00000364912	339-406	440-507	14	ENSMUSP00000066311	156-223	760-829	23
ENSP00000364912	339-406	535-587	22	ENSMUSP00000066311	285-351	403-458	19
ENSP00000364912	440-507	535-587	26	ENSMUSP00000066311	285-351	760-829	10
ENSP00000322016	88-158	185-254	41	ENSMUSP00000066311	403-458	760-829	30
ENSP00000322016	88-158	443-501	11	ENSMUSP00000084114	151-218	235-296	38
ENSP00000322016	185-254	443-501	8	ENSMUSP00000088365	113-180	230-332	5
ENSP00000246071	27-84	153-212	17	ENSMUSP00000093425	9-77	108-178	27
ENSP00000221419	120-174	211-268	27	ENSMUSP00000093425	9-77	216-280	36
ENSP00000221419	120-174	400-454	12	ENSMUSP00000093425	108-178	216-280	32
ENSP00000221419	211-268	400-454	18	ENSMUSP00000003501	41-111	127-190	37
ENSP00000352162	41-111	127-190	37	ENSMUSP00000003501	41-111	286-355	30
ENSP00000352162	41-111	286-355	30	ENSMUSP00000003501	127-190	286-355	28
ENSP00000352162	127-190	286-355	28	ENSMUSP00000105216	155-220	252-322	19
ENSP00000348345	14-84	113-182	42	ENSMUSP00000105216	155-220	452-505	7

ENSP00000348345	14-84	291-358	39
ENSP00000348345	113-182	291-358	42
ENSP00000258962	18-85	123-184	25
ENSP00000351409	19-84	110-175	33
ENSP00000351409	19-84	403-472	21
ENSP00000351409	110-175	403-472	24
ENSP00000363745	167-231	248-312	20
ENSP00000363745	167-231	346-407	25
ENSP00000363745	248-312	346-407	20
ENSP00000295470	151-218	235-296	38
ENSP00000260956	113-178	231-334	6
ENSP00000244020	4-64	112-177	24
ENSP00000228284	706-774	803-871	27
ENSP00000361949	13-83	101-168	29
ENSP00000361949	13-83	193-261	34
ENSP00000361949	13-83	296-363	30
ENSP00000361949	101-168	193-261	45
ENSP00000361949	101-168	296-363	42
ENSP00000361949	193-261	296-363	50
ENSP00000253363	155-220	252-322	19
ENSP00000253363	155-220	452-505	7
ENSP00000253363	252-322	452-505	12
ENSP00000313007	13-83	101-168	33
ENSP00000313007	13-83	193-261	31
ENSP00000313007	13-83	296-363	35
ENSP00000313007	101-168	193-261	47
ENSP00000313007	101-168	296-363	41
ENSP00000313007	193-261	296-363	52
ENSP00000349428	78-131	200-258	9
ENSP00000349428	78-131	383-435	18
ENSP00000349428	78-131	482-544	5
ENSP00000349428	200-258	383-435	24
ENSP00000349428	200-258	482-544	20
ENSP00000349428	383-435	482-544	22
ENSP00000240185	106-171	193-241	28
ENSP00000355089	56-123	154-216	41
ENSP00000355089	56-123	442-472	16
ENSP00000355089	154-216	442-472	29
ENSP00000355565	187-218	241-267	3
ENSP00000355565	187-218	289-358	34
ENSP00000355565	241-267	289-358	25
ENSP00000351168	227-296	328-400	37
ENSP00000351168	227-296	447-519	38
ENSMUSP00000105216	252-322	452-505	12
ENSMUSP00000107403	58-118	137-201	26
ENSMUSP00000007993	6-69	116-184	25
ENSMUSP00000007993	6-69	327-385	23
ENSMUSP00000007993	6-69	480-559	14
ENSMUSP00000007993	116-184	327-385	37
ENSMUSP00000007993	116-184	480-559	14
ENSMUSP00000007993	327-385	480-559	22
ENSMUSP00000030623	291-355	366-425	16
ENSMUSP00000048450	191-222	247-271	8
ENSMUSP00000048450	191-222	293-362	34
ENSMUSP00000048450	247-271	293-362	28
ENSMUSP00000005643	46-111	137-202	34
ENSMUSP00000005643	46-111	430-499	19
ENSMUSP00000005643	137-202	430-499	24
ENSMUSP00000035199	100-171	233-308	20
ENSMUSP00000045048	14-84	113-182	42
ENSMUSP00000045048	14-84	291-358	39
ENSMUSP00000045048	113-182	291-358	42
ENSMUSP00000105233	5-69	307-372	23
ENSMUSP00000105233	5-69	432-501	27
ENSMUSP00000105233	5-69	548-613	24
ENSMUSP00000105233	5-69	919-988	24
ENSMUSP00000105233	307-372	432-501	30
ENSMUSP00000105233	307-372	548-613	33
ENSMUSP00000105233	307-372	919-988	24
ENSMUSP00000105233	432-501	548-613	19
ENSMUSP00000105233	432-501	919-988	21
ENSMUSP00000105233	548-613	919-988	22
ENSMUSP00000101413	8-68	341-408	19
ENSMUSP00000101413	8-68	442-509	24
ENSMUSP00000101413	8-68	537-589	26
ENSMUSP00000101413	341-408	442-509	14
ENSMUSP00000101413	341-408	537-589	24
ENSMUSP00000101413	442-509	537-589	26
ENSMUSP00000075709	15-85	102-173	36
ENSMUSP00000103947	226-295	327-399	37
ENSMUSP00000103947	226-295	446-518	38
ENSMUSP00000103947	327-399	446-518	36
ENSMUSP00000017065	4-64	112-177	22
ENSMUSP00000101833	11-96	116-186	26
ENSMUSP00000101833	11-96	224-288	35
ENSMUSP00000101833	116-186	224-288	32

ENSP00000351168	328-400	447-519	36
ENSP00000341826	16-84	107-175	34
ENSP00000358089	11-96	116-186	26
ENSP00000358089	11-96	224-288	35
ENSP00000358089	116-186	224-288	32
ENSP00000382239	5-69	157-224	30
ENSP00000382239	5-69	286-353	24
ENSP00000382239	5-69	402-457	23
ENSP00000382239	5-69	928-997	20
ENSP00000382239	157-224	286-353	26
ENSP00000382239	157-224	402-457	28
ENSP00000382239	157-224	928-997	22
ENSP00000382239	286-353	402-457	14
ENSP00000382239	286-353	928-997	8
ENSP00000382239	402-457	928-997	30
ENSP00000310471	4-64	81-141	44
ENSP00000307863	151-225	261-330	18
ENSP00000307863	151-225	400-459	11
ENSP00000307863	261-330	400-459	18
ENSP00000313890	159-215	339-408	14
ENSP00000313890	159-215	420-480	8
ENSP00000313890	339-408	420-480	27
ENSP00000282574	9-77	108-178	27
ENSP00000282574	9-77	216-280	36
ENSP00000282574	108-178	216-280	32
ENSP00000343054	100-171	233-308	20
ENSP00000376309	37-106	128-196	36
ENSP00000349748	299-363	374-433	16
ENSP00000233078	12-80	115-183	31
ENSP00000363228	5-69	307-372	23
ENSP00000363228	5-69	432-501	26
ENSP00000363228	5-69	547-612	24
ENSP00000363228	5-69	859-928	24
ENSP00000363228	307-372	432-501	31
ENSP00000363228	307-372	547-612	33
ENSP00000363228	307-372	859-928	25
ENSP00000363228	432-501	547-612	22
ENSP00000363228	432-501	859-928	21
ENSP00000363228	547-612	859-928	22
ENSP00000223073	6-68	116-184	26
ENSP00000223073	6-68	337-395	25
ENSP00000223073	6-68	490-567	14
ENSP00000223073	116-184	337-395	37
ENSMUSP00000098096	136-206	233-302	41
ENSMUSP00000098096	136-206	491-549	13
ENSMUSP00000098096	233-302	491-549	11
ENSMUSP00000102733	70-140	156-222	38
ENSMUSP00000102733	70-140	308-377	31
ENSMUSP00000102733	156-222	308-377	26
ENSMUSP00000038256	4-64	81-141	44
ENSMUSP00000107595	37-106	128-196	36
ENSMUSP00000078792	18-85	123-184	25
ENSMUSP00000019118	706-775	803-872	27
ENSMUSP00000079070	13-83	101-168	29
ENSMUSP00000079070	13-83	193-261	34
ENSMUSP00000079070	13-83	296-363	30
ENSMUSP00000079070	101-168	193-261	44
ENSMUSP00000079070	101-168	296-363	42
ENSMUSP00000079070	193-261	296-363	48
ENSMUSP00000059330	12-81	93-153	27
ENSMUSP00000101476	167-231	248-309	22
ENSMUSP00000101476	167-231	343-404	25
ENSMUSP00000101476	248-309	343-404	20
ENSMUSP00000038113	106-171	193-242	24
ENSMUSP00000031590	4-72	297-362	12
ENSMUSP00000031590	4-72	402-471	24
ENSMUSP00000031590	4-72	587-639	18
ENSMUSP00000031590	4-72	724-796	27
ENSMUSP00000031590	4-72	826-896	31
ENSMUSP00000031590	297-362	402-471	4
ENSMUSP00000031590	297-362	587-639	9
ENSMUSP00000031590	297-362	724-796	21
ENSMUSP00000031590	297-362	826-896	27
ENSMUSP00000031590	402-471	587-639	28
ENSMUSP00000031590	402-471	724-796	30
ENSMUSP00000031590	402-471	826-896	25
ENSMUSP00000031590	587-639	724-796	28
ENSMUSP00000031590	587-639	826-896	37
ENSMUSP00000031590	724-796	826-896	25
ENSMUSP00000111483	56-123	153-215	41
ENSMUSP00000111483	56-123	422-490	22
ENSMUSP00000111483	153-215	422-490	25
ENSMUSP00000093109	77-130	199-257	9
ENSMUSP00000093109	77-130	381-433	7
ENSMUSP00000093109	77-130	480-542	5
ENSMUSP00000093109	199-257	381-433	24

ENSP00000223073	116-184	490-567	13	ENSMUSP00000093109	199-257	480-542	20
ENSP00000223073	337-395	490-567	25	ENSMUSP00000093109	381-433	480-542	22
ENSP00000316950	102-171	235-303	28	ENSMUSP00000090470	24-91	112-178	43
ENSP00000316950	102-171	525-592	39				
ENSP00000316950	235-303	525-592	44				

FlyShared	**Fly RRM1 vs RRM2**		**Fly_Scores**	**WormShared**	**Worm RRM1 vs RRM2**		**Worm_Scores**
FBpp0085916	4-74	92-160	27	CE18963	5-73	177-239	14
FBpp0085916	4-74	185-249	24	CE18963	5-73	281-349	28
FBpp0085916	4-74	289-357	30	CE18963	5-73	501-554	14
FBpp0085916	92-160	185-249	38	CE18963	5-73	643-719	13
FBpp0085916	92-160	289-357	43	CE18963	5-73	750-820	23
FBpp0085916	185-249	289-357	52	CE18963	177-239	281-349	11
FBpp0077781	556-626	658-724	14	CE18963	177-239	501-554	11
FBpp0077781	556-626	753-804	15	CE18963	177-239	643-719	17
FBpp0077781	658-724	753-804	19	CE18963	177-239	750-820	11
FBpp0080905	152-223	264-366	11	CE18963	281-349	501-554	14
FBpp0076875	233-297	312-378	29	CE18963	281-349	643-719	26
FBpp0078974	9-77	98-165	35	CE18963	281-349	750-820	21
FBpp0074012	304-368	378-441	14	CE18963	501-554	643-719	11
FBpp0079471	12-70	171-227	5	CE18963	501-554	750-820	22
FBpp0083366	166-231	247-308	29	CE18963	643-719	750-820	25
FBpp0083366	166-231	342-403	32	CE26020	7-73	125-194	32
FBpp0083366	247-308	342-403	22	CE26020	7-73	473-541	43
FBpp0083976	9-76	97-167	27	CE26020	125-194	473-541	36
FBpp0083976	9-76	223-287	20	CE20412	34-103	122-190	27
FBpp0083976	97-167	223-287	27	CE20412	34-103	215-282	25
FBpp0077324	112-182	198-270	33	CE20412	34-103	319-387	31
FBpp0077324	112-182	363-432	35	CE20412	122-190	215-282	33
FBpp0077324	198-270	363-432	30	CE20412	122-190	319-387	40
FBpp0085681	99-156	258-314	22	CE20412	215-282	319-387	47
FBpp0085681	99-156	369-429	12	CE29126	174-240	274-343	8
FBpp0085681	258-314	369-429	5	CE38662	31-99	112-180	31
FBpp0072706	5-65	362-423	26	CE27708	48-114	137-207	31
FBpp0072706	5-65	478-546	11	CE27708	48-114	243-305	31
FBpp0072706	5-65	741-773	12	CE27708	137-207	243-305	28
FBpp0072706	5-65	911-981	29	CE24110	5-70	876-943	16
FBpp0072706	362-423	478-546	25	CE03763	5-65	131-196	19
FBpp0072706	362-423	741-773	21	CE07355	28-85	145-204	10
FBpp0072706	362-423	911-981	20	CE08718	112-170	230-335	22
FBpp0072706	478-546	741-773	21	CE00983	44-114	130-196	31
FBpp0072706	478-546	911-981	23	CE00983	44-114	376-444	31
FBpp0072706	741-773	911-981	18	CE00983	130-196	376-444	31
FBpp0082724	9-75	117-178	22	CE03111	14-74	165-234	4

FBpp0087934	114-161	316-384	8
FBpp0087934	114-161	398-456	25
FBpp0087934	316-384	398-456	25
FBpp0072593	132-202	229-298	32
FBpp0072593	132-202	560-622	14
FBpp0072593	229-298	560-622	6
FBpp0088567	109-174	194-242	26
FBpp0082316	26-94	117-185	30
FBpp0081601	59-128	234-302	24
FBpp0081601	59-128	565-628	28
FBpp0081601	234-302	565-628	14
FBpp0085240	257-323	422-478	5
FBpp0085240	257-323	612-664	11
FBpp0085240	257-323	714-774	11
FBpp0085240	422-478	612-664	24
FBpp0085240	422-478	714-774	8
FBpp0085240	612-664	714-774	16
FBpp0076112	4-72	254-314	11
FBpp0076112	4-72	366-435	24
FBpp0076112	4-72	571-622	9
FBpp0076112	4-72	681-755	15
FBpp0076112	4-72	787-858	24
FBpp0076112	254-314	366-435	19
FBpp0076112	254-314	571-622	3
FBpp0076112	254-314	681-755	9
FBpp0076112	254-314	787-858	21
FBpp0076112	366-435	571-622	30
FBpp0076112	366-435	681-755	27
FBpp0076112	366-435	787-858	25
FBpp0076112	571-622	681-755	11
FBpp0076112	571-622	787-858	23
FBpp0076112	681-755	787-858	19
FBpp0088583	239-304	337-406	21
FBpp0088583	239-304	511-565	10
FBpp0088583	337-406	511-565	18
FBpp0088856	51-123	149-219	42
FBpp0088856	51-123	483-550	29
FBpp0088856	149-219	483-550	32
FBpp0076553	9-70	88-150	54
FBpp0082320	58-126	138-206	31
FBpp0070716	9-79	144-203	21
FBpp0082269	6-66	117-182	18
FBpp0074013	95-169	209-278	17
CE36388	202-266	283-350	10
CE36388	202-266	379-440	24
CE36388	283-350	379-440	14
CE31214	97-165	265-333	21
CE31214	97-165	414-484	23
CE31214	265-333	414-484	10
CE41585	58-122	144-208	36
CE41585	58-122	505-575	23
CE41585	144-208	505-575	27
CE35148	186-246	283-351	29
CE35148	186-246	391-451	18
CE35148	283-351	391-451	31
CE02231	176-242	261-328	10
CE36374	15-85	102-172	28
CE31089	11-77	125-186	19
CE42671	118-181	191-251	13
CE27339	186-260	293-361	20
CE27339	186-260	425-481	8
CE27339	293-361	425-481	14
CE26949	228-300	365-421	10
CE05167	595-660	685-753	21
CE39251	25-93	116-184	33
CE17724	36-99	171-230	15
CE17724	36-99	253-311	11
CE17724	171-230	253-311	23
CE34752	34-96	140-197	22
CE34752	34-96	325-379	5
CE34752	140-197	325-379	14
CE28911	145-231	248-317	27
CE08369	47-116	190-258	15
CE08369	47-116	387-449	25
CE08369	190-258	387-449	11
CE24366	104-174	209-278	28
CE24366	104-174	679-735	14
CE24366	209-278	679-735	10
CE29337	48-115	136-202	46
CE32618	33-68	293-335	13
CE36103	31-97	126-188	36
CE36103	31-97	431-500	19
CE36103	126-188	431-500	26
CE30508	137-193	248-306	28
CE30508	137-193	438-493	10
CE30508	137-193	540-600	3

FBpp0074013	95-169	344-399	8
FBpp0074013	209-278	344-399	10
FBpp0070859	15-85	102-172	29
FBpp0072239	50-118	232-299	17
FBpp0072239	50-118	381-449	11
FBpp0072239	232-299	381-449	17
FBpp0110257	298-364	384-448	38
FBpp0110257	298-364	810-880	28
FBpp0110257	384-448	810-880	23
FBpp0080510	157-235	262-329	30
FBpp0075557	57-119	339-408	25
FBpp0077631	260-326	390-468	16
FBpp0086479	281-350	382-458	35
FBpp0086479	281-350	548-619	24
FBpp0086479	382-458	548-619	26
CE30508	248-306	438-493	28
CE30508	248-306	540-600	15
CE30508	438-493	540-600	7

Table S 2.4 DAZ orthologs identified in different species (InParanoid and manual Blast)

	Chimpanzee InParanoid.H.sapiens-P.troglodytes.tgz					Macaque InParanoid.H.sapiens-M.mulatta.tgz				
	Ortholog_group	InParanoid Best Score or Blast Score	Species	Uniprot_ID	Bootstrap_support_as_seed_ortholog	Ortholog_group	InParanoid Best Score or Blast Score	Species	Uniprot_ID	Bootstrap_support_as_seed_ortholog
BOLL Q8N9W6 (BOLL_HUMAN)	12135	587	H.sapiens	Q8N9W6	100%	11313	584	H.sapiens	Q8N9W6	100%
	12135	587	P.troglodytes	H2QJ74	100%	11313	584	M.mulatta	F7HR30	100%
DAZL Q92904 (DAZL_HUMAN)	11774	609	H.sapiens	Q92904	100%	10944	605	H.sapiens	Q92904	100%
	11774	609	P.troglodytes	H2QM56	100%	10944	605	M.mulatta	F6WG03	100%
DAZ3 Q9NR90 (DAZ3_HUMAN)	Additional sequence search	2,425	H.sapiens	Q9NR90	Length: 510, Identity: 88.0% E-value: 0.0	Additional sequence search	1,658	H.sapiens	Q9NR90	Length: 551, Identity: 82.0% E-value: 0.0
	Additional sequence search	2,114	H.sapiens	Q9NR90	Length: 438, Identity: 82.0% E-value: 0.0	Additional sequence search	1,150	H.sapiens	Q9NR90	Length: 559, Identity: 49.0% E-value: 6.0×10-149
DAZ2 Q13117 (DAZ2_HUMAN)	7204	968	H.sapiens	Q13117	89%	Additional sequence search	1,662	H.sapiens	Q13117	Length: 551, Identity: 82.0% E-value: 0.0
	7204	968	P.troglodytes	H2RB60	100%	Additional sequence search	1139	H.sapiens	Q13117	Length: 559, Identity: 46.0% E-value: 3.0×10-146
DAZ4 Q86SG3 (DAZ4_HUMAN)	Additional sequence search	2,797	H.sapiens	Q86SG3	Length: 743, Identity: 91.0% E-value: 0.0	Additional sequence search	2,418	H.sapiens	Q86SG3	Length: 551, Identity: 83.0% E-value: 0.0
	Additional sequence search	2,054	H.sapiens	Q86SG3	Length: 438, Identity: 88.0% E-value: 0.0					
	Additional sequence search	2,033	H.sapiens	Q86SG3	Length: 414, Identity: 91.0% E-value: 0.0					
DAZ1 Q1RMF9 (Q1RMF9_HUMAN)	Additional sequence search	3605	H.sapiens	Q1RMF9	Length: 743, Identity: 92.0% E-value: 0.0					

	Gorilla InParanoid.G.gorilla-H.sapiens.tgz					Chicken InParanoid.G.gallus-H.sapiens.tgz				
	Ortholog_group	InParanoid Best Score or Blast Score	Species	Uniprot_ID	Bootstrap_support_as_seed_ortholog	Ortholog_group	InParanoid Best Score or Blast Score	Species	Uniprot_ID	Bootstrap_support_as_seed_ortholog
BOLL Q8N9W6 (BOLL_HUMAN)	12345	558	G.gorilla	G3RCN1	100%	8097	429	G.gallus	F1NDD0	100%
	12345	558	H.sapiens	Q8N9W6	100%	8097	429	H.sapiens	Q8N9W6	100%
DAZL Q92904 (DAZL_HUMAN)	11581	611	G.gorilla	G3SGI9	100%	8619	388	G.gallus	Q804A9	100%
	11581	611	H.sapiens	Q92904	100%	8619	388	H.sapiens	Q92904	100%
DAZ3 Q9NR90 (DAZ3_HUMAN)	Additional sequence search	841	H.sapiens G.gorilla	Q9NR90 G3S8R9	Length: 295, Identity: 83.0% E-value: 3.0×10-106					
DAZ2 Q13117 (DAZ2_HUMAN)	Additional sequence search	841	H.sapiens G.gorilla	Q13117 G3S8R9	Length: 295, Identity: 83.0% E-value: 3.0×10-105					
DAZ4 Q86SG3 (DAZ4_HUMAN)										
DAZ1 Q1RMF9 (Q1RMF9_HUMAN)										
	Zebrafish InParanoid.D.rerio-H.sapiens.tgz					Frog InParanoid.H.sapiens-X.tropicalis.tgz				

	Ortholog_group	InParanoid Best Score or Blast Score	Species	Uniprot_ID	Bootstrap_support_as_seed_ortholog	Ortholog_group	InParanoid Best Score or Blast Score	Species	Uniprot_ID	Bootstrap_support_as_seed_ortholog
BOLL Q8N9W6 (BOLL_HUMAN)										
DAZL Q92904 (DAZL_HUMAN)	11476	155	D.rerio	Q9YGW7	100%	10148	290	H.sapiens	Q92904	100%
	11476	155	H.sapiens	Q92904	99%	10148	290	X.tropicalis	Q76CY5	100%
DAZ3 Q9NR90 (DAZ3_HUMAN)						12219	88	H.sapiens	Q13117	100%
						12219	88	H.sapiens	Q9NR90	
						12219	88	H.sapiens	Q86SG3	
						12219	88	H.sapiens	Q9NQZ3	
						12219	88	X.tropicalis	F6SK87	100%
DAZ2 Q13117 (DAZ2_HUMAN)						12219	88	H.sapiens	Q13117	100%
						12219	88	H.sapiens	Q86SG3	
						12219	88	H.sapiens	Q9NQZ3	
						12219	88	X.tropicalis	F6SK87	100%
DAZ4 Q86SG3 (DAZ4_HUMAN)						12219	88	H.sapiens	Q13117	100%
						12219	88	H.sapiens	Q9NR90	
						12219	88	H.sapiens	Q86SG3	
						12219	88	H.sapiens	Q9NQZ3	
						12219	88	X.tropicalis	F6SK87	100%
DAZ1 Q1RMF9 (Q1RMF9_HUMAN)										

CHAPTER 3

TRANSCRIPTOME-WIDE IDENTIFICATION AND STUDY OF CANCER-SPECIFIC SPLICING EVENTS ACROSS MULTIPLE TUMORS[2]

3.1 Overview

Dysregulation of alternative splicing (AS) is one of molecular hallmarks of cancer, with splicing alteration of numerous genes in cancer patients. However, studying splicing mis-regulation in cancer is complicated by large noise of tissues-specific splicing. To obtain a global picture of cancer-specific splicing, we analyzed transcriptome sequencing data from 1149 patients in TCGA project, producing a core set of AS events significantly altered across multiple cancer types. These cancer-specific AS events are highly conserved, more likely to maintain protein reading frame, and mainly function in cell cycle, cell adhesion/migration, and insulin signaling pathway. Furthermore, these events can serve as new molecular biomarkers to distinguish cancer from normal tissues, to separate cancer subtypes, and to predict patient survival. We also found that most genes whose expression is closely associated with cancer-specific splicing are key regulators of the cell cycle. This study uncovers a common set of

[2] This chapter previously appeared as an article in the Journal of *Oncotarget*. The original citation is as follow: Tsai YS, Dominguez D, Gomez SM, Wang Z. "Transcriptome-wide identification and study of cancer-specific splicing events across multiple tumor," *Oncotarget*. (in press, Feb 2015)

cancer-specific AS events altered across multiple cancers, providing mechanistic insight into how splicing is mis-regulated in cancers.

3.2 Introduction

Most human genes undergo alternative splicing (AS) to produce multiple isoforms with different biological properties. This process is tightly controlled across different tissues and developmental stages, and dysregulation of AS is closely associated with various human diseases including cancer [14, 15]. The extensive alteration of AS is considered to be one of the molecular hallmarks of cancer [83], and affects numerous genes that are critical for tumor pathogenesis and progression (e.g. apoptosis, angiogenesis, tumor metastasis) [84]. While most genetic mutations occur at a low frequency in cancer patients (with a few exceptions like TP53), many identified cancer-specific AS events were found in more than half of the tumor samples, suggesting a predominant role of splicing dysregulation in cancer [14, 84]. For example, CD44 is a key mediator of cell-cell and cell-matrix interactions, migration and invasion [85], and different splicing isoforms of CD44 have been linked with tumor evasion and met Figure astasis in many cancers [86-88]. Other well-documented cases include the apoptosis regulator Bcl-x, which can shift its splicing from pro-apoptotic into anti-apoptotic isoforms in cancers [89].

AS is generally regulated by multiple *cis*-acting splicing regulatory elements (SREs) that are specifically bound by *trans*-acting splicing factors to enhance or inhibit the use of nearby splice sites [23, 90]. The same splicing factor may either activate or inhibit splicing by binding to its cognate SREs in different pre-mRNA regions, which is commonly referred to as context dependent activity [18, 20, 21]. Various cellular signaling pathways, such as the MEK/ERK or c-

Myc pathway [91-93], were found to control the expression level and activity of splicing factors, which in turn determine different splicing patterns in distinct tissues (reviewed in [90, 94]). Many splicing factors are involved in cancer pathogenesis through mediating AS of hundreds of genes [84]. For example, the splicing factor SRSF1 is found to act as a proto-oncogene to promote cell transformation [46], whereas the splicing suppressor RBM4 functions as a tumor suppressor to inhibit tumor progression[49]. These two antagonistic factors control a partially overlapping set of AS events that are involved in cell migration and apoptosis [49].

While the global change of splicing in cancers is being increasingly appreciated, the functional consequences and regulatory mechanisms of cancer-specific AS remain poorly understood. In addition, detailed analyses of cancer-associated splicing were previously focused on single tumor types or specific genes [95, 96], which is often dominated by the noise from tissue specific AS events. Since cancer is a highly heterogeneous disease, the genetic variation between samples has made the identification of cancer-specific splicing isoforms difficult. Recent advance in The Cancer Genome Atlas (TCGA) project has provided tremendous amounts of sequencing data from the transcriptome of thousands of samples in different cancer types [8], making it possible for an unbiased identification and a further analysis of "cancer-specific" splicing events across different cancer types.

In this study, we performed a transcriptome-wide splicing comparison between thousands of tumor samples and paired normal controls to identify a large number of splicing events with altered splicing in cancer. Most of these events were found to change in a single cancer type, and we further identified a core set of cancer-specific AS events across three different cancer types. The genes containing cancer-specific AS events are significantly enriched for functions in cell cycle, cell adhesion/migration, and insulin signaling pathway. Detailed analyses suggested

that cancer-specific cassette exons are more conserved among vertebrates and more likely to maintain the protein reading frame. The set of cancer-specific AS events can serve as reliable biomarkers to separate tumor from normal samples and to even distinguish different subtypes of breast cancer. Finally, we found that most genes whose expression is closely associated with cancer-specific splicing are also key regulators of cell cycle, providing a previously unknown link between cell cycle and splicing regulation in cancer cells.

3.3 Results

3.3.1 Identification of cancer-specific AS events common to multiple cancers.

It is well known that splicing is controlled in a tissue specific manner with a global change of the splicing landscape between different tissues [10, 11]. Thus identification of AS events altered in cancer vs. normal cells is often complicated by the tissue types used. It remains unclear what portion of AS events are commonly mis-regulated across all cancer types vs. those specific to certain cancers. To identify cancer-specific AS events that are changed within and/or across multiple cancers, we used RNA-seq data collected through TCGA project [8]. The aligned RNA-seq reads were processed through the MISO pipeline to estimate ratios of different splicing isoforms for each annotated splicing event. For each AS event, we calculated the PSI (Percent Spliced In) values between all normal and tumor samples, and identified potential cancer-specific AS events that are significantly altered in cancer vs. normal tissues (Figure 3.1A).

For a reliable comparison, we selected three types of cancers that have sufficient number of paired normal samples from TCGA, including breast invasive carcinoma (BRCA), lung squamous cell carcinoma (LUSC) and liver hepatocellular carcinoma (LIHC). We focused on

four major modes of AS for more detailed analysis: skipped exon (SE), retained intron (RI), alternative 3' splice site (A3SS) and alternative 5' splice site (A5SS). In each cancer type, we identified AS events that satisfy the following criteria: (i) the AS event is detected in at least 10 tumor samples and 10 normal samples; (ii) the distributions of PSI values of each event are significantly different between normal and tumor samples ($p < 0.05$ by t-test); (iii) the mean difference of PSI values between normal and tumor samples is large than 0.1 (Figure 3.1B). For each cancer type, we identified several thousand AS events that have significant changes of the splicing isoforms (table 3.1). Most of these events are specific to a single cancer type, with lung cancer having more altered AS events compare to breast and liver cancers, probably due to the higher mutation rate in lung cancer [97].

We considered the common events that changed in all three types of cancers as a core set of 163 cancer-specific AS events (Figure 3.1B and Supplementary Table S3.1). As a background control, we simulated the overlaps of AS events between different cancer types using 1000 randomly selected datasets with matched size (see methods). In all four AS modes, the overlaps between simulated datasets are significantly smaller than those between real set, especially for the overlaps of all three cancers (Figure 3.1C). This result suggests that although each cancer type has tissue-specific set of splicing events, there are indeed a significant number of splicing events shared by multiple cancers. Strikingly, 10 of the genes that show cancer-specific AS are also frequently mutated in cancers vs. normal samples [98]. Such overlap is significantly more than the overlap expected by chance ($p=10^{-6}$ by hypergeometric test), indicating that the function of these genes may be altered through either mutation or splicing changes to affect cancer progression. This result also suggests that, in addition to point mutations and copy number

variations, the alteration of splicing may serve as another important route to alter gene function in cancer.

Splicing of most introns is regulated in a co-transcriptional fashion [94], thus a change in gene expression may affect AS outcomes of corresponding genes. We examined the genes containing cancer-specific AS events, and found that most of them (68%-90%) do not have significant change of expression levels between normal vs. tumor samples (Supplementary Figure S3.1A), suggesting that splicing of these events are not biased by their gene expression. In addition, we found that similar numbers of genes have increased vs. decreased PSI values, with exception of RI events that tend to have more retained introns (i.e. PSI increased) in cancers (Supplementary Figure S3.1B). Intron retention is a relatively less studied mode of AS in mammals and usually changes the coding frame and triggers nonsense mediated mRNA decay (NMD) [99]. Therefore an increase in intron retention may represent an important mechanism for protein inactivation in tumors.

3.3.2 Consistent change of cancer-specific AS events across tumor types.

To determine if the cancer-specific AS events change consistently among different types of cancers, we compared the difference of PSI values between cancer and normal (ΔPSI) across three cancer types for each event (Figure 3.2A). The majority (i.e. 85%) of these cancer-specific AS events change consistently across different tumor types (i.e., with an increased or decreased PSI values in all three tumors) when comparing tumors to the cognate normal tissue, suggesting that the splicing change in these genes will likely generate similar functional consequences across different tumors. The remaining 15% of AS events, while being altered across all cancers, have different patterns of splicing changes depending on the cancer type.

Representative examples of cancer-specific AS events were arbitrarily selected to illustrate splicing changes between tumor and normal samples. We chose one example from each AS mode and used colored lines to represent the ΔPSI between the paired tumor *vs.* the adjacent normal tissue (Figure 3.2B, upper panel). We found that there is large heterogeneity of ΔPSI between the paired cancer-normal samples. In some cases of breast cancer, both an increase and decrease in PSI were found among different patients, which might reflect differences between breast cancer subtypes. In addition, we also plotted the distribution of the PSI for the same examples in all normal and tumor samples, and found that the changes of PSI are consistent with those found in paired samples (white and gray boxes, bottom of Figure 3.2B).

3.3.3 Biological functions of cancer-specific AS events.

We further examined the genes containing newly identified cancer-specific AS events, and found that many of these genes are known to play a key roles in different stages of tumor progression (Supplementary Table S3.1). Most of these genes are functionally related to each other and form closely connected protein interaction networks (Figure 3.2C). Based on the MCODE clustering algorithm [100], these genes can be clustered into three groups connected by several hub proteins (Figure 3.2C and Supplementary Table S3.2). The largest group contains genes involved in cell cycle regulation (such as AURKB, CDCA5), with genes in the other two groups having functions in mediating cell adhesion/migration (e.g. CD44 and Collagens) and involved in the insulin signaling pathway (e.g. INSR, PPARG). The functional clustering of genes with cancer-specific AS suggests that the regulation of AS in cancer plays important roles in key pathways related to cancer pathogenesis, including cell cycle and cell adhesion/migration. The association of insulin response pathway with splicing regulation in cancer has not been reported before, and its functional implication will be an interesting subject of future studies.

To further study the functional consequence of cancer-specific AS events in an unbiased fashion, we performed gene ontology (GO) analysis on genes containing cancer-specific AS events using the DAVID online tool (http://david.abcc.ncifcrf.gov/) [74, 101]. We found that the most enriched functional categories included cell adhesion, cell division, cell cycle and so on. (Figure 3.3A). We also did GO analysis on each individual cancer type and ranked enriched GO terms by their p-value (Supplementary Figure S3.2). The top enriched terms were cytoskeleton proteins and proteins associated with cell adhesion, ATP-binding, cell cycle.

3.3.4 Sequence characteristics of cancer-specific AS events.

The skipped exon is the most common mode of AS among all identified cancer-specific AS events (Figure 3.1B), providing a sufficient amount of data for detailed sequence analyses. We first measured the length of all skipped exons, and found no obvious difference between the cancer-specific SEs, the SEs that are altered in a single cancer type, and all annotated SEs as control (Figure 3.3B). We further examined if the lengths of these alternative exons can be divided by three, which is a good indication of how each AS event affects mRNA reading frame. We found that 42% of the alternative exons in control set of SEs are phase 0 exon (i.e. maintain their reading frame), while in the sets of SEs altered in single or multiple types of cancers, a notably increased fraction of exons maintain their reading frame (Figure 3.3C). In particular, 53% of the SEs shared by all three cancers are phase 0 exons, significantly more than what is expected by chance ($p=0.008$ by fisher's exact test). Since disruption of reading frame often introduces premature stop codons that lead to NMD, the increased tendency of cancer-specific SEs to maintain reading frame suggests that these events tend to produce proteins with different functions rather than disrupting protein function via changing reading frame.

Alternative splicing is generally regulated by *cis*-acting SREs that function as splicing enhancers or silencers [18]. These SREs usually function in the nearby region of alternative splice sites, and thus the pre-mRNA regions within and adjacent to the alternative exons are more conserved than corresponding regions near the constitutive exons. When further examining the conservation of pre-mRNA regions near 124 cancer-specific SEs shared by three cancer types, we found that these exons tend to be highly conserved across 100 vertebrate species in the adjacent regions. Such sequence conservation is even higher than alternative exons that are spliced in a cancer-independent manner (Figure 3.3D, comparing black and grey lines), suggesting that cancer-specific alternative exons are under additional evolutionary constraints. This result is consistent with the notion that alternative splicing of cancer-associated genes are tightly controlled in normal cells across different species to mediate critical and highly conserved processes in cell growth.

To further identify putative SREs that control cancer-specific SEs, we examined these highly conserved regulatory regions to measure whether there are enriched sequence motifs that could be potentially recognized by splicing factors (Supplementary Figure S3.3). Some of these motifs resemble the binding site of known splicing factor. For example, the AC-rich motifs are recognized by hnRNP L and the UG rich motifs resemble hnRNP H/F binding sites [21]. Consistently, the hnRNP H was shown to be up-regulated in certain cancer and control several cancer related splicing events [102, 103].

3.3.5 Splicing of cancer-specific AS events are highly fluctuated.

The ratios between different splicing isoforms of the same gene are tightly regulated to ensure precise control of gene function. In normal cells, splicing is usually controlled in a tissue-specific fashion with certain dominant isoforms in different tissues [10, 11]. However,

such dominance of certain tissue-specific isoforms is often absent in cancer cells. In another word, many splicing isoforms are found in the "wrong tissues", leading to a more dispersed spectrum of AS. However such deregulation of AS in cancer has only been observed in an anecdotal fashion, and a thorough investigation with correct controls is lacking.

To examine potential splicing deregulation in cancers, we directly test: (i) if the splicing of cancer-specific AS events have higher variability than control events, and (ii) if such variability is higher in cancer *vs.* normal samples. We calculated the standard deviation (SD) of PSI value for each AS event, which measures the amount of variation from the average. The SDs of cancer-specific AS events were compared to those of control AS events across both normal and tumor samples in each cancer type. Since the mean value of PSI dramatically affects its SD (Supplementary Figure S3.4, PSI values near 0 or 1 tend to have smaller SD), we randomly picked control AS events from the MISO database with PSI distribution matched to cancer-specific AS events. The selection of such controls can eliminate potential biases caused by different PSI distribution between cancer-specific AS events vs. all other AS events.

We found that all the cancer-associated AS events have higher PSI variation than controls in all three tissue types (Figure 3.3E, comparing 2 boxes at the right to the ones at left), suggesting that splicing of these events are indeed highly variable across different samples. In addition, when comparing the cancer-specific AS event in tumor sample with normal samples, those cancer-AS events still tend to have higher variability in tumors than in normal samples (p-value: 1.2×10^{-17}, 5.3×10^{-6} and 6.0×10^{-6} for BRCA, LIHC and LUSC respectively, Figure 3.3E). In each cancer type, we also plotted the distribution of PSI in histograms in supplementary figure S3.5 using both all AS events and the 163 cancer-specific events. This result is consistent with the popular hypothesis that the tissue specificity of AS in normal samples is disrupted in cancers,

probably due to extensive changes in the expression levels and/or activities of oncogenic splicing factors.

3.3.6 Cancer-specific AS events as molecular biomarkers.

Identification of a core set of cancer-specific AS events makes it possible to use this relatively small dataset as a new molecular biomarker of cancers. To this end, we conducted principal component analysis (PCA) using the 163 cancer-specific AS events. For each tumor or normal sample, we generated a vector with 163 variables using the PSI values of cancer-specific AS events. We constructed a data matrix consist of all tumor and normal samples in each cancer type and further analyzed with PCA. The first two principal components in PCA accounted for 30%, 25% and 24% of the total variance for LIHC, LUSC and BRCA samples respectively (Supplementary Figure S3.6). The distribution of all samples was plotted using the first two principal components, which show a clear separation between cancer and normal samples (Figure 3.4A). All analyses showed a reliable separation of samples into two clusters (labeled with red and blue for tumor and normal samples respectively), suggesting that the 163 cancer-specific AS can potentially serve as a reliable biomarker for cancer diagnosis.

In addition, we combined all samples from three types of cancers and analyzed combined data with a similar PCA procedure. Consistent with our analyses of single cancer type, the cancer and normal samples can be reliably separated with the first two principal components (Figure 3.4B, “C” for cancer and “N” for normal), indicating that cancer-specific AS events are useful molecular biomarkers to separate tumors from mixed samples. In addition, the samples from different tissue types can be roughly separated (Figure 3.4B, with orange, green and black representing breast, lung and liver respectively). This result suggests that although the 163 AS

events are identified based on their altered splicing in multiple cancers, their splicing patterns still partially reflect tissue of origin.

Breast cancer is a well-annotated cancer type in TCGA data and is classified into several subtypes based on histopathological criteria and expression of a core set of genes [104, 105]. The breast cancer cells in different subtypes (Claudin-low, Basal-like, HER2-enriched, Luminial B and Luminal A) resemble cells in different stages of normal mammary development, which is well correlated with tumor progression (Figure 3.5A) [105]. Since our BRCA dataset has a large number of samples with well-annotated subtype categories, we sought to determine if the cancer-specific AS events can be used to separate cancer subtypes. We conducted a similar PCA procedure using 818 breast samples that were independently classified into normal and four cancer subtypes by PAM50 [104]. By plotting all samples along the first two principal components, we found that different breast cancer subtypes tend to be clustered into different groups (Figure 3.5B). Certain subtypes of breast cancers, such as basal and luminal types, are particularly well separated. In addition, some normal samples that were misclassified as luminal A cancers by PAM50 were correctly distinguished using cancer-specific AS events, suggesting that the cancer-specific AS events can potentially serve as a cancer biomarker independent of current classification criteria using gene expression data. Although current separation by two PCA components is not very strong with some overlaps between subtypes, this analysis provided a proof-of-concept for splicing-based tumor classification. A more sophisticated statistical approach and analysis is needed to prove this. A representative example of cancer-specific AS events was selected to show that alteration of splicing patterns in the same gene could be different in distinct subtypes of breast cancer (Figure 3.5C), with the luminal A subtype having the largest variability between patients.

3.3.7 Ratio of different splicing variants can serve as predictor of cancer survival.

Until around 2000, the chance of survival for cancer patients was mainly predicted according to various histologic and clinical characteristics. The advance of microarray technology led to more accurate profiling of gene expression in cancers, allowing prediction of cancer survival by the gene-expression signature of cancer [106, 107].

The extensive splicing mis-regulation and frequent mutations of certain splicing factors in cancer have suggested that some AS events may directly affect tumor biogenesis and progression, however the consequence of splicing mis-regulation on patient survival remains unclear. The identification of cancer-specific AS events across a large number of patients makes it possible to test if splicing mis-regulation in certain genes can serve as a predictor of cancer prognosis.

We directly test this possibility using TCGA dataset of breast cancer which has largest number of patients. We separated 727 BRCA patients according to the ratio of different splicing isoform for each cancer-specific AS event (i.e. patients with high vs. low PSI values), and examine if such classification is correlated with the overall survival of BRCA patients using Kaplan-Meier analysis. We found that five of cancer-specific AS events indeed can be used as predictor of tumor survival (log rank $p<0.05$), with two examples shown in figure 3.5D. The first example, WBP1 (WW domain binding protein 1), is a binding partner of WWOX tumor suppressor that is frequently mutated in breast cancer [108]. We found that increased retention of intron 3 in WBP1 is associated with poor prognosis (Figure 3.5D). The second example, GPR116, is an adhesion G-protein-coupled receptor that promotes breast cancer metastasis [109]. The inclusion of an alternative exon at end of GPR116 will generate a non-canonical isoform (isoform 2) with a different C-terminal cytoplasmic domain that may change its ability

to interact with downstream signaling. We found that the increased production of isoform 2 is associated with poor prognosis. Taken together, our data show that, for the first time, the splicing ratio of some human genes in cancers is associated with cancer survival, suggesting the possibility to use gene splicing as a new molecular signature to predict cancer prognosis.

3.3.8 Possible regulators of cancer-specific AS.

AS is generally controlled by various *trans*-acting splicing factors that specifically recognize *cis*-acting SREs in pre-mRNA. The level and activity of splicing factors usually vary in different cells, leading to the distinct AS patterns in corresponding cell types. Since splicing factors are often controlled by their co-expressed and functionally interacted proteins in different cellular signaling pathways [84], the splicing profile in certain cells may be significantly correlated with the expression of genes that play regulatory roles in AS. Therefore, given a large set of mRNA-seq data across different samples, a global analysis of correlations between AS patterns and expression of all genes may reveal regulatory relationships.

To explore possible regulatory mechanisms of cancer-specific AS, we systematically calculated the correlation between the PSI value of the 163 cancer-specific AS events and the expression of all detectable genes (Figure 3.6A). We found that the set of cancer-specific AS events are indeed significantly correlated with expression of many genes, among which are 304 genes highly correlated with more than 30 cancer-specific AS events. This set of genes are either positively or negatively correlated with the PSI values of many cancer-specific AS events, and thus may reflect potential regulatory pathways for the associated AS events.

We further conducted GO analysis on genes significantly associated with cancer-specific AS events, and found that the vast majority of them function in multiple pathways related to cell cycle regulation (Figure 3.6B). However, the functional category of RNA-binding is not

significantly enriched in the set of genes associated with cancer-specific AS. We were a little surprised by the small number of correlated splicing factors. However, splicing factor activities can also be regulated in the level of protein modification, and indeed we found that the phosphorylation of several splicing factors is cell cycle dependent (Dominguez at al, unpublished data). This result suggests that the activity of splicing factors may be controlled in a cell cycle dependent manner (e.g., through protein phosphorylation), and thus cell cycle proteins can indirectly affect splicing in tumor cells.

In addition, we analyzed protein-protein interaction among the genes correlated with cancer-specific AS using the STRING database. We found that 173 out of the 304 genes are highly connected with each other, and surprisingly the two largest clusters in the interaction network consist predominantly of genes that mediate the two major cell cycle checkpoints (i.e. checkpoint for G1-S and G2-M transition, Figure 3.6C). The genes in the largest cluster remarkably form a complete graph with 32 nodes, of which each connects with all others to function in kinetochore formation, cell division and mitosis (red cluster in Figure 3.6C). The second biggest group has 18 genes that are mostly involved in DNA replication. Even the smaller clusters have cell cycle related functions such as P53 signaling, M phase, DNA repair, and condensin complex. Such a high degree of correlation between cell cycle regulation and cancer-specific splicing has not been previously reported but has profound implication in how AS is controlled in multiple cancers.

3.3.9 Cancer-specific AS events common to 13 cancers

After the completion of three cancer comparisons, we then expanded our study to a more sophisticated list of cancers which includes BLCA (Bladder Urothelial Carcinoma), BRCA (Breast invasive carcinoma), COAD (Colon adenocarcinoma), HNSC (Head and Neck squamous

cell carcinoma), KICH (Kidney Chromophobe), KIRC (Kidney renal clear cell carcinoma), KIRP (Kidney renal papillary cell carcinoma), LIHC (Liver hepatocellular carcinoma), LUAD (Lung adenocarcinoma), LUSC (Lung squamous cell carcinoma), PRAD (prostate adenocarcinoma), THCA (Thyroid carcinoma) and UCEC (Uterine Corpus Endometrial Carcinoma). Same analysis pipeline was used but with a more stringent read coverage criteria. We required every splicing event to have at least 20 reads coverage and at least 5 reads supporting one of the two isoforms, which allow us to eliminate some false-positive splicing events. The number of cancer-specific AS events in each individual cancer are shown in figure 3.7A. Again, most of these events are specific to a single cancer type, with UCEC, COAD and LUSC having more altered AS events compare to the rest. We then compared every pair of cancers to see whether there are some cancers have similar AS profiles (Figure 3.7B). We found among all the cancer pairs, HNSC and LUSC have the highest similarity, which is consistent with the finding in [98] that these two types of cancer also have similar mutation profiles. Finally, we searched for AS events that were changed in multiple cancer types. Interestingly, there are 161 AS events altered among at least eight types of cancer (Figure 3.7C). The genes containing these events are enriched in regulation of cell motion, muscel cell differentiation, cytoskeleton and etc (Figure 3.7D). In addition, 23 out of the 161 AS events are altered among more than 10 types of cancer (Figure 3.7C): *ASAP2, SLK, NUMB, ADD3, EXOC7, NIN, RPS24, GAB1, CCDC50, LRRFIP2, MYH11, TNC, MBNL1, TUBA1A, HERC2P3, PRICKLE4, PILRB, RBM6, CDK10, ATG16L2, MYO15B, TUBA1A* and *PPCS*. Many of these genes are know to play an important role in some cancers. For example, a cancer-specific transcript of *MYH11* in AML was reported in many studies [110, 111]. In addition, cancer-specific isoform of *SLK*, *NUMB* and *ADD3* were also found in non-small cell lung cancer [112]. All these studies identified the

cancer-specific AS in one type of cancer. In our study, we used high-throughtput sequencing data to identify these AS event and confirm they are not only specific to one type of cancer, but also change among a lot of cancers.

3.4 Discussion

Here we performed a systematic identification and analysis of cancer-specific AS events using thousands of patient samples from TCGA data. To increase the statistical power of our analyses, we selected three types of cancers that have a relatively large number of paired normal controls. These tissue types are sufficiently different to enable us to filter out tissue-specific splicing, as most identified AS events are altered in only a single type of cancer (Figure 3.1B). The AS events significantly altered in all three cancer types include many genes whose splicing was known to play critical roles in cancer development, such as the CD44 [88], NUMB [113, 114], and FN1 [115]. These genes probably represent a core set of cancer-specific AS events that affect key pathways in cancer progression.

On the other hand, several AS events that have well-known roles in cancer development were not identified by our procedure, probably due to the high stringency used in our filters. For example, our set of cancer-specific AS does not include Bcl-x, whose splicing is known to control cell apoptosis in multiple tumors [89] and can be used as a potential therapeutic target [116, 117]. However the ΔPSI of Bcl-x is not large enough to pass the thresholds in our pipeline, and we expect that additional cancer-specific AS events can be identified when the criteria are relaxed. We also require the AS events to be detected in ~30% of normal liver samples (10 out of 36, see table 1), which may cause some uncommon events to be omitted in our pipeline.

Although most cancer-specific AS events are involved in cellular pathways critical to cell growth and migration, they may not directly drive the initial stages of tumorigenesis, as we could

not detect obvious mutations near the splice sites of these alternatively spliced exons. Instead we speculate that they are more likely to be a result of mis-regulated splicing factors that potentially change splicing of many pre-mRNA targets. Consistently, several splicing factors were found to be mutated or significantly changed in expression between cancers and normal tissue, including SRSF1, QK1, RBM4, RBM5/6/10, and hnRNP A2 [46, 47, 49, 113, 114]. These results imply that cancer-specific AS events will be more useful as cancer biomarkers, whereas the splicing factors may better serve as potential therapeutic targets to restore mis-regulated splicing in cancer.

To study potential regulatory mechanisms for the cancer-specific AS events, we used an association study to identify genes whose expression is correlated with these events across thousands of tumor and normal samples. This large dataset size enables a statistically reliable identification of genes that directly or indirectly regulate AS. Such analyses only identified a small number of putative splicing factors including hnRNP L and snRPA1 (Figure 3.6C, marked at the bottom). We speculate that this is due to the large heterogeneity among tumor samples, as the known cancer-related splicing factors are found to be altered in only a subset of tumor samples. Remarkably, the majority of genes whose expression is associated with cancer-specific AS events are those involved in cell cycle regulation, revealing an unknown link between cell cycle and splicing regulation. An unbiased clustering of these associated genes recapitulated two major cell cycle checkpoints (i.e., G1 to S and G2 to M transition) and several main control pathways for cell cycle progression (e.g., DNA repair and P53 signaling). Although the reason of such high correlation is not clear, there are several interesting implications and predictions. For example, this result may suggest that genes controlling cell cycle progression also play a central regulatory role in pre-mRNA splicing and processing.

Since cancer cells undergo fast growth and division compared to normal cells, there may be an increasing pressure for cancer cells to transcribe and splice certain genes at a high rate in some cell cycle stages. Because most introns are spliced co-transcriptionally, the increased transcription rate may directly affect AS of a certain set of genes in cancer. There may also be epigenetic factors that bridge the regulation of cell cycle with alternative splicing. A careful examination of this link requires integration of the changes in various epigenetic markers and transcription factors with splicing alteration, which will be an important direction for future investigation. Another interesting implication is that AS may be temporally regulated during the cell cycle. Although periodic gene transcription during cell cycle is well documented [118], there are limited reports on temporal regulation of splicing at different cell cycle stages. Our result implies that such a regulation mode is likely to exist and may even be a major mechanism responsible for cancer-specific AS.

In summary, this study generated a common set of cancer-specific AS events across different cancer types, which can be used as novel cancer biomarkers. We provided a detailed picture of unique features for these AS events and mechanistic insights on how splicing is mis-regulated in cancer. Because dysregulation of splicing in cancer can often serve as a cancer progression indicator, the identification of a core set of cancer-specific AS events will likely help early cancer detection and thus improve the chance of cure. Finally, this relative small set of AS events will facilitate direct discovery of key regulators that are responsible for splicing dysregulation in cancers and thus can potentially be used as new therapeutic targets.

3.5 Materials and Methods

3.5.1 Data acquisition and sequence processing.

Pair-ended RNA-seq data were acquired from the TCGA consortium, with all reads being pair-ended (length: 50, 48, and 48 for breast, lung, and liver cancer respectively). Each sample has an average of >150 million reads. The reads were aligned to the human genome version hg19 with MapSplice V2.0 [119], and the gene expression values were estimated using the RSEM pipeline [120] and normalized to the upper quartile of all expressed genes [121].

To analyze AS events on a genomic scale, we used the MISO event-centric pipeline [122] with the hg19 v2.0 annotation to calculate the inclusion ratio of all annotated AS isoforms (http://genes.mit.edu/burgelab/miso/annotations/ver2/miso_annotations_hg19_v2.zip). Further analyses were carried out for four major modes of AS: skipped exon (SE), retained intron (RI), alternative 3' splice site (A3SS) and alternative 5' splice site (A5SS). Based on the coverage of different splicing isoforms, each AS event was assigned with a PSI (Percent Spliced In) value ranging from 0 to 1. To qualify as a valid AS events, we require that both isoforms are detectable in at least 10 normal samples and 10 tumor samples for each cancer type.

3.5.2 Determination of AS events shared between cancer types.

To examine the statistical significance for the number of AS events that are in common between different cancers, 1000 simulated datasets were generated by randomly selecting a control set of AS events with matched size in each cancer type (number of AS events for lung, liver and breast cancer respectively: SE: 3111, 1308, 1804; RI: 378, 109, 201; A3SS: 614, 317,

331; A5SS: 533, 244, 290). We then computed the number of events that were common across multiple cancer types. In each AS mode (SE, RI, A3SS and A5SS), we generated 1000 simulated datasets and calculate the mean overlaps between different cancers, which were then compared to the overlaps of real data using rank test.

3.5.3 Analyses of protein-protein-interaction among cancer-specific AS events.

The genes containing cancer-specific AS events (or genes whose expression is associated with cancer-specific AS events) were obtained and submitted to the STRING database [123, 124] (http://string-db.org/) for protein-protein interactions (PPI) analysis. We used the combined score of 0.4 as a cutoff and included five white nodes for network continuity. We used Cytoscape [125] to visualize the PPI network and the MCODE algorithm [100] to identify highly connected clusters within the network. See supplementary Table S3.2 and S3.3 for detailed parameters.

3.5.4 Calculation of evolutionary score.

Sequence evolutionary score was downloaded from UCSC phastCons100 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons100way/) [126]. Based on multiple sequence alignments of 100 vertebrate species, each nucleotide was given an evolutionary conservation score ranging from 0 to 1. Highly conserved regions are assigned with a higher score. PhastCons estimates the probability that each nucleotide belongs to a conserved element based on multiple alignments using a hidden Markov model. For each SE event, we extracted sequences from different regions near the alternative exon to calculate average conservation score in a sliding window of 8 nt across all cancer-specific SE events and control events.

3.5.5 Motif enrichment analysis.

To analyze the enriched sequence motifs near the splice sites of the 163 cancer-specific AS events, we first obtained nucleotide sequences from three splicing regulatory regions: upstream intron (300 nt), exon and downstream intron (300 nt) as shown in Figure S3.3. When obtaining the sequences, we excluded the first 25 nucleotides right upstream of the skipped exon, the first 10 nucleotides right downstream of the skipped exon and the first and the last two nucleotides within the exon. We then calculated the frequency and Z-score of all 5-nt "words" near the alternative exons of the 163 sequences in three regulatory regions using methods described in [127]. All 5-mers with Z-score larger than 2.5 were then clustered by sequence similarity and multiply aligned by using CLUSTALW to identify candidate motifs. At a cutoff dissimilarity score of 2.65, 2.7 and 2.7, we obtained 5, 7 and 5 clusters of at least four sequences in each cluster for upstream intron, exon and downstream intron respectively. Finally, we plotted the consensus sequence for each cluster for all three regulatory regions (Figure S 3.3).

3.5.6 Principal Component analysis (PCA).

PCA is a data analysis technique commonly applied for dimension reduction, exploratory analysis and feature selection. PSI values of the 163 cancer-specific AS events were used to form the data vector for PCA. For each cancer type, the PSI vectors across all normal and tumor samples were then combined and used as the input data matrix to perform PCA using the prcomp() function in R. We also conducted PCA by combining the PSI values across all samples from three cancer types. The distributions of normal and cancer samples across the first two components were plotted.

3.5.7 Survival analysis for breast cancer patients.

We obtained the overall survival data of breast cancer patients from the UCSC Cancer Browser (727 patients). If a patient deceased (event happened), the "days_to_death" was used as the time variable; if a patient is still living, the time variable is the maximum of "days_to_last_known_alive" and "days_to_last_followup". The patient samples were split into two groups according to the top or bottom quartile of PSI values for each of the 163 cancer-specific events. The resulted two patient groups are compared for their probability of survival using a Kaplan-Meier survival plot and the log rank *P* values are calculated. This process was repeated for every cancer-specific event.

3.5.8 Correlation between gene expression and AS.

Correlations between genes and AS events were calculated using two matrices. The first matrix consists of the PSI values of 163 cancer-specific AS events across 1319 cancer and normal samples. Another matrix contains the expression level of every gene across 1319 samples. We computed the spearman rank correlation, ρ (rho), between every two vectors from the two matrices using cor.test() in R. Each pair with $|\rho| >= 0.4$ and $p <= 0.005$ was considered as a highly correlated event-gene pair. We considered genes that are highly correlated with more than 30 cancer-specific AS events as potential regulators through a direct or indirect regulation. We then used STRING database [123, 124] (http://string-db.org/) to extract PPIs between these potential regulators (304 genes), and Cytoscape and MCODE to visualize and cluster the interaction networks.

Table 3.1 Summary of cancer dataset

Tumor Type	No. of normal samples	No. of tumor samples	Total mapped Reads	No. of AS events detected	No. of AS events changed in cancer vs. normal
BRCA (breast cancer)	91	727	1.3E+11	65,152	2,626
LIHC (liver cancer)	36	74	1.6E+10	70,342	1,978
LUSC (lung cancer)	43	348	6.4E+10	70,637	4,636

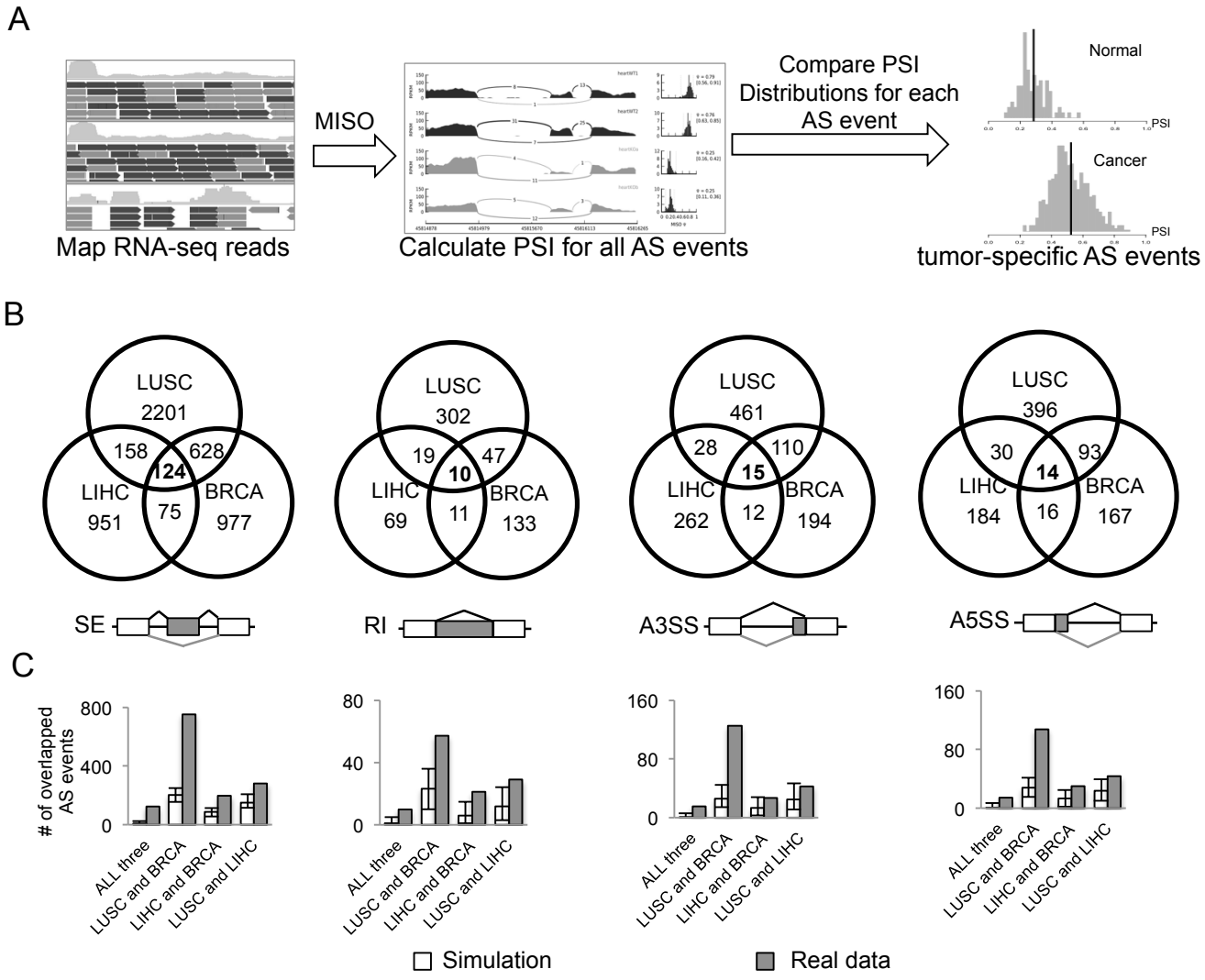

Figure 3.1 Identification of AS events altered in cancers.

(A) Analysis pipeline to identify AS events that are differentially spliced in cancer vs. normal tissues. (B) Venn diagram of differentially spliced AS events that are specific to one or multiple types of cancers in four different AS modes (SE, RI, A3SS and A5SS). (C) Numbers of AS events overlapped in the simulated dataset (white boxes) and the real dataset (gray boxes). We randomly selected the matched number of AS events from all the detectable AS events for each cancer and then calculated the overlaps. This procedure was repeated 1000 times, and the mean and range of numbers of overlapped events are shown.

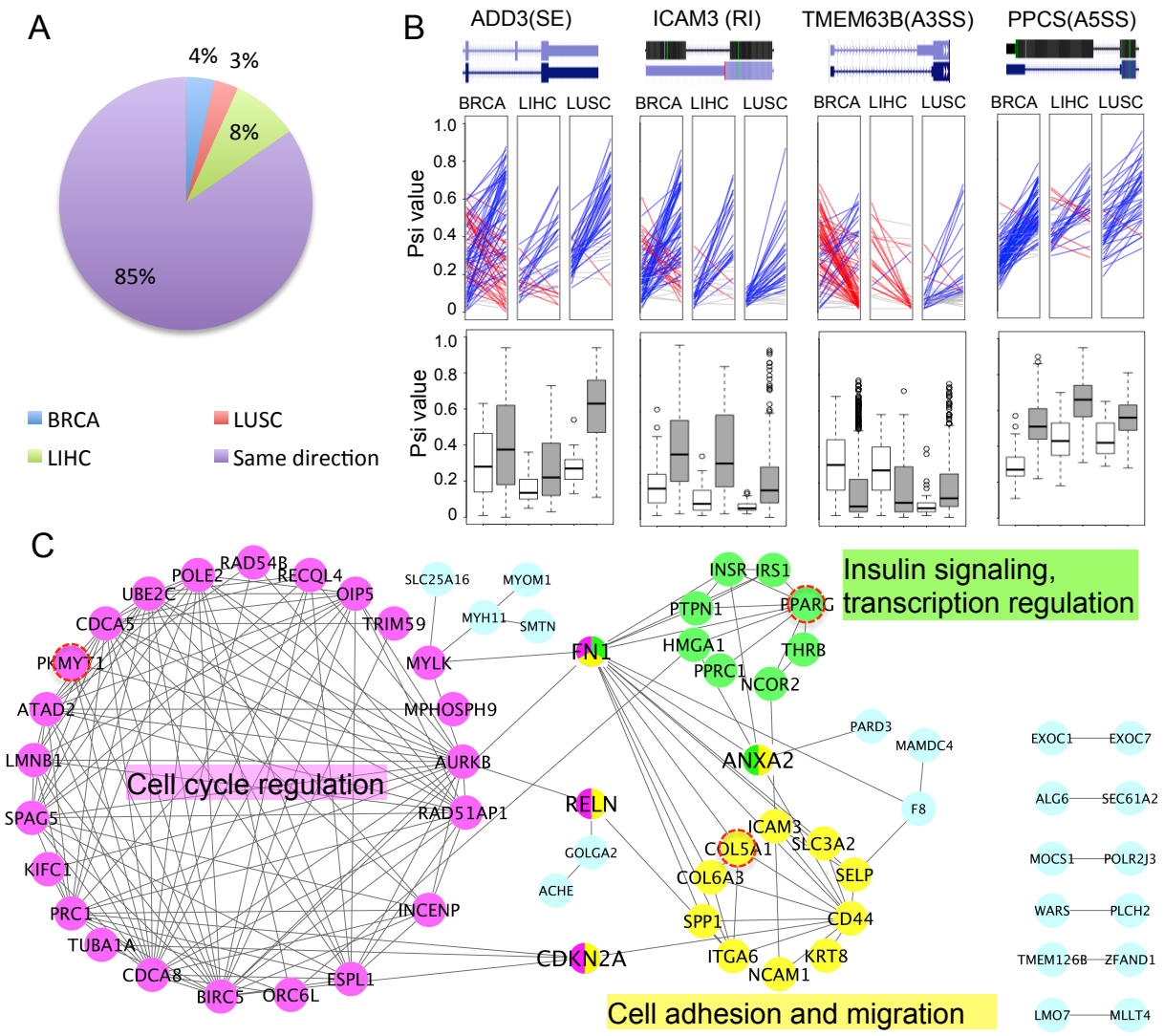

Figure 3.2 Examples of cancer-specific AS events.

(A) Changes of splicing across different cancer types. For all cancer-specific AS events, the differences of mean PSI values between cancer and normal samples were calculated. We plotted the percent of AS events with PSI changed in same direction among all cancers (purple) or in different directions for one specific type of cancer (BRCA, LUSC or LIHC). (B) Change of splicing in selected examples of cancer-specific AS events. The PSI values of paired samples are marked on the left (normal) and right (tumor) panel in each ladder plot with a colored line linking two PSI values. Blue lines represent AS events with increased PSIs in tumor, whereas red lines represent events with decreased PSIs and grey lines represent events with negligible change of PSI (<=0.05). Box plots in the bottom are comparisons between all normal samples (white boxes) and all tumor samples (grey boxes). (C) PPI networks of genes containing cancer-specific AS events. The networks have 3 highly connected clusters defined by MCODE (color coded in pink, yellow and green). The hub proteins interacting with multiple clusters were coded with multiple colors. Three genes that are also frequently mutated in tumors were marked by red circles. The most enriched function/GO-term was labeled next to each cluster.

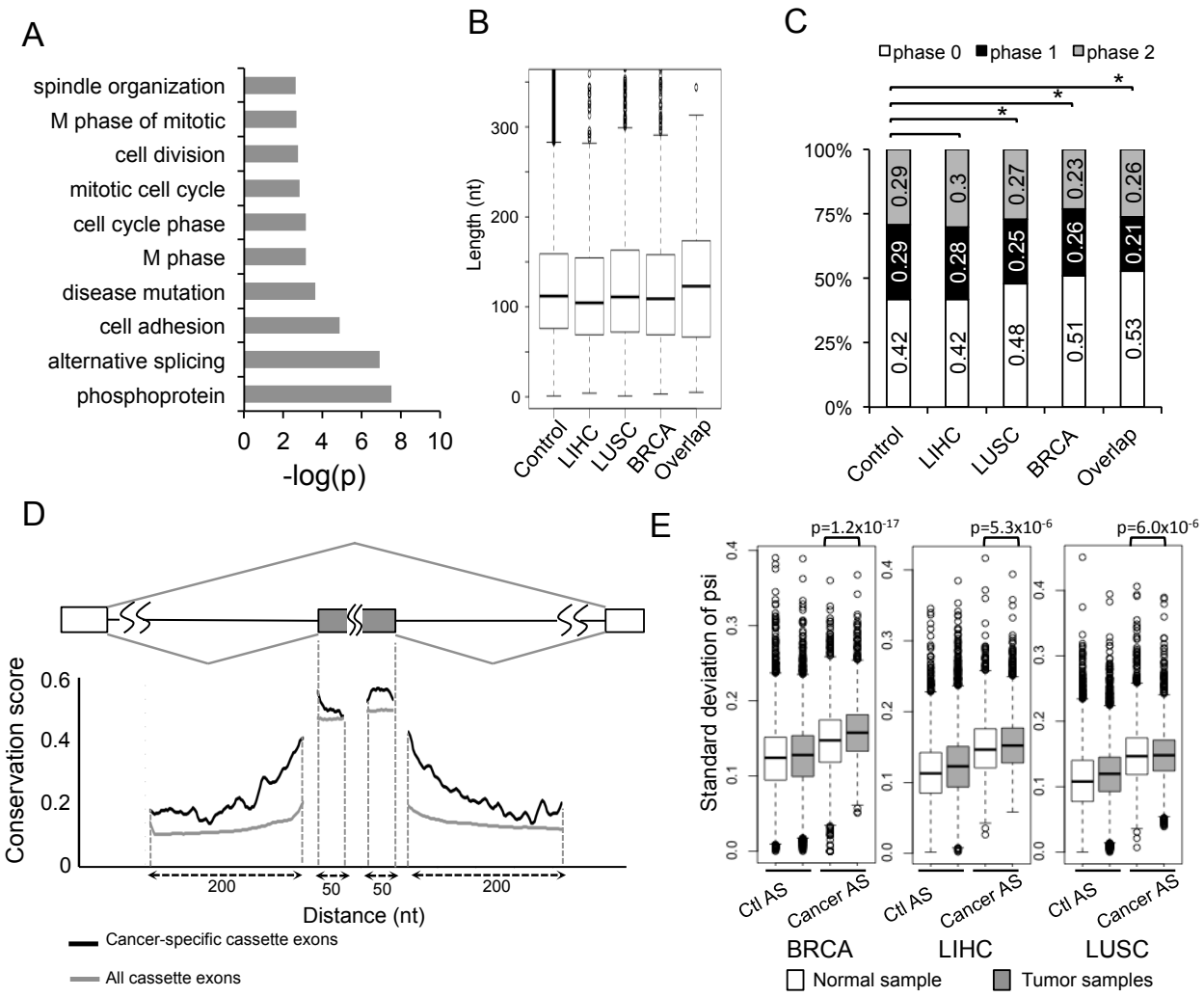

Figure 3.3 Molecular features of AS events changed in cancers.

(A) The genes containing 163 cancer-specific AS events were analyzed by gene ontology, and the significantly enriched ($p<0.005$) GO terms are plotted. (B) Box plots of the length of alterative exons in the SE events from the MISO database, the SE events that were changed in LIHC, LUSC, BRCA, and the SE events altered in all cancers (from the left to the right). Asterisks indicate significant increase of phase 0 exon ($P<0.05$ by Fisher's exact test). (C) Percent of skipped exons in each exon phase. Exons are classified into phase 0, 1, and 2 depending on the reminders when dividing their length by 3. Phase 0 (white boxes): events without frame-shift; Phase1 and 2 (black and gray boxes): events with frame-shift. The order is same as (B). (D) Sequence conservation near the cancer-specific skipped exons and all skipped exons. Black line represents average conservation score from the 124 cancer-specific SE events; grey line represents average conservation score from all the SE events in MISO database (control). (E) We compared the distribution of PSI standard deviation between control AS (left) and AS events that change in each cancer (right). We also compared those between normal (white box) and tumor samples (grey box).

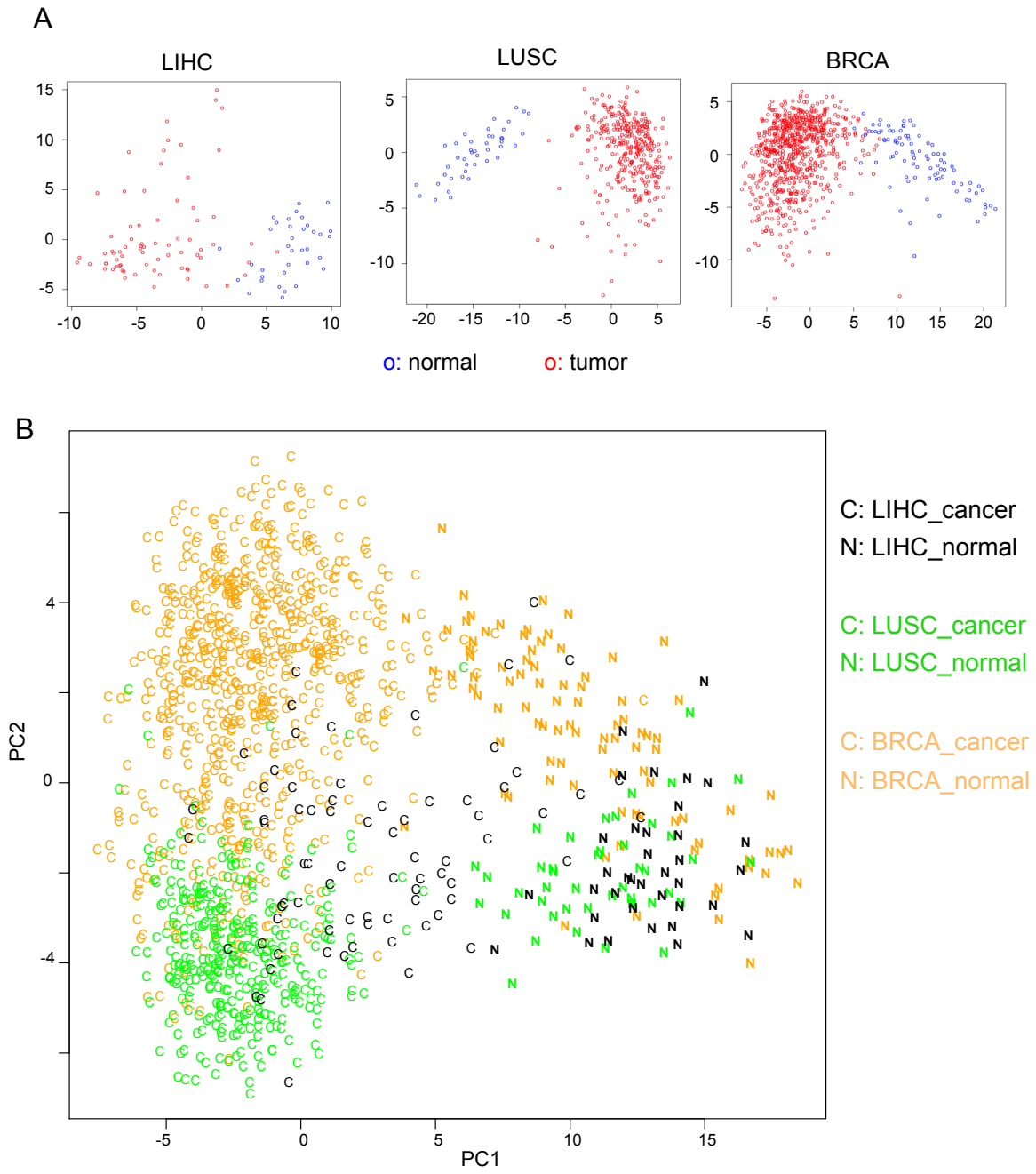

Figure 3.4 PCA analysis using cancer-specific AS events.

(A) PCA analysis of the 110 liver tissue samples (LIHC: 36 normal and 74 tumor), 391 lung tissue samples (LUSC: 43 normal and 348 tumor) and 818 breast tissue (BRCA: 91 normal and 727 tumor) using the 163 cancer-specific AS events. Tumor samples are in red circles and normal samples are in blue circles. (B) PCA analysis of samples from all three tissues using the 163 cancer-specific AS events. Samples were color-coded as its origin tissue. The cancer samples are labeled as "C" and the normal samples are labeled as "N".

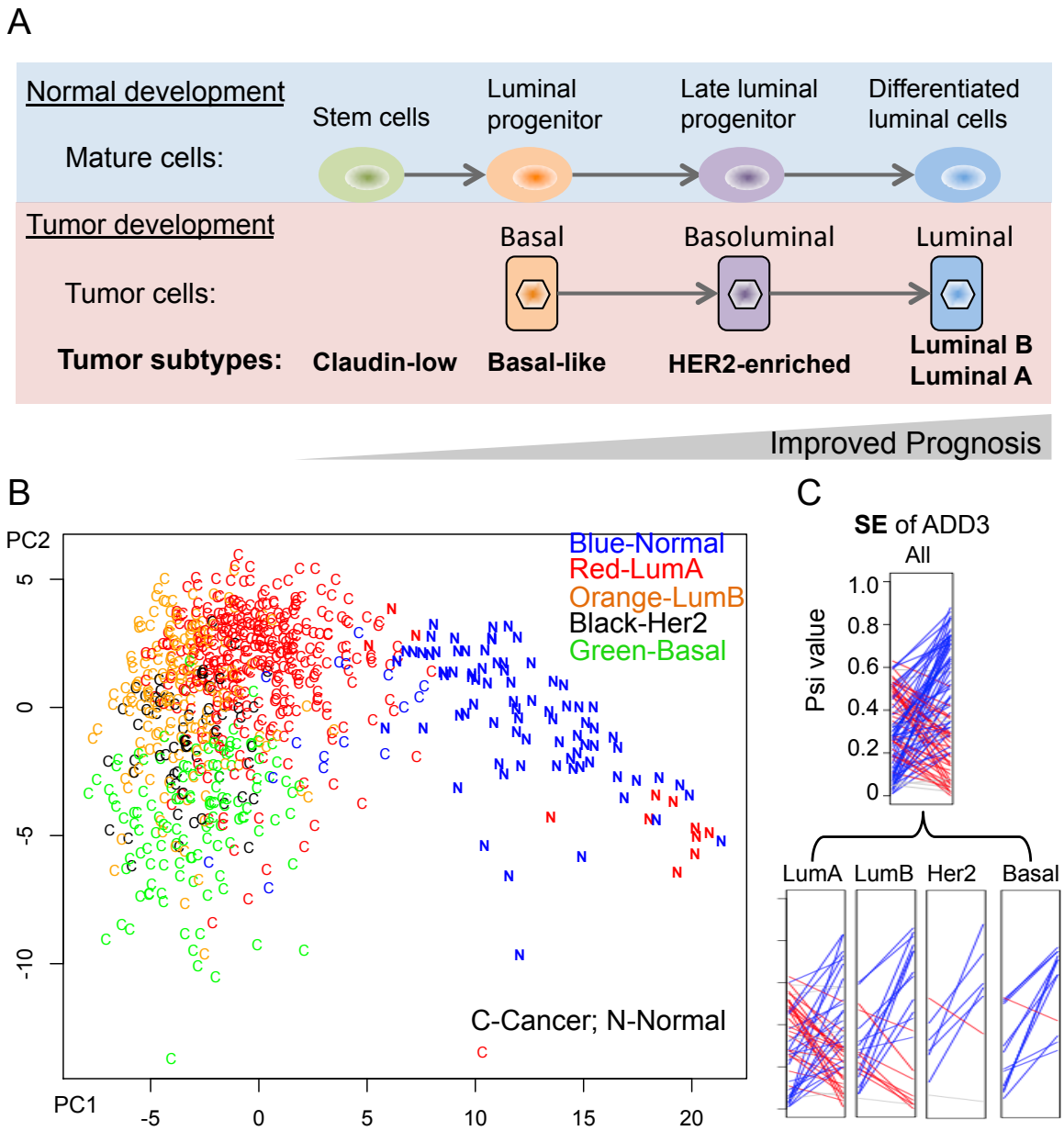

Figure 3.5 Using cancer-specific AS events to separate breast cancer subtypes.

(A) An overview of four different breast cancer subtypes in tumor developments. (B) PCA analysis of the 818 breast tissue using the 163 cancer-specific AS events. Different BRCA samples were labeled according to each subtype as classified by PAM50. (C) An example SE event where the different BRCA subtypes have different splicing patterns. The ladder plots were generated as described in figure 3.2B. (D) Two examples of cancer-specific AS events whose PSI value can be used to predict survival of breast cancer patients.

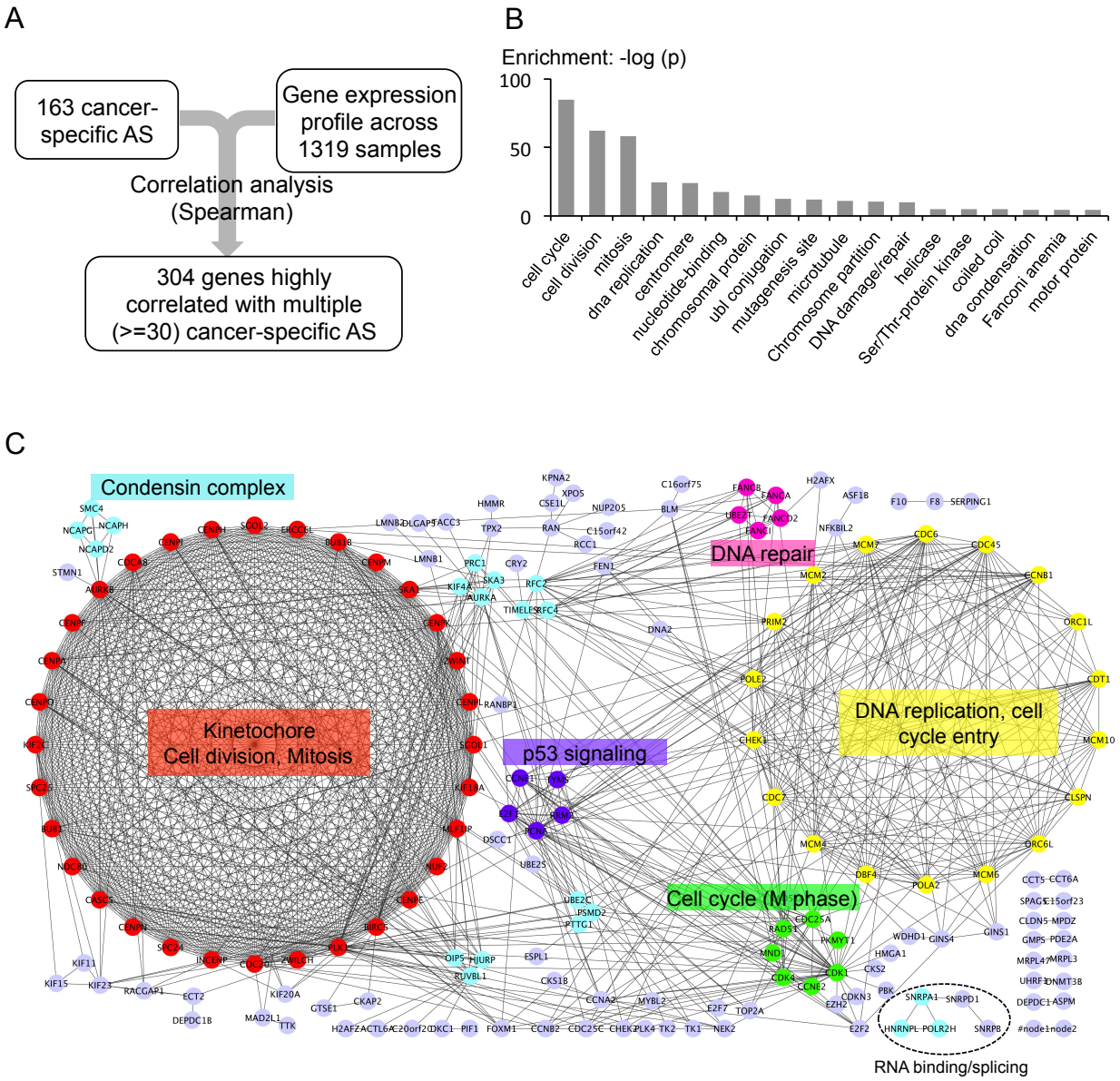

Figure 3.6 Genes associated with cancer-specific AS events.

(A) Flow chart of identifying possible regulators of cancer-specific AS. (B) Gene ontology analysis of the 304 genes that are highly correlated with the 163 tumor AS events. (Spearman rank correlation >=0.4 across 1319 samples, p-value<0.005, and correlated with more than 30 out of the 163 tumor AS). (C) PPI networks of genes that are highly correlated with the 163 cancer-specific AS events. Color-coded proteins are clustered by MCODE. Light purple-colored nodes were proteins that were not clustered into any group by MCODE. The most enriched function/GO-term was labeled next to each cluster. The genes involved in RNA binding and splicing regulation are also indicated at the bottom.

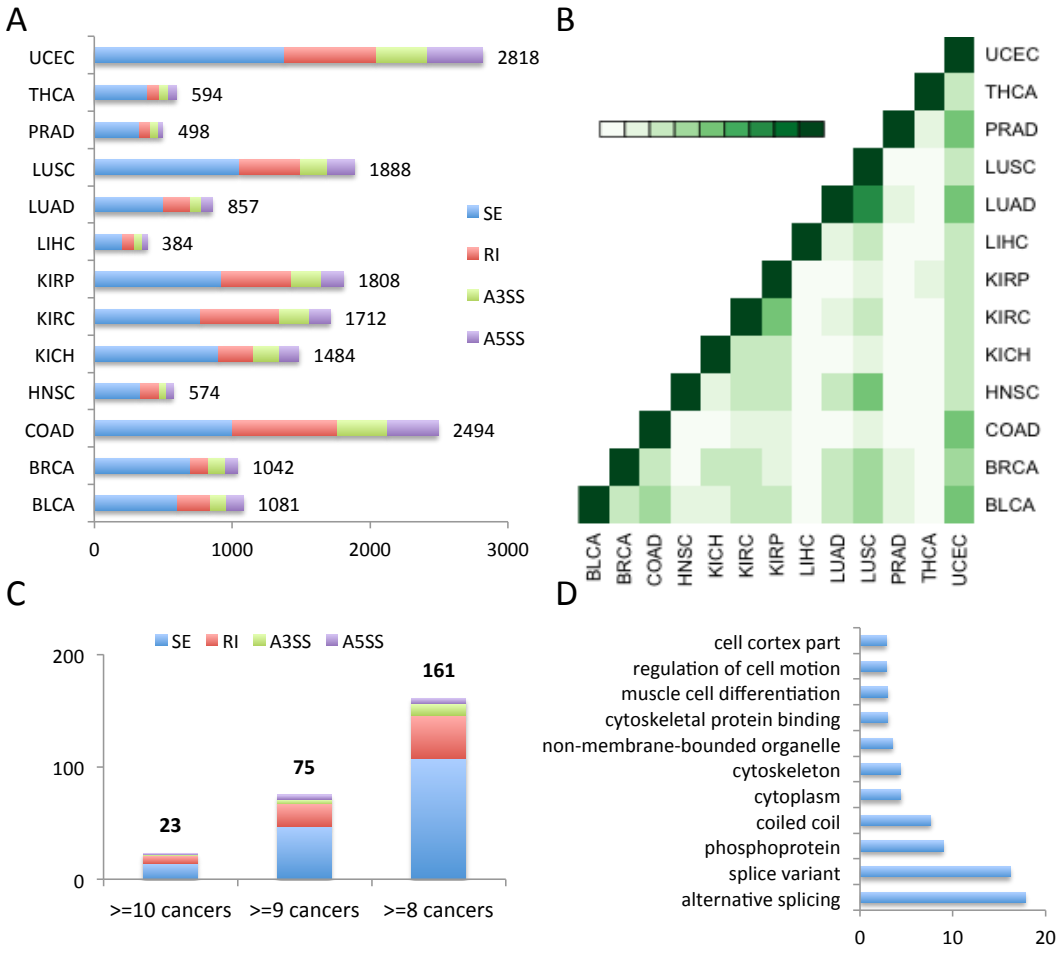

Figure 3.7 Cancer-specific AS events among 13 cancer types.

(A) Cancer-specific AS events in each individual cancer type and in each splicing mode (as color specified). (B) Heatmap of similarity between every paired cancer. The similarity was measured as the percentage of AS events altered in both cancers (normalized to the leftmost cancer type in each row). (C) Number of AS events altered among at least 10, 9 and 8 types of cancer. (D) Gene ontology analysis of the genes contacting cancer-specific AS events that are altered among at least eight types of cancer (161 AS events).

3.6 Supplementary Material

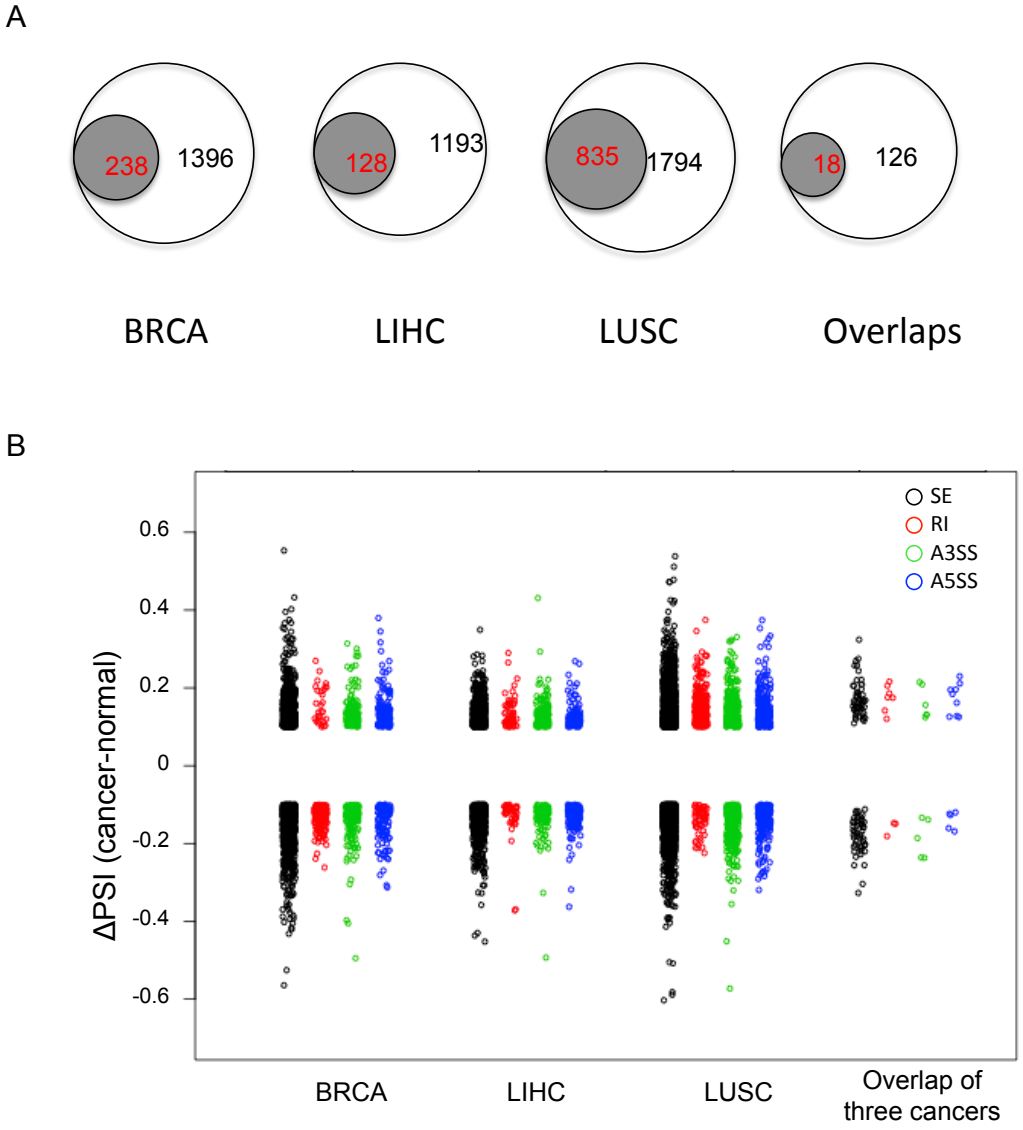

Figure S 3.1 The percentage of genes change in both expression level and splicing and the splicing isoform change in four AS modes.

(A) For the genes containing cancer-specific AS (white circle) in each cancer type, a small fraction also showed significant changes in expression level between tumor and normal samples (grey circle). We used the following threshold for expression changes: the expression levels of each gene have to change by at least two fold between cancers *vs* normal with p-value <= 0.005. (B) The change of PSI values between paired tumor and normal samples were plotted in each cancer type and in all three cancers separately. ΔPSI was calculated as the PSI value in cancer sample minus the PSI value in paired normal control. Each dot represents a paired of cancer and normal sample.

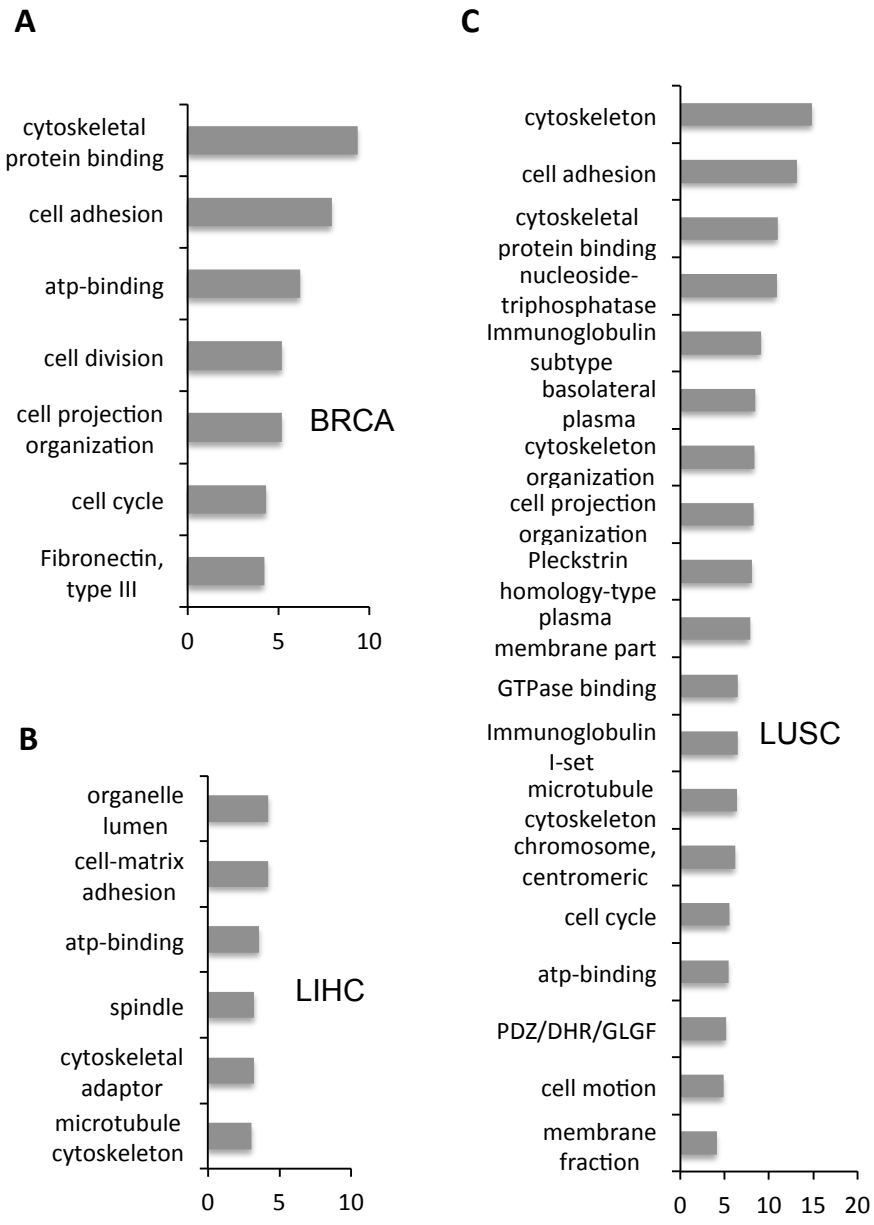

Figure S 3.2 Gene ontology analysis of AS events altered in three cancer types: BRCA (A), LIHC (b) and LUSC (C).

We obtained the list of genes containing AS events that change significantly in their PSI values between tumor and normal samples in breast, liver and lung cancer datasets, and listed the highly enriched GO terms with p-value less than 0.005 using DAVID gene ontology tool. The x-axis is the –log (P) of the enriched GO term.

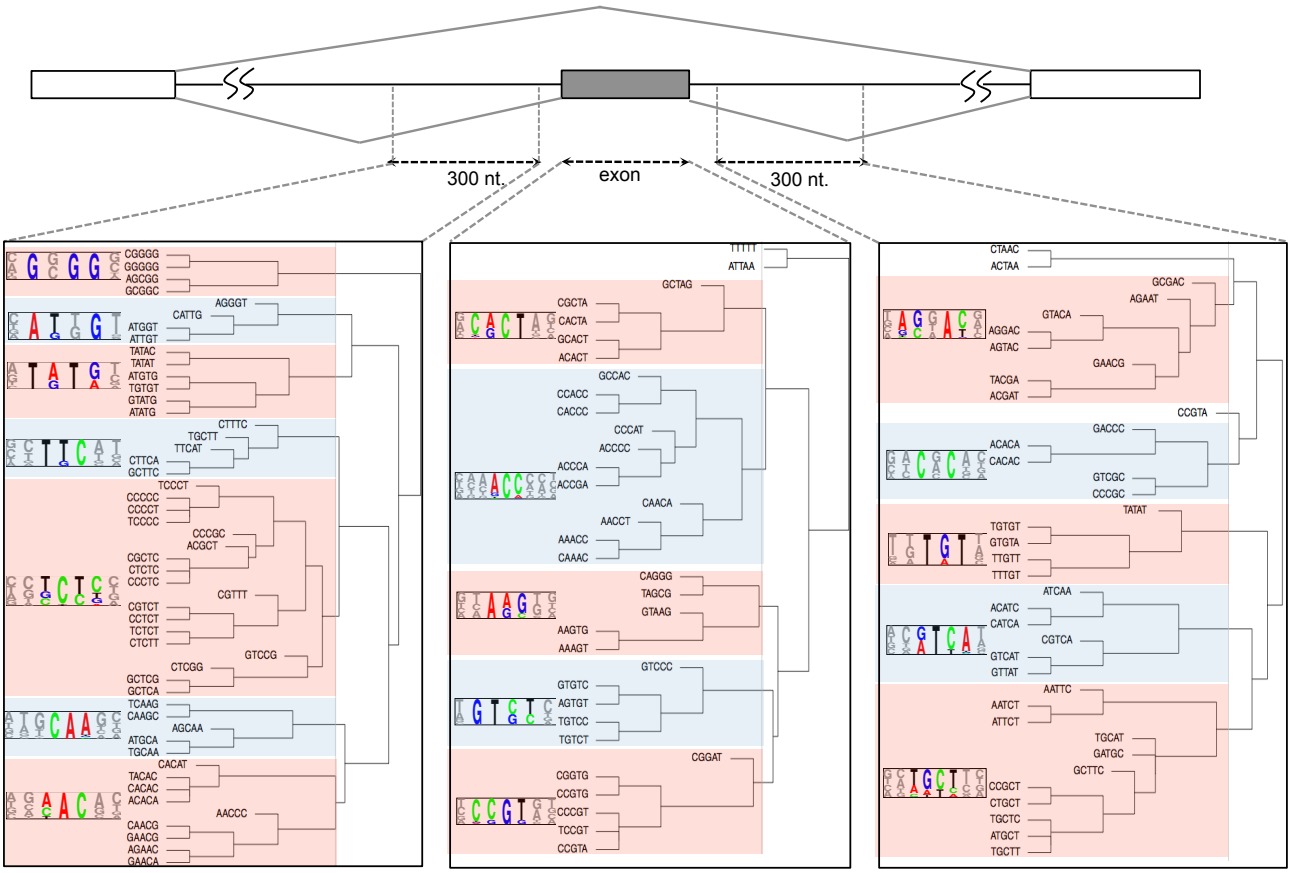

Figure S 3.3 Enriched motifs near cancer-specific skipped exons.

The pentamers significantly enriched in each pre-mRNA region were identified and clustered into different groups according sequence similarity. The consensus motif in each group was represented with pictogram. Upstream intron: 300 nucleotides upstream of the skipped exon. The enriched Exon region: the whole exon sequences were used except the first and the last two nucleotides. Downstream intron: 300 nucleotides downstream of the skipped exon. The regions overlapping with splice sites (the first 10 nt and the last 25 nt of intorns) were excluded to avoid strong splicing signals.

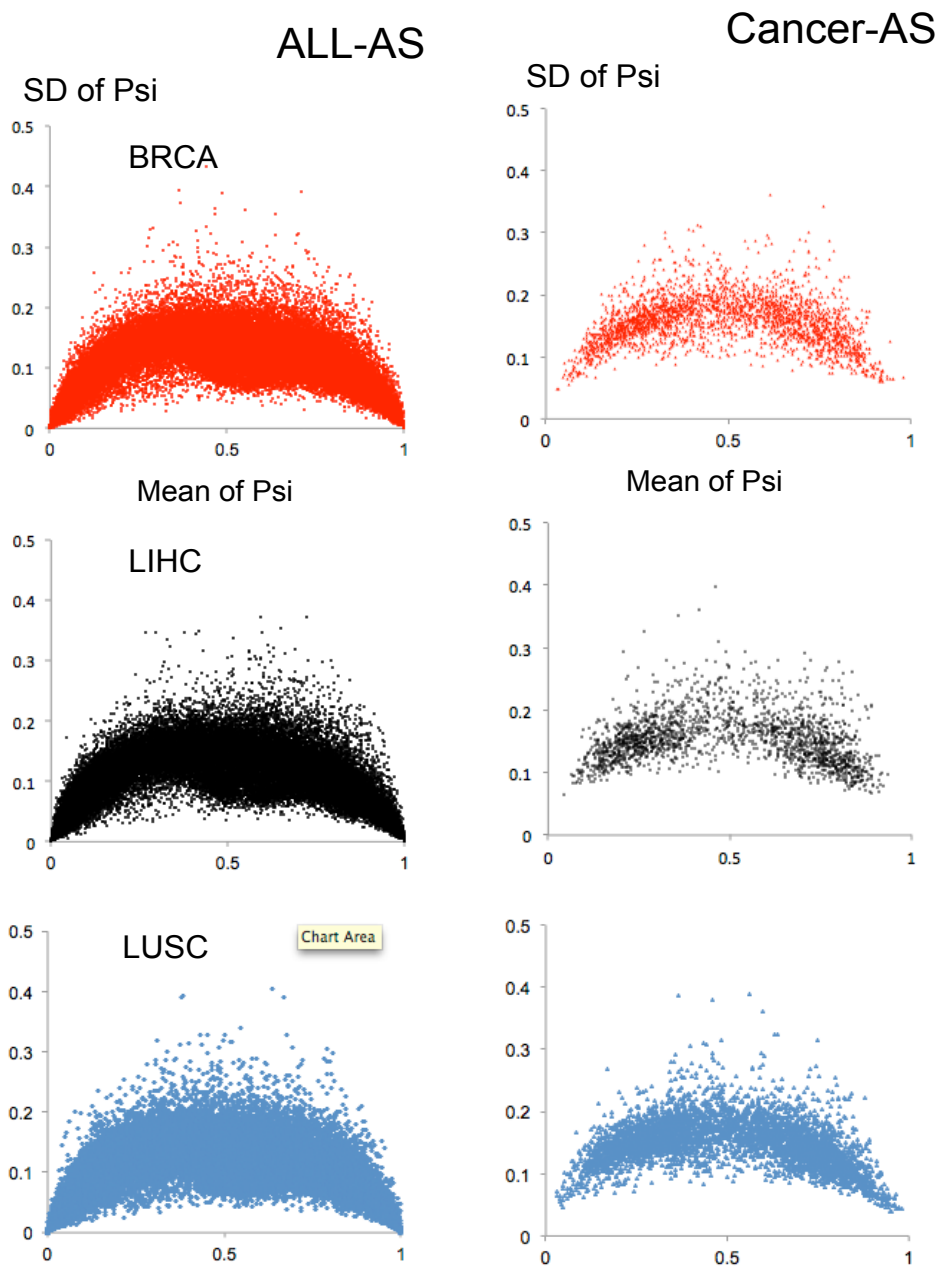

Figure S 3.4 Scatter plots of the standard deviation of PSI vs. mean of PSI.

For each AS event, the PSI values and the standard deviation of PSI were plotted among all samples in breast, liver and lung cancer datasets. The distribution of all AS events (left) were compared to the AS events that significantly change between tumor and normal (right), and the control set was selected from all AS events with matched distribution of PSI values.

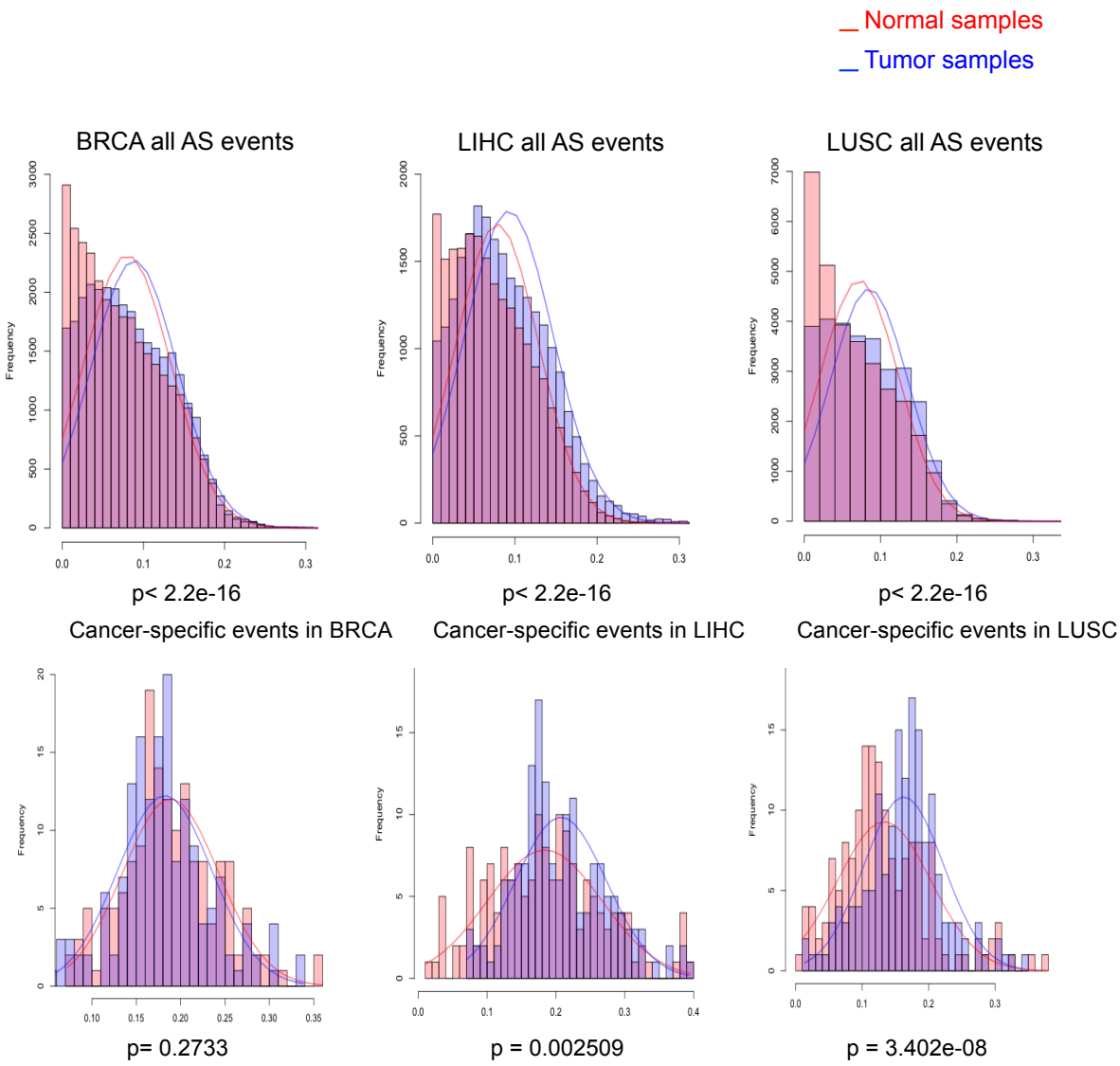

Figure S 3.5. Histograms of the standard deviation of PSI for all AS events (top) or for 163 cancer-specific AS events (bottom).

The normal and tumor samples (in BRCA, LIHC and LUSC cancer) are plotted in different colors, and we found that for both types of AS events, the SD of PSI is larger (right-skewed) in tumor samples, suggesting that splicing in tumors are more dispersed. See also Fig 3E.

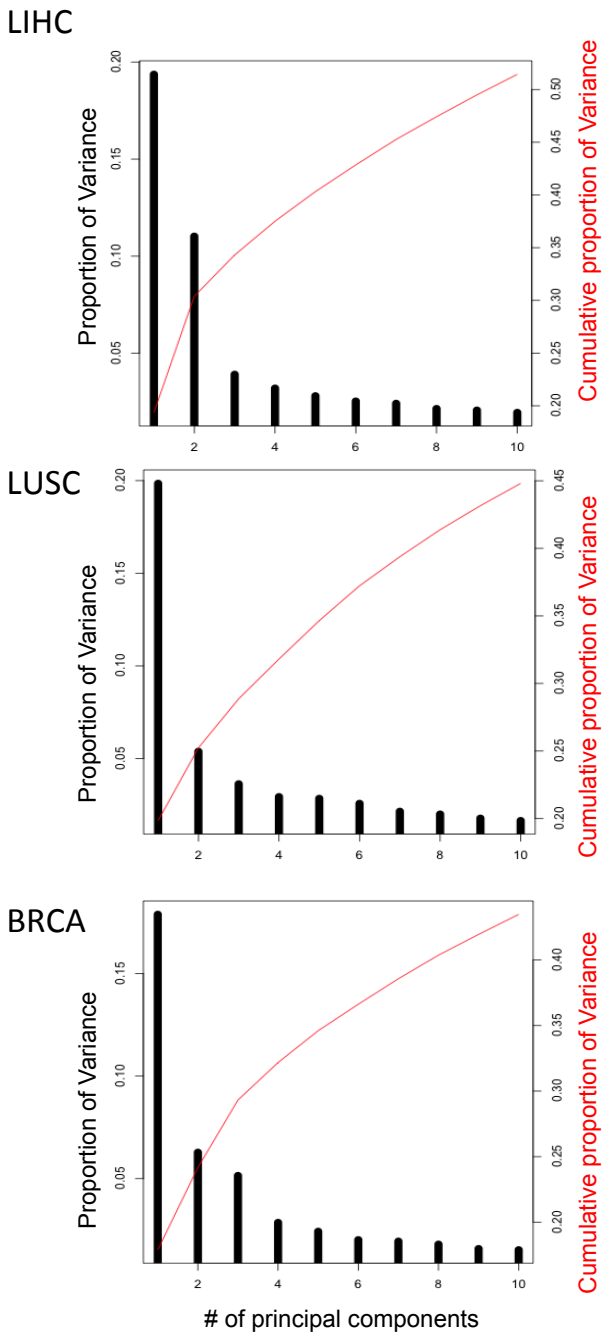

Figure S 3.6. The proportion of variance explained by the first ten principal components.

Variances of the first 10 PCA compoents are shown in black boxes and the cumulative proportions of them are shown in red lines.

Table S 3.1 162 Cancer-specific AS events and their average PSI values in three types of normal and tumor samples

	Mean of PSI							
SE	**BRCA normal**	**BRCA tumor**	**LUSC normal**	**LUSC tumor**	**LIHC normal**	**LIHC tumor**	**ENSG ID**	**Gene Name**
chr10:34663802:34663930:-@chr10:34661426:34661464:-@chr10:34648999:34649187:-	0.56	0.43	0.59	0.42	0.63	0.49	ENSG00000148498	PARD3
chr17:76210761:76210870:+@chr17:76218909:76219073:+@chr17:76219546:76221716:+	0.21	0.03	0.16	0.02	0.30	0.06	ENSG00000089685	BIRC5
chr8:82630417:82630459:-@chr8:82629484:82629523:-@chr8:82627222:82627349:-	0.75	0.62	0.76	0.59	0.46	0.57	ENSG00000104231	ZFAND1
chr6:45881965:45882146:-@chr6:45881515:45881549:-@chr6:45866190:45870992:-	0.07	0.25	0.02	0.24	0.45	0.35	ENSG00000112782	CLIC5
chr17:5329291:5329402:+@chr17:5329556:5329619:+@chr17:5335862:5336340:+	0.63	0.49	0.72	0.52	0.68	0.52	ENSG00000129197	RPAIN
chr8:26721604:26722922:-@chr8:26627798:26628183:-@chr8:26613913:26614296:-	0.79	0.59	0.80	0.61	0.95	0.76	ENSG00000120907	ADRA1A
chr6:39877579:39877699:-@chr6:39876816:39876878:-@chr6:39872034:39874893:-	0.80	0.67	0.82	0.72	0.84	0.71	ENSG00000124615	MOCS1
chr15:60688350:60688626:-@chr15:60685237:60685639:-@chr15:60678227:60678285:-	0.45	0.29	0.34	0.18	0.37	0.23	ENSG00000182718	ANXA2
chr9:131036129:131036251:-@chr9:131035064:131035144:-@chr9:131030699:131030803:-	0.67	0.36	0.72	0.49	0.65	0.51	ENSG00000167110	GOLGA2
chr11:64850836:64850871:-@chr11:64846825:64847259:-@chr11:64835960:64836073:-	0.63	0.83	0.65	0.88	0.54	0.74	ENSG00000146670	CDCA5
chr5:126112853:126112944:+@chr5:126113053:126113559:+@chr5:126140468:126140624:+	0.63	0.78	0.72	0.84	0.68	0.78	ENSG00000113368	LMNB1
chr11:65307716:65307853:-@chr11:65307484:65307624:-@chr11:65307191:65307352:-	0.40	0.28	0.43	0.24	0.50	0.35	ENSG00000168056	LTBP3
chr6:131199244:131199390:-@chr6:131193511:131193678:-@chr6:131191468:131191521:-	0.06	0.23	0.09	0.28	0.41	0.30	ENSG00000079819	EPB41L2
chr18:3131373:3131494:-@chr18:3129230:3129517:-@chr18:3126699:3126895:-	0.08	0.20	0.11	0.58	0.20	0.42	ENSG00000101605	MYOM1
chr3:58817412:58817615:-@chr3:58792121:58792182:-@chr3:58739496:58739590:-	0.49	0.63	0.55	0.73	0.51	0.65	ENSG00000163689	C3orf67
chr4:38869354:38869455:+@chr4:38870019:38870167:+@chr4:38879692:38880047:+	0.22	0.33	0.22	0.35	0.23	0.36	ENSG00000197712	FAM114A1
chr11:85339622:85339732:+@chr11:85342189:85342360:+@chr11:85342731:85342852:+	0.64	0.40	0.69	0.56	0.69	0.59	ENSG00000171204	TMEM126B
chr6:46823711:46823795:-@chr6:46822452:46822518:-@chr6:46820242:46821808:-	0.23	0.36	0.07	0.26	0.08	0.32	ENSG00000069122	GPR116
chr3:194134488:194134568:-@chr3:194132928:194133017:-@chr3:194123403:194126845:-	0.79	0.60	0.82	0.67	0.68	0.57	ENSG00000133657	ATP13A3
chr10:111890121:111890244:+@chr10:111892063:111892158:+@chr10:111893084:111895323:+	0.30	0.41	0.28	0.60	0.16	0.26	ENSG00000148700	ADD3

chr14:100842597:100842680:-@chr14:100841620:100841740:-@chr14:100835424:100835595:-	0.42	0.28	0.62	0.42	0.67	0.40	ENSG00000140105	WARS
chr3:98241386:98241910:-@chr3:98240497:98240562:-@chr3:98239977:98240286:-	0.10	0.25	0.10	0.28	0.36	0.48	ENSG00000080822, ENSG00000080819	CLDND1
chr9:21994820:21995300:-@chr9:21993881:21994052:-@chr9:21970901:21971207:-	0.40	0.29	0.35	0.20	0.43	0.23	ENSG00000147889	CDKN2A
chr17:74090495:74090662:-@chr17:74086410:74086478:-@chr17:74085256:74085401:-	0.20	0.45	0.20	0.31	0.44	0.29	ENSG00000182473	EXOC7
chr19:1358393:1358463:+@chr19:1358587:1358699:+@chr19:1360135:1361031:+	0.07	0.22	0.10	0.26	0.13	0.23	ENSG00000160953	MUM1
chr20:37076109:37076266:+@chr20:37076573:37076736:+@chr20:37077305:37077373:+	0.36	0.24	0.38	0.16	0.33	0.15	ENSG00000174365	SNHG11
chr8:26721604:26722922:-@chr8:26716549:26716722:-@chr8:26605667:26606265:-	0.28	0.46	0.25	0.46	0.06	0.29	ENSG00000120907	ADRA1A
chr10:79796952:79797062:+@chr10:79799959:79799983:+@chr10:79800373:79800473:+	0.94	0.75	0.85	0.55	0.72	0.51	ENSG00000138326	RPS24
chr10:103902802:103902855:+@chr10:103904007:103904064:+@chr10:103908129:103908278:+	0.70	0.56	0.73	0.55	0.65	0.53	ENSG00000148840	PPRC1
chr7:103123320:103123418:-@chr7:103113449:103113453:-@chr7:103112231:103113355:-	0.15	0.33	0.24	0.39	0.05	0.16	ENSG00000189056	RELN
chr1:160109683:160109774:+@chr1:160110441:160110564:+@chr1:160111084:160113374:+	0.03	0.22	0.08	0.32	0.36	0.25	ENSG00000018625	ATP1A2
chr1:155170491:155170617:+@chr1:155172914:155173062:+@chr1:155174826:155175286:+	0.13	0.28	0.37	0.25	0.40	0.28	ENSG00000231064	RP11-263K19.4
chr2:3605976:3606588:+@chr2:3607038:3607319:+@chr2:3608907:3609340:+	0.38	0.50	0.40	0.54	0.20	0.40	ENSG00000234171	RNASEH1-AS1
chr14:73749067:73749213:-@chr14:73745989:73746132:-@chr14:73741918:73744001:-	0.19	0.48	0.07	0.35	0.16	0.32	ENSG00000133961	NUMB
chr10:79796952:79797062:+@chr10:79799962:79799982:+@chr10:79800373:79800473:+	0.82	0.43	0.63	0.30	0.53	0.27	ENSG00000138326	RPS24
chr3:24338717:24338862:-@chr3:24270429:24270492:-@chr3:24231565:24231825:-	0.87	0.72	0.70	0.59	0.63	0.74	ENSG00000151090	THRB
chr16:46727005:46727094:+@chr16:46729474:46729586:+@chr16:46729929:46729997:+	0.66	0.82	0.66	0.88	0.62	0.73	ENSG00000091651	ORC6
chr7:129710349:129710649:+@chr7:129736761:129736847:+@chr7:129756285:129756506:+	0.53	0.64	0.45	0.59	0.53	0.63	ENSG00000128607	KLHDC10
chr15:91512754:91512853:-@chr15:91512309:91512350:-@chr15:91509268:91510432:-	0.22	0.10	0.25	0.12	0.39	0.21	ENSG00000198901	PRC1
chr3:98241386:98241910:-@chr3:98240497:98240540:-@chr3:98239977:98240286:-	0.14	0.30	0.13	0.32	0.39	0.52	ENSG00000080822, ENSG00000080819	CLDND1
chr17:79865430:79865474:-@chr17:79865080:79865133:-@chr17:79864636:79864774:-	0.42	0.52	0.50	0.62	0.53	0.39	ENSG00000185813	PCYT2
chr2:216238045:216238134:-@chr2:216236832:216237023:-@chr2:216236632:216236738:-	0.69	0.80	0.70	0.90	0.72	0.87	ENSG00000115414	FN1
chr2:175351601:175351816:-@chr2:175347738:175347886:-@chr2:175346225:175346715:-	0.30	0.48	0.44	0.56	0.44	0.58	ENSG00000163328	GPR155
chr2:238303230:238303847:-@chr2:238296225:238296827:-@chr2:238289558:238290142:-	0.23	0.60	0.79	0.63	0.48	0.32	ENSG00000163359	COL6A3

chr13:76378425:76378677:+@chr13:76383290:76383319:+@chr13:76391297:76391414:+	0.58	0.48	0.86	0.59	0.69	0.54	ENSG00000136153	LMO7
chr3:37136283:37136399:-@chr3:37132958:37133029:-@chr3:37125127:37125297:-	0.69	0.30	0.80	0.64	0.77	0.56	ENSG00000093167	LRRFIP2
chr12:123694603:123694709:-@chr12:123687797:123687922:-@chr12:123687171:123687631:-	0.34	0.50	0.39	0.57	0.40	0.54	ENSG00000051825	MPHOSPH9
chr7:30618847:30618930:-@chr7:30618622:30618744:-@chr7:30617111:30617707:-	0.58	0.30	0.50	0.27	0.34	0.23	ENSG00000196295	AC005154.6
chr1:171196790:171196927:-@chr1:171173030:171173127:-@chr1:171168386:171168654:-	0.97	0.77	0.99	0.83	0.67	0.53	ENSG00000225243, ENSG00000231424	RP1-127D3.4
chr7:30793347:30793554:+@chr7:30805961:30806177:+@chr7:30818048:30818167:+	0.13	0.31	0.02	0.27	0.16	0.34	ENSG00000254959	INMT-FAM188B
chr11:35211382:35211519:+@chr11:35232793:35232996:+@chr11:35236399:35236461:+	0.67	0.80	0.48	0.83	0.31	0.42	ENSG00000026508	CD44
chr5:122181160:122181388:+@chr5:122226722:122226820:+@chr5:122272429:122272512:+	0.32	0.20	0.22	0.33	0.40	0.28	ENSG00000064652	SNX24
chr9:139304780:139305054:-@chr9:139304542:139304691:-@chr9:139302278:139302390:-	0.44	0.55	0.46	0.58	0.41	0.51	ENSG00000165689	SDCCAG3
chr3:180630234:180630524:+@chr3:180632440:180632783:+@chr3:180651122:180651174:+	0.46	0.28	0.47	0.37	0.49	0.35	ENSG00000114416	FXR1
chr17:65870933:65871136:+@chr17:65871672:65871860:+@chr17:65882244:65882432:+	0.73	0.36	0.73	0.40	0.65	0.44	ENSG00000171634	BPTF
chr11:120195838:120196077:+@chr11:120197831:120198349:+@chr11:120200686:120204388:+	0.63	0.76	0.64	0.85	0.59	0.75	ENSG00000181264	TMEM136
chr15:41624113:41624179:-@chr15:41611856:41611978:-@chr15:41605471:41605552:-	0.64	0.81	0.64	0.85	0.59	0.77	ENSG00000104147	OIP5
chr16:69166387:69166493:-@chr16:69154956:69155073:-@chr16:69151912:69154552:-	0.69	0.56	0.75	0.63	0.58	0.48	ENSG00000168802	CHTF8
chr16:69155339:69155396:-@chr16:69154956:69155073:-@chr16:69151912:69154552:-	0.66	0.54	0.69	0.58	0.57	0.46	ENSG00000168802	CHTF8
chr8:145738025:145738154:-@chr8:145737775:145737944:-@chr8:145737527:145737707:-	0.82	0.92	0.82	0.94	0.76	0.89	ENSG00000160957	RECQL4
chr14:100842597:100842680:-@chr14:100841620:100841687:-@chr14:100835424:100835595:-	0.49	0.34	0.69	0.48	0.71	0.45	ENSG00000140105	WARS
chr20:44442076:44442103:+@chr20:44443023:44443109:+@chr20:44444180:44444384:+	0.82	0.96	0.82	0.98	0.69	0.92	ENSG00000175063	UBE2C
chr3:160166514:160166583:-@chr3:160165520:160165630:-@chr3:160153291:160156974:-	0.19	0.05	0.21	0.05	0.41	0.29	ENSG00000213186, ENSG00000248710	TRIM59
chr8:26721604:26722922:-@chr8:26716549:26716722:-@chr8:26613913:26614296:-	0.26	0.45	0.21	0.43	0.05	0.26	ENSG00000120907	ADRA1A
chr1:78963559:78963629:+@chr1:78997898:78998038:+@chr1:79002091:79006386:+	0.15	0.31	0.22	0.36	0.39	0.21	ENSG00000122420	PTGFR
chr3:152163071:152163328:+@chr3:152164493:152164546:+@chr3:152165409:152165562:+	0.20	0.44	0.13	0.46	0.19	0.39	ENSG00000152601	MBNL1
chr19:7152737:7152938:-@chr19:7150508:7150543:-@chr19:7142827:7143101:-	0.63	0.30	0.53	0.38	0.77	0.61	ENSG00000171105	INSR
chr6:168311964:168312169:+@chr6:168314083:168314103:+@chr6:168314848:168314993:+	0.43	0.23	0.66	0.53	0.42	0.22	ENSG00000130396	MLLT4
chr15:101827113:101827215:-@chr15:101826419:101826498:-	0.60	0.83	0.74	0.89	0.66	0.78	ENSG00000131876	SNRPA1

@chr15:101825931:101826006:-								
chr12:53677842:53677951:+@chr12:53679708:53680696:+@chr12:53681756:53682125:+	0.73	0.86	0.70	0.83	0.51	0.71	ENSG00000135476	ESPL1
chr20:44442076:44442103:+@chr20:44444180:44444384:+@chr20:44444493:44444552:+	0.65	0.88	0.68	0.91	0.60	0.79	ENSG00000175063	UBE2C
chr14:102698872:102699045:-@chr14:102691432:102692745:-@chr14:102690837:102691131:-	0.75	0.63	0.70	0.58	0.57	0.46	ENSG00000080823	MOK
chr10:79796952:79797062:+@chr10:79799962:79799983:+@chr10:79800373:79800473:+	0.96	0.79	0.89	0.62	0.80	0.58	ENSG00000138326	RPS24
chr5:131705401:131706057:+@chr5:131714070:131719993:+@chr5:131721020:131721191:+	0.75	0.59	0.57	0.42	0.40	0.30	ENSG00000197375	SLC22A5
chr20:50418818:50419048:-@chr20:50407872:50408891:-@chr20:50405400:50405680:-	0.62	0.73	0.59	0.75	0.61	0.73	ENSG00000101115	SALL4
chr13:50007454:50007529:-@chr13:49975269:49975406:-@chr13:49956936:49957077:-	0.65	0.46	0.51	0.32	0.40	0.17	ENSG00000102547	CAB39L
chr14:100842597:100842680:-@chr14:100841620:100841743:-@chr14:100835424:100835595:-	0.41	0.28	0.61	0.41	0.68	0.40	ENSG00000140105	WARS
chr4:338124:338219:+@chr4:366453:366796:+@chr4:376884:377760:+	0.81	0.67	0.75	0.65	0.59	0.41	ENSG00000131127	ZNF141
chr14:73753818:73754022:-@chr14:73749067:73749213:-@chr14:73741918:73744001:-	0.80	0.52	0.89	0.69	0.67	0.49	ENSG00000133961	NUMB
chr2:63206323:63206470:+@chr2:63215066:63215173:+@chr2:63217851:63217975:+	0.26	0.54	0.31	0.63	0.22	0.40	ENSG00000115504	EHBP1
chr10:70287003:70287280:-@chr10:70276841:70277022:-@chr10:70276508:70276600:-	0.23	0.47	0.42	0.53	0.33	0.53	ENSG00000122912	SLC25A16
chrX:108939373:108939425:-@chrX:108928656:108928673:-@chrX:108926365:108926895:-	0.24	0.36	0.15	0.33	0.37	0.26	ENSG00000068366	ACSL4
chr15:44673004:44673174:+@chr15:44695085:44695252:+@chr15:44705534:44707959:+	0.54	0.42	0.54	0.42	0.53	0.39	ENSG00000166734	CASC4
chr10:12191851:12191960:+@chr10:12197777:12197930:+@chr10:12198906:12199066:+	0.39	0.61	0.45	0.61	0.48	0.65	ENSG00000065665	SEC61A2
chr1:143910078:143910155:-@chr1:143906012:143906136:-@chr1:143896452:143897641:-	0.57	0.72	0.61	0.79	0.46	0.62	ENSG00000215784	FAM72D
chr20:50418818:50419048:-@chr20:50406561:50408891:-@chr20:50405400:50405680:-	0.68	0.81	0.66	0.82	0.67	0.79	ENSG00000101115	SALL4
chr3:98241386:98241910:-@chr3:98240497:98240547:-@chr3:98239977:98240286:-	0.12	0.28	0.13	0.32	0.38	0.50	ENSG00000080822, ENSG00000080819	CLDND1
chr9:21994820:21995300:-@chr9:21993881:21994067:-@chr9:21970901:21971207:-	0.38	0.28	0.34	0.19	0.36	0.23	ENSG00000147889	CDKN2A
chr8:95479633:95479783:-@chr8:95470496:95470664:-@chr8:95423349:95423543:-	0.70	0.86	0.65	0.87	0.55	0.68	ENSG00000197275	RAD54B
chr1:78958357:78959226:+@chr1:78963559:78963629:+@chr1:79002091:79006386:+	0.11	0.30	0.15	0.34	0.39	0.20	ENSG00000122420	PTGFR
chr6:30294131:30294927:-@chr6:30282046:30282259:-@chr6:30263909:30264014:-	0.58	0.69	0.44	0.66	0.43	0.53	ENSG00000231074	HCG18
chr8:67579787:67579936:+@chr8:67589877:67590189:+@chr8:67591956:67592259:+	0.44	0.59	0.43	0.61	0.48	0.62	ENSG00000213865	C8orf44
chr4:56749989:56750094:+@chr4:56755054:56755098:+@chr4:56756389:56756552:+	0.41	0.74	0.14	0.61	0.34	0.51	ENSG00000090989	EXOC1
chr17:62747925:62748074:-	0.30	0.14	0.34	0.11	0.27	0.16	ENSG00000215769	hsa-mir-

@chr17:62746712:62746841:- @chr17:62745780:62746126:-								6080
chr7:102185153:102185223:- @chr7:102183972:102184108:- @chr7:102181948:102182109:-	0.71	0.83	0.79	0.94	0.68	0.78	ENSG00000168255, ENSG00000228049	POLR2J3
chr20:44463598:44463755:+@chr20:44469087:4 4469097:+@chr20:44469278:44471914:+	0.40	0.23	0.36	0.25	0.40	0.28	ENSG00000124104	SNX21
chr14:50116983:50117159:- @chr14:50114011:50114078:- @chr14:50110270:50110388:-	0.71	0.83	0.68	0.85	0.70	0.82	ENSG00000100479	POLE2
chr3:12330436:12330633:+@chr3:12353879:123 53952:+@chr3:12421203:12421430:+	0.92	0.70	0.93	0.77	0.69	0.80	ENSG00000132170	PPARG
chr5:122181160:122181388:+@chr5:122189369: 122189462:+@chr5:122272429:122272512:+	0.31	0.20	0.19	0.33	0.39	0.28	ENSG00000064652	SNX24
chr9:19336268:19336412:+@chr9:19336684:193 36830:+@chr9:19340990:19341112:+	0.67	0.82	0.64	0.79	0.59	0.74	ENSG00000137145	DENND4C
chr14:100842597:100842680:- @chr14:100841620:100841883:- @chr14:100835424:100835595:-	0.30	0.19	0.46	0.28	0.57	0.30	ENSG00000140105	WARS
chr11:62623484:62623853:+@chr11:62639049:6 2639141:+@chr11:62648491:62648919:+	0.49	0.67	0.12	0.48	0.25	0.52	ENSG00000168003	SLC3A2
chr11:113142482:113142598:+@chr11:1131436 55:113143770:+@chr11:113145989:113149158: +	0.05	0.24	0.11	0.24	0.17	0.31	ENSG00000149294	NCAM1
chr17:8213556:8213661:+@chr17:8214146:8214 181:+@chr17:8215309:8215958:+	0.07	0.17	0.06	0.30	0.40	0.28	ENSG00000198844	ARHGEF1 5
chr22:31495731:31495882:+@chr22:31496871:3 1496939:+@chr22:31500302:31500610:+	0.76	0.92	0.67	0.91	0.73	0.84	ENSG00000183963, ENSG00000100330	SMTN
chr7:102185153:102185223:- @chr7:102183972:102184149:- @chr7:102181948:102182109:-	0.72	0.84	0.79	0.94	0.67	0.79	ENSG00000168255, ENSG00000228049	POLR2J3
chr8:26721604:26722922:- @chr8:26627798:26628183:- @chr8:26605667:26606265:-	0.82	0.59	0.83	0.60	0.94	0.78	ENSG00000120907	ADRA1A
chr1:53387225:53387591:- @chr1:53379567:53379767:- @chr1:53377395:53377462:-	0.72	0.59	0.51	0.66	0.20	0.38	ENSG00000121310	ECHDC2
chr6:138768138:138768330:- @chr6:138763120:138763251:- @chr6:138751530:138754817:-	0.22	0.47	0.24	0.36	0.40	0.54	ENSG00000135540	NHSL1
chr5:148619322:148619451:+@chr5:148622054: 148622101:+@chr5:148624444:148624578:+	0.69	0.49	0.77	0.56	0.75	0.61	ENSG00000173210	ABLIM3
chr1:9801152:9801314:- @chr1:9797556:9797612:- @chr1:9795943:9796100:-	0.51	0.19	0.45	0.34	0.51	0.36	ENSG00000171603	CLSTN1
chr4:88898211:88898249:+@chr4:88901198:889 01278:+@chr4:88901545:88901586:+	0.75	0.86	0.71	0.89	0.65	0.87	ENSG00000118785	SPP1
chr3:123452534:123453069:- @chr3:123451743:123451949:- @chr3:123444791:123444925:-	0.69	0.56	0.71	0.51	0.56	0.42	ENSG00000065534	MYLK
chr2:173362703:173362828:+@chr2:173366500: 173366629:+@chr2:173368819:173371181:+	0.69	0.44	0.56	0.74	0.32	0.21	ENSG00000091409	ITGA6
chr6:33366065:33366164:+@chr6:33368169:333 68291:+@chr6:33371091:33371144:+	0.34	0.10	0.35	0.13	0.50	0.19	ENSG00000237649	KIFC1
chrX:154255055:154255215:- @chrX:154250685:154250771:- @chrX:154227754:154227875:-	0.65	0.40	0.72	0.42	0.55	0.44	ENSG00000185010	F8
chr7:79088163:79088315:+@chr7:79088723:790 88769:+@chr7:79088852:79096783:+	0.04	0.15	0.04	0.19	0.09	0.28	ENSG00000234456	MAGI2- AS3
chr3:12328867:12329174:+@chr3:12353879:123 53952:+@chr3:12421203:12421430:+	0.96	0.74	0.96	0.79	0.72	0.84	ENSG00000132170	PPARG
chr7:102185153:102185223:- @chr7:102183972:102184144:- @chr7:102181948:102182109:-	0.72	0.84	0.78	0.94	0.68	0.78	ENSG00000168255, ENSG00000228049	POLR2J3

chr15:91512754:91512853:-@chr15:91512309:91512350:-@chr15:91509280:91509883:-	0.19	0.04	0.21	0.04	0.39	0.11	ENSG00000198901	PRC1
chr1:2433551:2433848:+@chr1:2435056:2435082:+@chr1:2435361:2436964:+	0.24	0.40	0.39	0.19	0.48	0.34	ENSG00000149527	PLCH2
chr3:37396592:37396678:+@chr3:37402734:37402796:+@chr3:37407571:37408370:+	0.40	0.71	0.29	0.51	0.43	0.55	ENSG00000144674	GOLGA4
chr12:49580394:49580616:-@chr12:49580093:49580241:-@chr12:49578578:49579773:-	0.92	0.75	0.89	0.67	0.56	0.40	ENSG00000167552	TUBA1A
chr1:110213909:110214004:+@chr1:110214095:110214205:+@chr1:110217369:110217908:+	0.86	0.70	0.86	0.73	0.33	0.52	ENSG00000213366	GSTM2
RI								
chr2:74686565:74686679:+@chr2:74686770:74686872:+	0.48	0.62	0.49	0.69	0.46	0.64	ENSG00000239779	WBP1
chr21:45751897:45751726:-@chr21:45750799:45750346:-	0.39	0.50	0.34	0.49	0.16	0.26	ENSG00000160226	C21orf2
chr3:132298402:132298336:-@chr3:132297725:132297640:-	0.43	0.53	0.45	0.62	0.09	0.25	ENSG00000240303	ACAD11
chr1:28907158:28907072:-@chr1:28906493:28906045:-	0.39	0.25	0.50	0.32	0.43	0.31	ENSG00000197989	SNHG12
chr19:10446652:10446347:-@chr19:10446029:10445742:-	0.20	0.40	0.07	0.22	0.11	0.38	ENSG00000076662	ICAM3
chr9:139753661:139753769:+@chr9:139754338:139754516:+	0.12	0.32	0.25	0.45	0.04	0.16	ENSG00000177943	MAMDC4
chr2:74686565:74686689:+@chr2:74686770:74686872:+	0.35	0.52	0.38	0.62	0.39	0.55	ENSG00000239779	WBP1
chr3:52005908:52005828:-@chr3:52005714:52005476:-	0.28	0.49	0.32	0.61	0.18	0.33	ENSG00000114779	ABHD14B
chr1:161146896:161146702:-@chr1:161146369:161146224:-	0.55	0.31	0.48	0.37	0.53	0.34	ENSG00000158850	B4GALT3
chr1:28907158:28906937:-@chr1:28906493:28906045:-	0.58	0.43	0.66	0.53	0.63	0.46	ENSG00000197989	SNHG12
A3SS								
chr8:67364188:67364340:+@chr8:67365945|67366294:67366392:+	0.29	0.50	0.38	0.64	0.11	0.28	ENSG00000147576	ADHFE1
chr16:1843638:1843734:-@chr16:1842402|1842516:1840414:-	0.39	0.27	0.18	0.39	0.02	0.14	ENSG00000099769	IGFALS
chr19:19144419:19144822:+@chr19:19153520|19153525:19153686:+	0.67	0.78	0.66	0.81	0.59	0.70	ENSG00000105676	ARMC6
chr12:53343499:53343608:-@chr12:53343362|53343369:53343240:-	0.60	0.75	0.71	0.57	0.76	0.64	ENSG00000170421	KRT8
chr17:27045140:27045286:-@chr17:27044611|27044665:27044231:-	0.51	0.62	0.50	0.64	0.53	0.67	ENSG00000109113	RAB34
chr6:44095376:44095415:+@chr6:44102157|44102298:44102480:+	0.31	0.17	0.08	0.19	0.28	0.16	ENSG00000137216	TMEM63B
chr2:74685527:74685798:+@chr2:74686565|74686605:74686689:+	0.52	0.40	0.46	0.29	0.36	0.24	ENSG00000115274, ENSG00000239779	INO80B
chr1:63872733:63872797:+@chr1:63876811|63876817:63877002:+	0.22	0.11	0.31	0.19	0.38	0.21	ENSG00000088035	ALG6
chr16:3024001:3024158:-@chr16:3023216|3023254:3023139:-	0.41	0.19	0.44	0.16	0.45	0.24	ENSG00000127564	PKMYT1
chr16:3024001:3024158:-@chr16:3023216|3023446:3023139:-	0.81	0.67	0.81	0.54	0.76	0.62	ENSG00000127564	PKMYT1
chr12:49580394:49580616:-@chr12:49580197|49580241:49580093:-	0.99	0.85	0.97	0.52	0.56	0.44	ENSG00000167552	TUBA1A
chr3:136676707:136677043:+@chr3:136699173|136699308:136699434:+	0.34	0.44	0.39	0.23	0.29	0.41	ENSG00000174564	IL20RB
chr9:137717638:137717750:+@chr9:137721822|137721954:137722022:+	0.42	0.64	0.43	0.71	0.46	0.59	ENSG00000130635	COL5A1
chr17:27045140:27045286:-@chr17:27044611|27044638:27044231:-	0.45	0.56	0.48	0.63	0.49	0.62	ENSG00000109113	RAB34

chr7:100488790:100488959:-@chr7:100487956|100488709:100487615:-	0.43	0.60	0.39	0.52	0.30	0.47	ENSG00000087085	ACHE
A5SS								
chr1:169559440:169559385|169559386:-@chr1:169558090:169558699:-	0.81	0.68	0.89	0.65	0.65	0.55	ENSG00000174175	SELP
chr21:44293801:44293661|44293713:-@chr21:44283550:44283706:-	0.68	0.81	0.72	0.86	0.74	0.85	ENSG00000160193	WDR4
chr19:58055951:58055019|58055605:-@chr19:58053208:58054611:-	0.43	0.57	0.42	0.55	0.43	0.53	ENSG00000105132, ENSG00000251369	ZNF550
chr17:26925653:26925483|26925528:-@chr17:26919000:26920084:-	0.23	0.06	0.25	0.06	0.25	0.10	ENSG00000076382, ENSG00000258472	SPAG5
chr8:124360513:124360423|124360427:-@chr8:124359332:124359646:-	0.81	0.93	0.76	0.92	0.76	0.87	ENSG00000156802	ATAD2
chr12:124812179:124811955|124812093:-@chr12:124810737:124810916:-	0.54	0.78	0.42	0.65	0.64	0.75	ENSG00000196498	NCOR2
chr1:42922173:42922299|42922744:+@chr1:42922918:42923021:+	0.28	0.52	0.43	0.65	0.44	0.56	ENSG00000127125	PPCS
chr16:15802698:15802659|15802660:-@chr16:15796992:15797980:-	0.03	0.33	0.01	0.29	0.29	0.41	ENSG00000133392	MYH11
chr19:38806445:38806357|38806387:-@chr19:38800045:38800283:-	0.36	0.48	0.45	0.62	0.53	0.79	ENSG00000167645	YIF1B
chr9:137721822:137721890|137722022:+@chr9:137726817:137727050:+	0.18	0.06	0.17	0.05	0.21	0.09	ENSG00000130635	COL5A1
chr12:4665519:4665668|4665716:+@chr12:4668023:4669213:+	0.18	0.07	0.16	0.06	0.30	0.14	ENSG00000111247	RAD51AP1
chr6:34204650:34204738|34204740:+@chr6:34208514:34208659:+	0.54	0.69	0.62	0.92	0.49	0.67	ENSG00000137309	HMGA1
chr16:1733510:1733593|1733621:+@chr16:1735440:1735588:+	0.21	0.09	0.30	0.14	0.32	0.21	ENSG00000206053, ENSG00000261732	HN1L
chr16:16315688:16315470|16315506:-@chr16:16313679:16313804:-	0.46	0.59	0.43	0.58	0.19	0.40	ENSG00000091262	ABCC6

Table S 3.2 MCODE Cluster Results of corresponding proteins of cancer-specific AS

Cluster	Score (Density*#Nodes)	Nodes	Edges	Node IDs
1	9.28	26	116	POLE2, MYLK, LMNB1, FN1, CDCA8, ORC6L, RAD51AP1, SPAG5, PRC1, UBE2C, TUBA1A, CDKN2A, ESPL1, OIP5, AURKB, PKMYT1, TRIM59, MPHOSPH9, BIRC5, RAD54B, ATAD2, RECQL4, CDCA5, INCENP, RELN, KIFC1
2	4	14	26	FN1, SELP, ICAM3, KRT8, SPP1, COL5A1, COL6A3, CD44, NCAM1, CDKN2A, SLC3A2, RELN, ANXA2, ITGA6
3	4	10	18	HMGA1, PPARG, INSR, FN1, PTPN1, PPRC1, THRB, NCOR2, IRS1, ANXA2

*Parameters:
Network Scoring:
Include Loops: false
Degree Cutoff: 2
Cluster Finding:
Node Score Cutoff: 0.2
Haircut: false
Fluff: true
Fluff Density Cutoff: 0.1
K-Core: 2
Max. Depth from Seed: 100

Table S 3.3 MCODE Cluster Results of proteins that are highly correlated with the cancer-specific AS

Cluster	Score (Density*#Nodes)	Nodes	Edges	Node IDs
1	32	32	496	CENPO, SPC25, NDC80, AURKB, CENPH, MLF1IP, ERCC6L, CENPE, SPC24, BUB1B, ZWILCH, ZWINT, CENPK, CENPL, KIF18A, SGOL2, BIRC5, NUF2, BUB1, CENPA, SGOL1, SKA1, INCENP, CDCA8, KIF2C, CDC20, CENPN, CASC5, CENPI, CENPM, PLK1, CENPF
2	13.529	18	115	MCM6, CLSPN, CDT1, POLE2, POLA2, MCM10, CDC6, PRIM2, MCM4, DBF4, MCM7, CDC45, ORC1L, CDC7, ORC6L, CHEK1, CCNB1, MCM2
3	5	5	10	CCNE1, TYMS, E2F1, PCNA, RRM2
4	5	5	10	UBE2T, FANCA, FANCB, FANCD2, FANCI
5	4	4	6	KIF4A, SKA3, AURKA, PRC1
6	4	4	6	NCAPG, NCAPH, NCAPD2, SMC4
7	3	3	3	HJURP, OIP5, RUVBL1
8	3	3	3	UBE2C, PSMD2, PTTG1
9	3	3	3	RFC4, TIMELESS, RFC2
10	3	7	9	MND1, CDK1, RAD51, PKMYT1, CDC25A, CDK4, CCNE2
11	3	3	3	POLR2H, SNRPA1, HNRNPL

*Parameters:
Network Scoring:
Include Loops: false
Degree Cutoff: 2
Cluster Finding:
Node Score Cutoff: 0.2
Haircut: true
Fluff: false
K-Core: 2
Max. Depth from Seed: 100

CHAPTER 4

GLOBAL INTRON RETENTION IN HUMAN KIDNEY TUMORS CORRELATES WITH PATIENT SURVIVAL[3]

4.1 Overview

The identification of molecular traits that drive clinical outcomes is critical to a complete understanding of cancer and the development of effective therapeutics. Gene regulatory networks are well-established predictors of patient survival and drug response. However, most current analyses have focused almost exclusively on examining mRNA levels, making it unclear if other gene regulatory mechanisms like alternative splicing (AS) are clinically informative molecular traits. We carried out a comprehensive analysis of the clear-cell renal carcinoma (ccRCC) transcriptome and uncovered widespread changes in alternative splicing. A significant increase in global intron retention in tumor *vs.* matched normal kidney emerged as the predominant abnormality. Variability in global intron retention levels was observed across tumors, that is, some tumors generally splice out introns better than others. We define a novel class of renal tumors based on global intron retention (IR-class). While the IR-class does not associated with known genetic alterations, they do correlate with previously identified kidney cancer subtypes, have significantly shorter overall survival and exhibit high expression of non-coding RNAs and

[3] This work is currently under preparation as a manuscript to the *Cancer Research* journal.

key splicing regulators. Further analysis revealed that intron retention-based classifications are specific to kidney cancers and not other tumor types. This work demonstrates that intron retention patterns are molecular traits with predictive power in kidney tissues, and that these tumors have co-opted the splicing landscape to alter pathways that provide an advantage to tumors.

4.2 Introduction

Generally, the choice of splice sites within a pre-mRNA is regulated by *cis*-acting sequence elements that interact with *trans*-acting protein factors that promote or inhibit recruitment of the splicing machinery. This process is tightly regulated during development, differentiation, circadian rhythms, and varies dramatically across different tissues [10, 11]. Mis-regulated AS has been observed in wide range of human cancers [14, 15, 128], however the consequence on patient survival and tumor biology remains unclear. Known drivers of these AS alterations include changes in the expression level of specific splicing factors including the proto-oncogene SRSF1 or the tumor suppressor RBM4, which promote or inhibit a "cancer state" via splicing regulation [46, 49]. While some tumors harbor genetic alterations in splicing factors like U2AF and SF3B [129-132], these mutations account for only a small fraction of genetic alterations across tumors, whereas emerging evidence suggests that many tumors have altered splicing.

ccRCC is the most prevalent type of kidney cancer, and is considered a heterogeneous disease with incompletely understood molecular pathogenesis. Genetic alterations in chromosome 3p leading to bi-allelic loss of *VHL* are the most common feature of ccRCC [133, 134]. Secondary alterations in *SETD2*, *PBRM1* and *BAP1* have also emerged as frequent

recurrent mutations which reside in the same genetic locus [135-137]. These proteins, as well as general chromatin features, have been implicated in splicing regulation [138]. Previous, studies have shown a correlation between *SETD2* loss and aberrant splicing in tumors and cell lines with defects in intron retention being the most affected [138]. Furthermore, mutations in various splicing factors were identified in ccRCC by whole genome sequencing [134], however their consequence on the splicing landscape is not known. Together these data suggest that mis-regulation of AS in ccRCC could be an alteration underlying its pathogenesis. Here we report widespread alterations of AS in ccRCC and define a class of tumors with high global intron retention (IR-class). These intron-based classifications appear to be specific to ccRCC, and not other tumor types including breast, lung and liver cancers.

4.3 Results

4.3.1 Intron retention is widespread in kidney tumors

Through analysis in the previous chapter (3.3.9), we noticed that kidney cancer has a widespread splicing mis-regulation, especially in the intron retention category. A thorough transcriptome analysis was then applied to the ccRCC (i.e. KIRC in the TCGA data). About five hundred kidney tumors and 72 normal paired samples were analyzed by the MISO pipeline to measure the skipped exons (SE), retained introns (RI), alternative 3' (A3SS) and 5'splice sites (A5SS) [122] (See 3.5.1 for details). In total over 15,000 splicing events were analyzed. Over a thousand events were altered between normal and tumor tissues (Figure 3.7). The most striking splicing change was a statistically significant increase in about 2/3 retained introns analyzed. To make sure this phenomenon is specific to retained introns and kidney cancer only, we compared the results with the other three types of splicing and other types of cancer. We did not see the

significant increase or decrease in other types of splicing (i.e. they changed in both directions, both inclusion and skipping of exons. Figure 4.1A). We also compared it with breast and lung cancer and found no significant increase in retained intron in these two types of cancer (Figure 4.1B). We also examined the levels of global intron retention in non-transformed normal kidney and compared this to tumor kidney and found kidney cancer showed much higher variability in intron retention compared to the normal kidney (Figure 4.1C), indicating that the widespread intron retention may be a feature of kidney tumors.

4.3.2 Some kidney tumors express significant more retained introns than the others

Remarkably, we noted that some tumors exhibited pervasive intron retention, meaning many detected introns in single tumor were co-retained. To further examine such common intron retentions in some kidney tumors, we stratified tumors by calculating a retention score, which is the mean PSI value of all 3190 detected introns. Based on the retention scores, 124 cases (top quartile) were defined as having high intron retentions and 373 cases were defined as having normal intron retentions (within the distribution of normal kidney retention scores). We refer them as HIR (High Intron Retention) and NIR (Normal Intron Retention) classes, respectively (Figure 4.2A). We also classified all detected retained intron events into four subgroups using correlation analysis. For every RI event, we calculated a Spearman's rank correlation between its PSI values and the average PSI values across all tumors (Figure 4.2A, right color bar). We identified a set of 858 retained introns that were highly co-regulated across tumors (Spearman's $\rho > .75$), while only a small fraction (30%) of introns showed little or no co-regulation (Spearman's $|\rho| < 0.2$). We also found that the length of the introns of those highly co-regulated events are longer than those of events with low or no correlation ($P = 2e^{-16}$ and $7e^{-14}$ when

comparing ‘high positive’ to ‘positive’ and ‘high positive’ to ‘no correlation’ respectively, Figure 4.2B), consistent with the fact that longer introns are generally more retained [139].

Furthermore, genes harboring these highly co-regulated introns were highly enriched for RNA processing and protein degradation functions (Figure 4.2C) when we conducted a functional enrichment analysis by he DAVID annotation tool (http://david.abcc.ncifcrf.gov/) [74]. This may imply a non-random selection of these genes.

To further identify putative SREs that may control these highly co-regulated introns, we examined their regulatory regions to measure whether there are enriched sequence motifs that could be potentially recognized by known splicing factors (Figure 4.2D). We found that motifs that have been previously shown to regulate splicing from an intronic position were identified in these co-regulated introns. These include a CTTC motif, which is similar to the previously published intronic splicing silencer (ISS group C) [20] as well as a G-rich motif, which is similar to the intronic splicing enhancer (ISE group D) [21].

4.3.3 High intron retention is not correlated with any known mutations, but correlated with a recently defined subclass in kidney cancer and expression levels of some genes

To determine if any of the common mutations found in kidney cancer correlated with intron retention levels, we then analyzed prevalence of specific mutations in the HIR class including *VHL, PBRM1, SETD2, KDM5C, BAP1, PTEN, MTOR, TP53* and *PIK3CA* (Figure 4.3A). Surprisingly no specific mutations correlated with a global increase in intron retention as the percentage of the observed HIR classes with these mutations are similar to what is expected (Figure 4.3B). This indicates that the major drivers of these phenotypes may be related to gene expression changes. Kidney tumors were previously classified as M1-M4 class using mRNA expression, with the M2 and M3 being the most aggressive tumor subtype [140]. Our HIR class

is mostly composed of the M2 tumors (~60%) and it is significant more than expected (Figure 4.3C, P =0.0001), while the NIR class is comprised of the M1-M3 classes. These data indicate that, widespread changes in the mRNA expression may drive the retention phenotype.

Furthermore, when comparing gene expression differences between the HIR and NIR class, we identified 388 transcripts that are expressed at significantly higher levels in the HIR class (cutoff of $P < 9.0e^{-5}$ and greater than 2-fold expression change, Figure 4.3A bottom). Some regulators of AS, including the CLK1 kinase, which has been shown to regulate intron retention, was found within these differentially expressed genes (Figure 4.3D). Interestingly, a significant increase in some long non-coding RNAs (lncRNAs), such as MALAT1 ($P < 4.2e^{-48}$) and NEAT1 ($P < 1.9e^{-53}$) was also observed in the HIR class (Figure 4.3D). Some reports have indicated that MALAT1 may regulate alternative splicing and cell cycle control, while other studies have shown no difference in the splicing of exons. HIR tumors represent a highly specific context wherein these lncRNAs may be promoting tumorigenesis through alternative splicing.

Finally, when we look at the overlap between the genes whose expression levels altered between HIR and NIR (388 genes) and genes containing the highly co-regulated introns (588 genes), there were only 26 genes overlaps (Figure 4.3E). This indicates that in kidney cancer, aberrant gene regulation via mRNA levels and splicing are affecting nearly mutually exclusive genes sets. This mutual exclusivity is an emerging theme in various biological settings including disease states.

4.3.4 Intron retention in kidney cancer is highly correlated with patient survival

Given the substantial increase of global intron retention in these tumors we determined if these changes would bear any significance on patient survival. Strikingly, we found that HIR tumors had a significantly shorter median overall survival as compared to NIR tumors (Figure

4.4A, P=0.0004). To the best of our knowledge this is the first analysis of patient survival based on global splicing patterns. These data are also consistent with tumors of M2 class being most aggressive. However, even when this analysis was carried out with HIR tumors that were not of the M2 class, a significant survival difference was still observed, indicating that HIR represents a new molecular classification. Furthermore, no such difference in survival was seen in other tumor types when the same classifications based on intron retention were attempted in liver, lung and breast cancers (Figure 4.4B-D).

Genes whose expression levels can predict patient survival in kidney have been previously published [140]. We set out to perform a similar analysis using splicing PSI values for over 40,000 splicing events using the same RNA-sequencing data from the same tumors (Figure 4.4E). We identified over 1,100 (P <0.005 and FDR = 5) splicing events whose pattern was predictive of patient survival in kidney tumors, the majority of these events were retained introns (898 total events), but cassette exons and alternative 5' and 3' splice sites were also identified (Figure 4.4F). Given that intron retention is not as common as other alternative splicing types, but still emerged as the most powerful and prevalent predictor, is both highly significant (P = Fisher's exact test) and indicative of the role of intron spicing in kidney cancer progression. This finding was not seen in breast, lung and liver cancer since only 34, 80 and 40 introns, respectively, correlated with patient survival (Figure 4.4G). Consistent with the observation that global intron retention leads to poor survival, we found that typically high levels of specific introns correlated with poor survival. However, a small set of introns correlated with better survival. These data suggest that for some genes less retention is observed in more aggressive tumors, albeit at much lower frequency.

Intron retention is the least understood form of alternative splicing, but is the one most likely to have drastic consequences on mRNA stability and protein production by activation of the nonsense mediated decay pathway (NMD), which leads to the degradation of the mRNA [99].

We selected the genes whose retained introns most significantly correlated with median survival differences and performed gene ontology analysis (~500 genes) using DAVID online tool. The most enriched terms was alternative splicing, atp-binding, mutagenesis and etc. (Figure 4.4H). This result provides evidence that mis-spliced genes are not being randomly affected, rather they belong to specific classes of genes.

4.4 Discussion

Our work marks the first comprehensive analysis of alternative splicing and patient survival. We reveal a previously unappreciated layer of mis-regulated gene expression and more specifically intron retention. Splicing alterations across human tumors are being increasingly uncovered due to the advent RNA-sequencing. The Cancer Genome Atlas network has provided a powerful resource for the identification of mis-regulated features. Other examples of RNA-processing defects in tumors have emerged including alternative polyadenalytion and 3'UTR usage[141, 142], which is known to affect translation and micro-RNA mediated gene suppression. We predict that, RNA processing is equally mis-regulated as are mRNA levels and that measurement of specific RNA processing events (i.e. splicing or alternative polyadenylation) in human tumors will have predictive value in patient survival and drug response. Ultimately, combing alternative splicing, poly-adenylation and RNA abundance will help paint a more accurate portrait of the tumor transcriptome that can be used for meaningful

tumor classifications. Given that these RNA processing pathways affect mostly non-overlapping genes it is likely that combining these types of analysis will significantly increase the number of genes that are mis-regulated in cancer. Additionally mis-regulation of lncRNA and mircoRNA expression are altered in cancer [143-145], putting forth the possibility that most of the transcribed genome is altered in some fashion in human cancers.

4.5 Materials and Methods

4.5.1 Kidney tumor classification using retention score and splicing event correlation analysis

For every tumor sample we calculated its retention score by taking the average PSI values of all detected RI events (3190 events in KIRC). Therefore, the higher the retention score, the more retained introns it has. We also used the same method to compute the retention score for other types of cancer for analysis.

Correlations between all detected RI events and the average PSI values were calculated using spearman rank correlation. Each of the 3190 detected RI events has a vector of PSI values among all kidney tumors. Another vector is composed of the average PSI values (retention score mentioned above) among all kidney tumors. We computed the spearman rank correlation, ρ (rho), between these two vectors and repeated this procedure for every detected RI events. RI events with $\rho \geq 0.75$ were classified as “high positive”, events with $0.4 \leq \rho < 0.75$ were classified as “positive”, events with $0.2 \leq \rho < 0.4$ were classified as “weak positive”, and the rest, $\rho < 0.2$, were classified as “no correlation or negative”.

4.5.2 Motif enrichment analysis

To analyze the enriched sequence motifs near the splice sites of the 868 high positive correlated retained intron events, we first obtained nucleotide sequences from two splicing regulatory regions: upstream intron (300 bp downstream the 5'SS) and downstream intron (300 bp upstream the 3'SS) as shown in Figure 4.2D. When obtaining the sequences, we excluded the first 10 nucleotides right downstream of the upstream exon and the first 25 nucleotides right upstream of the downstream exon because these regions overlap with the splice signal motifs. We then calculated the frequency and Z-score of all 6-nt "words" near the splice sites of the 868 high positive correlated retained introns in two regulatory regions using methods described in [127]. All 6-mers with Z-score larger than 3 were then clustered by sequence similarity and multiply aligned by using CLUSTALW to identify candidate motifs. At a cutoff dissimilarity score of 2.5, we obtained 6 and 5 clusters of at least five sequences in each cluster for upstream intron and downstream intron respectively. Finally, we plotted the consensus sequence for each cluster for the two regulatory regions (Figure 4.2D).

4.5.3 Survival analysis for kidney, liver, lung and breast cancer patients.

We obtained the overall survival data of cancer patients from the UCSC Cancer Browser for all four types of cancer. For figure 4.4A-D, the patient samples were split into two groups according to their retention scores (see method 4.5.1): HIR (above the third quartile) and NIR (below the third quartile). The resulted two patient groups were compared for their probability of survival using a Kaplan-Meier survival analysis and the log rank *P* values were calculated. This process was repeated for all four types of cancer. For figure 4.4E-G, samples were separated into two groups according to each event's PSI value: high PSI (above the third quartile) and low PSI (below the third quartile). Then we tested if there is a significant difference of their probability of

survival between the two groups using the Kaplan-Meier survival analysis and the log rank *P* values were calculated. It the log rank $P < 0.005$, we considered the event has potential prediction power for patients survival. We repeated this procedure for all four types of splicing event in kidney cancer and for all RI events in other types of cancer.

In the parameter setting, if a patient deceased (event happened), the "days_to_death" was used as the time variable; if a patient is still living, the time variable is the maximum of "days_to_last_known_alive" and "days_to_last_followup".

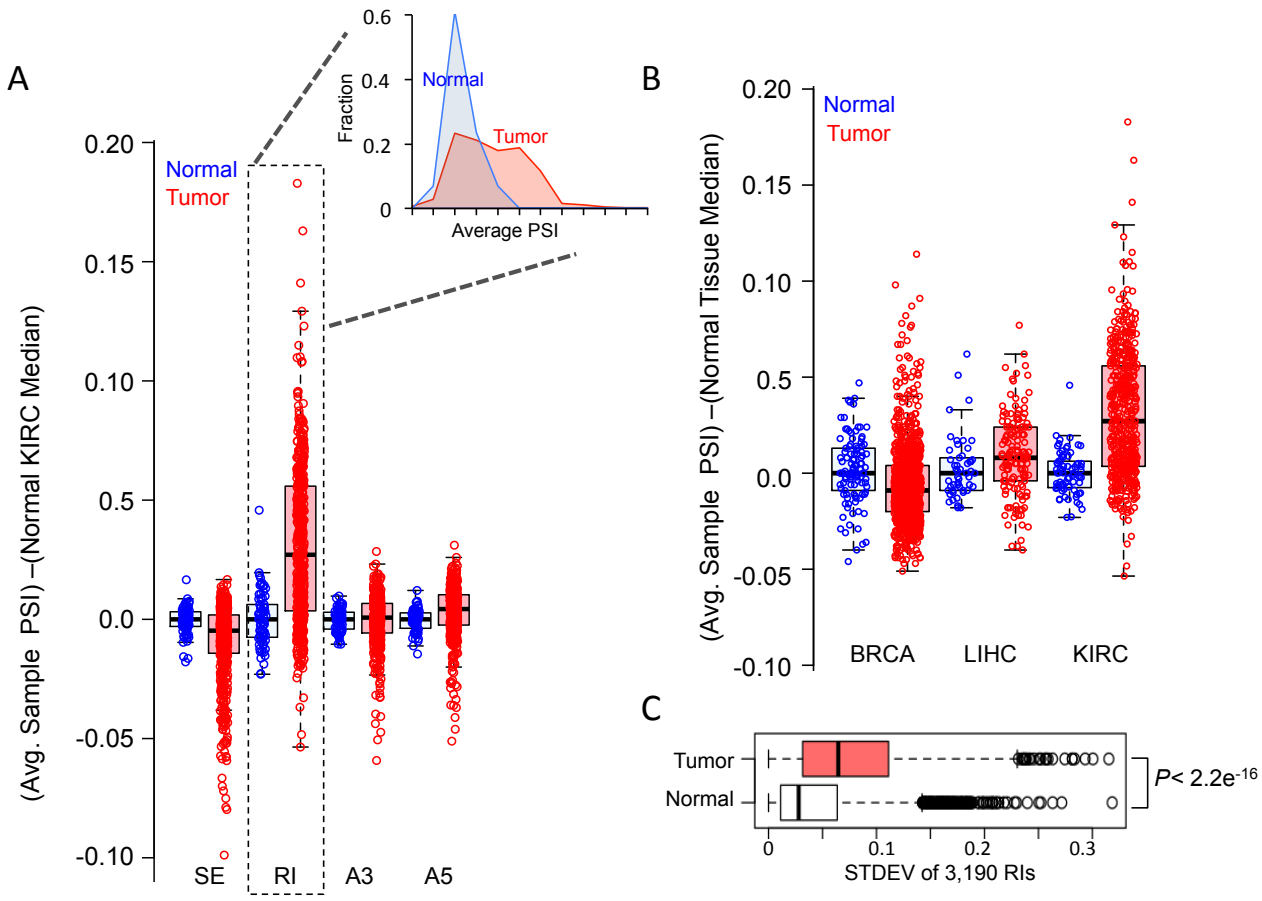

Figure 4.1 Comparing AS variability among normal vs. tumor in KIRC and other cancers.

(A) Boxplots of average PSI values of all AS events between normal (blue) vs. tumor (red) in kidney cancer in four different types of AS. Each circle represents a sample and its average psi value across all AS events minus the median of the average PSI values in normal KIRC. The insert shows the average PSI distribution of RIs shift between normal samples (blue) and tumor samples (red). (B) Boxplots of average PSI values of all RI events between normal (blue) vs. tumor (red) in breast, liver and kidney cancer. Each circle represents a sample and its average psi value across all retained intron events minus the median of the average psi values in normal tissue. (C) Boxplots of standard deviation of PSIs among all detected RIs in kidney tumors (red box) and in normal kidney tissues (white box)

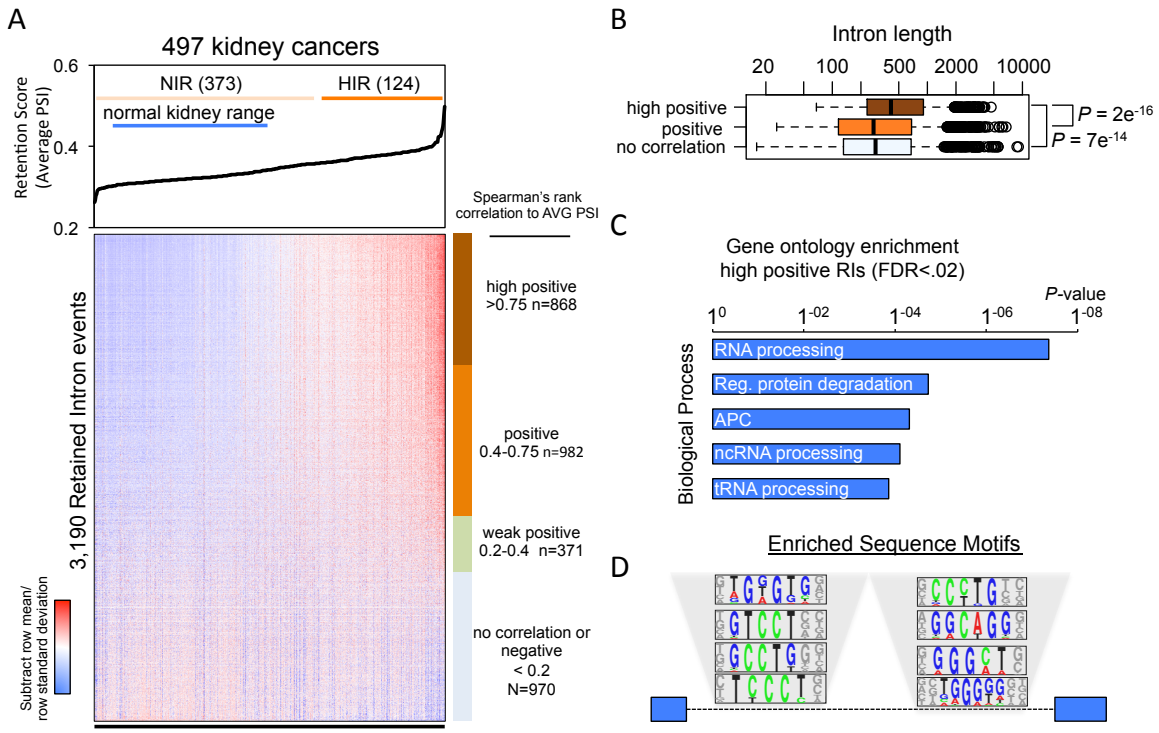

Figure 4.2 Intron retention is widespread in kidney cancers

(A) Heatmap of PSI values of the 3190 retained intron events in 497 kidney cancer samples (Red to blue: high to low; the value was subtracted by the mean and normalized to the row standard deviation). The average psi values across 3190 events in each sample were plotted on top of the heatmap. Tumors were sorted by their retention score (i.e. average PSI value) in columns (from left to right: lower score to higher score) and the top quartile was classified as HIR class and the rest were NIR class. Retained intron events was sorted by their correlation to the average PSI in rows (from top to bottom: higher correlation to lower correlation) and they are clustered into four groups according to their Spearman's rank correlation coefficient ρ with different color labeled on the right. (B) The length distribution of retained introns in the three sub-groups: high positive, positive and no correlation or negative (from top to bottom). (C) Enriched gene ontology terms of genes containing the 868 high positive correlated retained intron events. (D) Enriched sequence motifs near the splice sites of the 868 high positive correlated retained intron events.

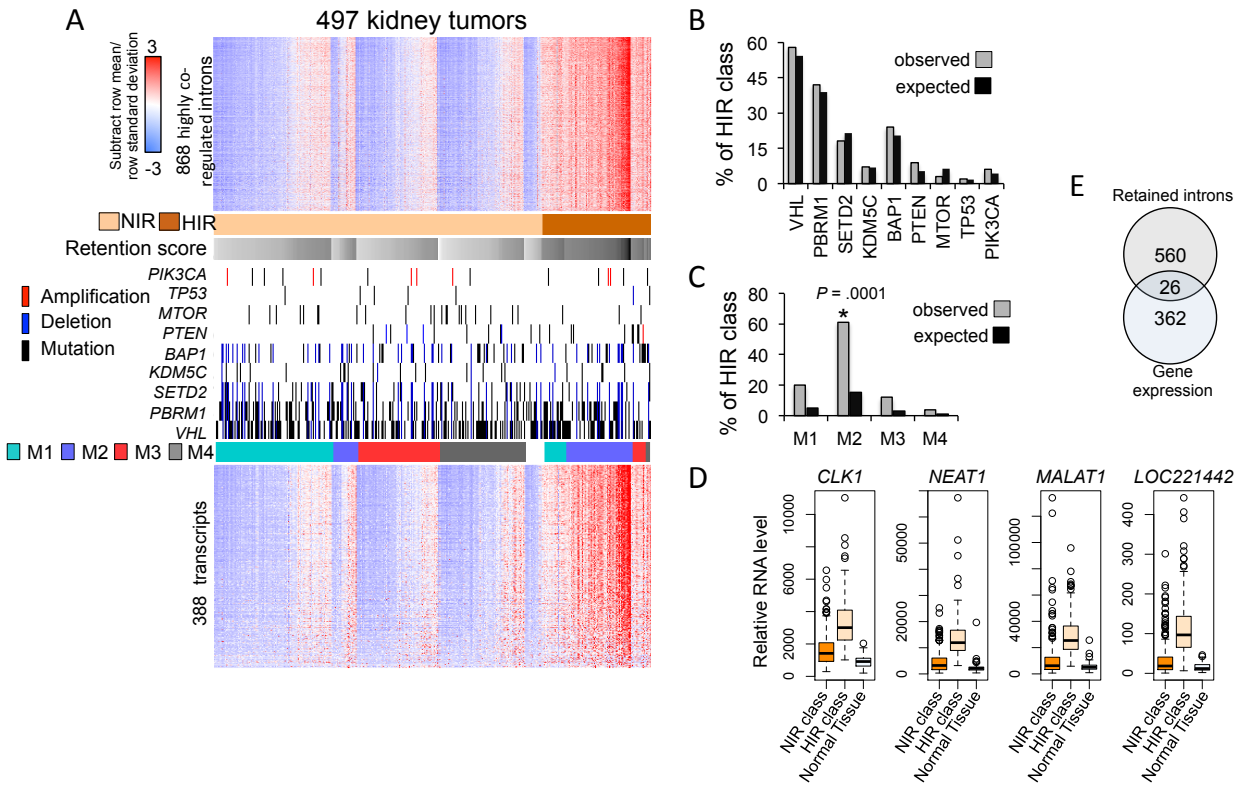

Figure 4.3 Mutation profiles of kidney cancer in the two defined classes and genes up-regulated in the HIR class

(A) Heatmap of PSI values of the 868 highly co-regulated intron retention events in 497 kidney tumors (Red to blue: high to low; the value was subtracted by the mean and normalized to the row standard deviation). The 497 kidney tumors were grouped into two classes first: NIR (light brown) and HIR (darker brown), and then clustered into four subgroups: M1-M4 cancer classes (cyan, slateblue, red and gray), and finally sorted by the retention score (from black to white: higher to lower score). The frequently mutated genes were labeled in different color boxes according to their mutation types in each tumor sample if existing. The bottom heatmaps are transcripts that are highly up-regulated in the HIR class. (B) The percentage of observed (grey) and expected (black) samples in HIR class with each frequently mutated genes. (C) The percentage of samples in HIR class that are classified in M1-M4 classes as observed (grey) and as expected (black). (D) Box plot of four example genes that are highly up-regulated in the HIR class compared with the NIR class and normal tissue. (E) Venn diagram of the overlapping genes among genes containing the highly co-regulated RIs and genes that are highly up-regulated in the HIR class.

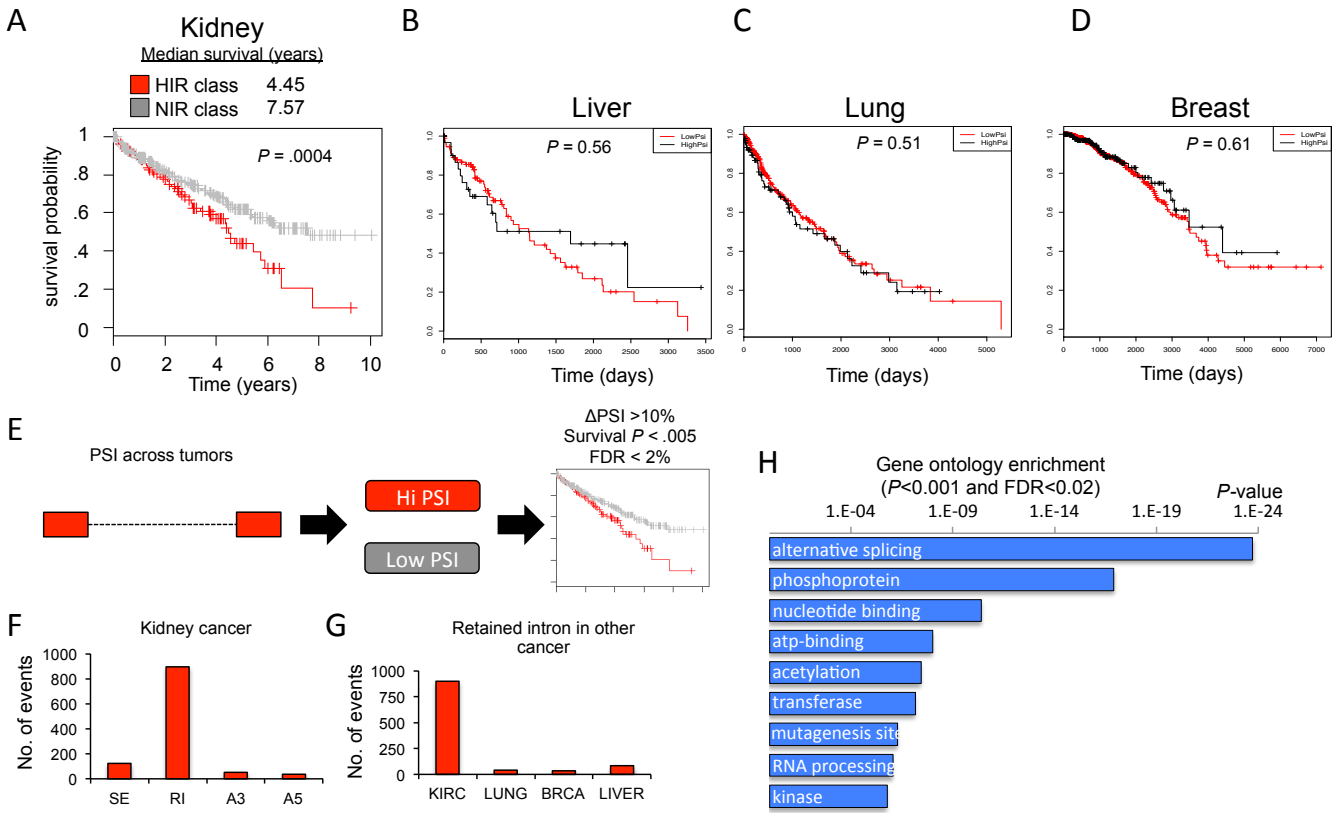

Figure 4.4 Using retention score as a predictor for patient survival

(A) The kidney tumor samples were separated into two groups according to their retention score: HIR and NIR classes. The group with higher score (i.e. more retained intron, HIR class) has significant shorter survival (median survival years: 4.45) than the group with lower score (i.e. less retained intron, NIR class, median survival years: 7.57) as shown in the Kaplan–Meier plot (p-value=0.0004). Same analyses were repeated for liver tumors (B), lung tumors (C) and breast tumors (D). (E) The flow chart of identifying potential predictor (i.e. AS event) for patient survival. Samples were separated into two groups according to each event's PSI value: high PSI (above the third quartile) and low PSI (below the third quartile). Then we tested if there is a significant difference between the two survival curves using the log rank test (p<0.005). (F) We repeated this procedure for every splicing event in kidney cancer and plotted the number of events with potential prediction power (i.e. significant difference between the two survival curves). (G) Same as (F), but repeated in retained intron events in other types of cancer. (H) Enriched gene ontology terms of genes containing RI events that are most significantly correlated with median survival differences.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

This study was set out to explore RNA binding proteins, alternative splicing, and their connection and mis-regulation in cancers. The main findings are chapter specific and we summarized them according to the respective chapters below.

5.1 Work summary/Important findings

In the part 1 of our study (chapter 2), using genome-wide sequence analysis of the proteins containing RNA recognition motif (RRM), we have found extensive RRM duplications in RNA binding proteins. This unexpected finding suggested a new model for the evolution of RNA binding proteins and provided important implications on their function. The analyses of various annotated genomes suggest that the number of proteins with canonical RBDs has expanded significantly in mammals. In addition, the number of proteins with multiple RNA binding domains has also increased. In the study, we seek to study the mechanism and functional consequence of such expansion. The novel and important points of this study are: 1) The RRMs within same RBP are more similar than what would be expected in random pairs. Such increased similarity was maintained even after various shuffles. 2) The sibling RRM pairs have even

higher sequence similarity in species-specific RBPs compared to the ancient RBPs shared by multiple species. Together these observations suggest an extensive RRM duplication in many mammalian proteins. 3) We presented two examples of RRM duplication: a recent RRM duplication found only in some primates and an ancient duplication shared from yeast to human. 4) The RBPs with different similarities between their sibling RRMs belong to distinct functional groups. 5) The RBP sequences outside the RNA binding domain are enriched with distinct low complexity domains that may be involved in protein interaction. In summary, this comprehensive study helps us better understand the evolution of RBPs and its functional implication.

We then move forward to study alternative splicing in cancer (chapter 3). With a systematic analysis of alternative splicing (AS) in cancer using the TCGA data, we have made a significant step forward in revealing the identity and function of alternative splicing events altered across multiple cancers. Prevalent splicing alteration is one molecular hallmark of cancer. However, previous analyses of cancer-associated splicing were focused on specific genes or single tumor type, which is often dominated by the noise from tissue specific splicing events. In the study we performed a transcriptome-wide splicing comparison between thousands of samples in three different cancer types and performed detailed analyses of the function and regulation of cancer-specific AS events. The novel and important points of this study are: 1) A core set of 163 cancer-specific AS events are identified, which is commonly found in multiple cancer types of different tissue origins. 2) Genes containing cancer-specific AS events are enriched for functions of cell cycle regulation, cell adhesion/migration, and insulin signaling pathway. 3) The cancer-specific AS events are more conserved and more likely to maintain reading frame to produce different protein isoforms with distinct functions. 4) The newly identified splicing events can be used as molecular biomarkers to distinguish tumors from normal samples, to separate different

cancer subtypes, and to predict cancer survival. 5) The genes whose expression is closely associated with cancer-specific AS events are mostly key regulators of cell cycle progression, revealing an unknown link between the cell cycle and alternative splicing in cancer. In addition to providing a global picture of cancer-specific AS, this comprehensive study helps us to better understand the functional consequence of and mechanism of splicing mis-regulation in cancer.

Through the analysis of 13 cancers, an interesting finding of retained introns in kidney cancer was identified, which leads to another extended study in chapter 4. With detailed sequence analysis, we found that some kidney tumors harbor a significant more retained introns in their mRNA transcripts. When further investigating other types of cancer, we did not find such elevation in intron retention, implying the phenomenon is specific to kidney cancer. The novel and important points of this work are: 1) Kidney tumors express a global intron retention which is specific to kidney cancers. 2) Kidney tumors with increased intron retention are more likely to have poor prognosis. We found those tumors have significant overlaps with recently defined kidney cancer subtype, M2, which is one of the most aggressive subtypes. 3) There is no clear link between such high intron retention and known cancer mutations through our statistical analysis. However, high gene expression of some lncRNAs and splicing factors are correlated with these elevated retained introns. 4) Intron retention in kidney cancer is highly correlated with patient survival, implying that global splicing patterns could be used as a predictor in kidney cancer progression. Together we have shown that intron retention patterns are molecular traits with predictive power in kidney tissues and that these tumors have adopted the splicing landscape to alter pathways that provide an advantage to tumors.

5.2 Weaknesses/limitations of the study and future direction

At this point, we have summarized the most important details and findings from previous chapters, and it is necessary to discuss the weakness of our studies and limitations of the approaches we used. Additionally, it is important to emphasize and outline the future work that needs to be done in extending the research described. The main points are chapter specific and we described them according to the respective chapters below.

5.2.1 Weaknesses/limitations in chapter 2 and future works

In chapter 2, we found extensive RRM duplications in RBPs, and describe a potential model for their evolution. However, our study only covers about half of the RBPs in human. Although RRM-containing RBP is the largest group, in terms of protein number, among all human RBPs, there are still other RNA binding domains/motifs playing important roles in alternative splicing regulation, such as RS domains in serine-arginine (SR) proteins which are often the functional sequence interacting with other proteins and K homology (KH) domains in hnRNP proteins which can function in RNA recognition. We chose RRM domains because it is the most prevalent domain, however it will only make our study more completed when we include other RBDs that play important roles in the splicing regulation.

In the end of chapter 2, we have provided important implications of functional correlation of the domain duplication and our results suggest that multiple RRMs allow a protein to bind RNA with higher sequence specificity and/or affinity than those RBPs with a single binding domain. In the future work, we would like to identify RNA binding targets of the RBPs we are interested in (e.g. Splicing factors). Now, with the availability of many crosslinking immunoprecipitation (CLIP)-seq data, which have revealed transcriptome-wide binding sites of RBPs at the single-nucleotide level. We can use them to identify the RBP binding targets at a

high-throughput manner. A recently published work has constructed a public available database, CLIPdb, which combines results from about 400 CLIP-seq datasets [146]. We will use data from the CLIPdb to explore the potential binding targets of our interested splicing factors.

5.2.2 Weaknesses/limitations in chapter 3 and future works

In chapter 3, we used MISO as our splicing annotation tools when analyzing RNA-seq data, however there are some limitations in our analyzing pipeline. First, only the annotated splicing events included in the MISO annotation database will be identified, those novel splicing events in the sample will never been detected. Future work in this part of the study will be trying different statistical models to detect differential transcription, which do not depend on a pre-defined annotation database. For example both FDM [147] and EBseq [148] identify differential expressed splicing isorforms in an RNA-seq experiment. Using these methods we expect to identify many more novel splicing events that were not included in the MISO annotation database in the cancer samples. Secondly, we have analyzed the SE, RI, A3SS and A5SS splicing events among the cancer samples, but there are also alternative first exon (AFE), alternative last exon (ALE), tandem un-translated region (TandemUTR) and mutually exclusive exons (MXE) events that are also important, especially the TandemUTR events that have been shown to affect translation in tumor cells [141, 142]. Therefore including those types of splicing events in our study can help us identify more potential interesting splicing change in cancers. Finally, in our study, we have chose a filtering criteria to identify cancer-specific AS events that we think it's good for the purpose of this study. However there is always no perfect filtering, in terms of balancing the sensitivity and specificity. In our results, there are still some known splicing events not picked up by our method due to the stringent requirement we have setup. In the future work, we will tune our filtering criteria to increase the sensitivity for some low

abundant cancer-specific events.

5.2.3 Weaknesses/limitations in chapter 4 and future works

In chapter 4, we have shown that intron retentions are widespread in kidney cancer and it is highly correlated with patient survival. However we only analyzed kidney tumor RNA-seq data with the annotated MISO database which includes only the known retained introns. It is fair to speculate that this phenomenon could happen globally in the splicing machinery and many other previously undetected introns are also retained in kidney cancer. Future work in this part of the study will be focusing on checking whether the intron retentions actually happen more than what we have found. In addition, we would like to further investigate the 388 transcripts that are highly expressed in the HIR class. We have shown that there are some splicing factors and lncRNAs within the highly expressed transcripts. For some of them, we would like to do experiments to test the connection between their gene expression level and intron retention rate in kidney cancer cell.

Characterization of cancer splicing and how it is regulated is still far from complete. The work we presented here has provided a start point and some clues that how splicing alteration link to tumors. Together, our studies advance the understanding of splicing dysregulation in cancer in multiple aspects. Though much work will be needed, our works have provided many insights for cancer diagnosis and therapy.

BIBLIOGRAPHY

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.

2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.

3. Ponting CP, Hardison RC: **What fraction of the human genome is functional?** *Genome Res* 2011, **21:**1769-1776.

4. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al: **Defining functional DNA elements in the human genome.** *Proc Natl Acad Sci U S A* 2014, **111:**6131-6138.

5. **DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)** [http://www.genome.gov/sequencingcosts]

6. Consortium EP: **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306:**636-640.

7. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061-1073.

8. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project.** *Nat Genet* 2013, **45:**1113-1120.

9. Nagalakshmi U, Waern K, Snyder M: **RNA-Seq: a method for comprehensive transcriptome analysis.** *Curr Protoc Mol Biol* 2010, **Chapter 4:**Unit 4 11 11-13.

10. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456:**470-476.

11. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.** *Nat Genet* 2008, **40:**1413-1415.

12. Cascino I, Fiucci G, Papoff G, Ruberti G: **Three functional soluble forms of the human apoptosis-inducing Fas molecule are produced by alternative splicing.** *J Immunol* 1995, **154:**2706-2713.

13. Demir E, Dickson BJ: **fruitless splicing specifies male courtship behavior in Drosophila.** *Cell* 2005, **121:**785-794.

14. Biamonti G, Catillo M, Pignataro D, Montecucco A, Ghigna C: **The alternative splicing side of cancer.** *Semin Cell Dev Biol* 2014, **32C:**30-36.

15. Venables JP: **Aberrant and alternative splicing in cancer.** *Cancer Res* 2004, **64:**7647-7654.

16. Lim LP, Burge CB: **A computational analysis of sequence features involved in recognition of short introns.** *Proc Natl Acad Sci U S A* 2001, **98:**11193-11198.

17. Black DL: **Mechanisms of alternative pre-messenger RNA splicing.** *Annu Rev Biochem* 2003, **72:**291-336.

18. Wang Z, Burge CB: **Splicing regulation: from a parts list of regulatory elements to an integrated splicing code.** *RNA* 2008, **14:**802-813.

19. McCullough AJ, Berget SM: **G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection.** *Mol Cell Biol* 1997, **17:**4562-4571.

20. Wang Y, Xiao X, Zhang J, Choudhury R, Robertson A, Li K, Ma M, Burge CB, Wang Z: **A complex network of factors with overlapping affinities represses splicing through intronic elements.** *Nat Struct Mol Biol* 2013, **20:**36-45.

21. Wang Y, Ma M, Xiao X, Wang Z: **Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules.** *Nat Struct Mol Biol* 2012, **19:**1044-1052.

22. Konarska MM, Query CC: **Insights into the mechanisms of splicing: more lessons from the ribosome.** *Genes Dev* 2005, **19:**2255-2260.

23. Matlin AJ, Clark F, Smith CW: **Understanding alternative splicing: towards a cellular code.** *Nat Rev Mol Cell Biol* 2005, **6:**386-398.

24. Blencowe BJ: **Alternative splicing: new insights from global analyses.** *Cell* 2006, **126:**37-47.

25. House AE, Lynch KW: **Regulation of alternative splicing: more than just the ABCs.** *J Biol Chem* 2008, **283:**1217-1221.

26. Janga SC, Mittal N: **Construction, structure and dynamics of post-transcriptional regulatory network directed by RNA-binding proteins.** *Adv Exp Med Biol* 2011, **722:**103-117.

27. Lukong KE, Chang KW, Khandjian EW, Richard S: **RNA-binding proteins in human genetic disease.** *Trends Genet* 2008, **24:**416-425.

28. Tolino M, Kohrmann M, Kiebler MA: **RNA-binding proteins involved in RNA localization and their implications in neuronal diseases.** *Eur J Neurosci* 2012, **35:**1818-1836.

29. King OD, Gitler AD, Shorter J: **The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease.** *Brain Res* 2012, **1462:**61-80.

30. Lunde BM, Moore C, Varani G: **RNA-binding proteins: modular design for efficient function.** *Nat Rev Mol Cell Biol* 2007, **8:**479-490.

31. Maris C, Dominguez C, Allain FH: **The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression.** *FEBS J* 2005, **272:**2118-2131.

32. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, et al: **Insights into RNA biology from an atlas of mammalian mRNA-binding proteins.** *Cell* 2012, **149:**1393-1406.

33. Dembowski JA, Grabowski PJ: **The CUGBP2 splicing factor regulates an ensemble of branchpoints from perimeter binding sites with implications for autoregulation.** *PLoS Genet* 2009, **5:**e1000595.

34. Goo YH, Cooper TA: **CUGBP2 directly interacts with U2 17S snRNP components and promotes U2 snRNA binding to cardiac troponin T pre-mRNA.** *Nucleic Acids Res* 2009, **37:**4275-4286.

35. Zhu J, Mayeda A, Krainer AR: **Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins.** *Mol Cell* 2001, **8:**1351-1361.

36. Talukdar I, Sen S, Urbano R, Thompson J, Yates JR, 3rd, Webster NJ: **hnRNP A1 and hnRNP F modulate the alternative splicing of exon 11 of the insulin receptor gene.** *PLoS One* 2011, **6:**e27869.

37. Caceres JF, Stamm S, Helfman DM, Krainer AR: **Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors.** *Science* 1994, **265:**1706-1709.

38. Caputi M, Mayeda A, Krainer AR, Zahler AM: **hnRNP A/B proteins are required for inhibition of HIV-1 pre-mRNA splicing.** *EMBO J* 1999, **18:**4060-4067.

39. Tang YH, Han SP, Kassahn KS, Skarshewski A, Rothnagel JA, Smith R: **Complex evolutionary relationships among four classes of modular RNA-binding splicing regulators in eukaryotes: the hnRNP, SR, ELAV-like and CELF proteins.** *J Mol Evol* 2012, **75:**214-228.

40. Martinez-Contreras R, Cloutier P, Shkreta L, Fisette JF, Revil T, Chabot B: **hnRNP proteins and splicing control.** *Adv Exp Med Biol* 2007, **623:**123-147.

41. Busch A, Hertel KJ: **Evolution of SR protein and hnRNP splicing regulatory factors.** *Wiley Interdiscip Rev RNA* 2012, **3:**1-12.

42. Pascale A, Govoni S: **The complex world of post-transcriptional mechanisms: is their deregulation a common link for diseases? Focus on ELAV-like RNA-binding proteins.** *Cell Mol Life Sci* 2012, **69:**501-517.

43. Dasgupta T, Ladd AN: **The importance of CELF control: molecular and biological roles of the CUG-BP, Elav-like family of RNA-binding proteins.** *Wiley Interdiscip Rev RNA* 2012, **3:**104-121.

44. Jin W, Bruno IG, Xie TX, Sanger LJ, Cote GJ: **Polypyrimidine tract-binding protein down-regulates fibroblast growth factor receptor 1 alpha-exon inclusion.** *Cancer Res* 2003, **63:**6154-6157.

45. Fregoso OI, Das S, Akerman M, Krainer AR: **Splicing-factor oncoprotein SRSF1 stabilizes p53 via RPL5 and induces cellular senescence.** *Mol Cell* 2013, **50:**56-66.

46. Karni R, de Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR: **The gene encoding the splicing factor SF2/ASF is a proto-oncogene.** *Nat Struct Mol Biol* 2007, **14:**185-193.

47. Golan-Gerstl R, Cohen M, Shilo A, Suh SS, Bakacs A, Coppola L, Karni R: **Splicing factor hnRNP A2/B1 regulates tumor suppressor gene splicing and is an oncogenic driver in glioblastoma.** *Cancer Res* 2011, **71:**4464-4472.

48. Oh JJ, Razfar A, Delgado I, Reed RA, Malkina A, Boctor B, Slamon DJ: **3p21.3 tumor suppressor gene H37/Luca15/RBM5 inhibits growth of human lung cancer cells through cell cycle arrest and apoptosis.** *Cancer Res* 2006, **66:**3419-3427.

49. Wang Y, Chen D, H. Q, Tsai YS, Shao S, Liu Q, Dominguez D, Wang Z: **The splicing factor RBM4 controls apoptosis, proliferation, and migration to suppress tumor progression.** *Cancer Cell* 2014, **26:**374-389.

50. Millevoi S, Bernat S, Telly D, Fouque F, Gladieff L, Favre G, Vagner S, Toulas C: **The c.5242C>A BRCA1 missense variant induces exon skipping by increasing splicing repressors binding.** *Breast Cancer Res Treat* 2010, **120:**391-399.

51. Kim YJ, Kim HS: **Alternative splicing and its impact as a cancer diagnostic marker.** *Genomics Inform* 2012, **10:**74-80.

52. Omenn GS, Menon R, Zhang Y: **Innovations in proteomic profiling of cancers: alternative splice variants as a new class of cancer biomarker candidates and bridging of proteomics with structural biology.** *J Proteomics* 2013, **90:**28-37.

53. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al: **A compendium of RNA-binding motifs for decoding gene regulation.** *Nature* 2013, **499:**172-177.

54. Hamy F, Asseline U, Grasby J, Iwai S, Pritchard C, Slim G, Butler PJ, Karn J, Gait MJ: **Hydrogen-bonding contacts in the major groove are required for human immunodeficiency virus type-1 tat protein recognition of TAR RNA.** *J Mol Biol* 1993, **230:**111-123.

55. Shen H, Kan JL, Green MR: **Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly.** *Mol Cell* 2004, **13:**367-376.

56. Auweter SD, Oberstrass FC, Allain FH: **Sequence-specific binding of single-stranded RNA: is there a code for recognition?** *Nucleic Acids Res* 2006, **34:**4943-4959.

57. Birney E, Kumar S, Krainer AR: **Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors.** *Nucleic Acids Res* 1993, **21:**5803-5816.

58. Bentley RC, Keene JD: **Recognition of U1 and U2 small nuclear RNAs can be altered by a 5-amino-acid segment in the U2 small nuclear ribonucleoprotein particle (snRNP) B" protein and through interactions with U2 snRNP-A' protein.** *Mol Cell Biol* 1991, **11:**1829-1839.

59. Caceres JF, Krainer AR: **Functional analysis of pre-mRNA splicing factor SF2/ASF structural domains.** *EMBO J* 1993, **12:**4715-4726.

60. Oberstrass FC, Auweter SD, Erat M, Hargous Y, Henning A, Wenter P, Reymond L, Amir-Ahmady B, Pitsch S, Black DL, Allain FH: **Structure of PTB bound to RNA: specific binding and implications for splicing regulation.** *Science* 2005, **309:**2054-2057.

61. Hargous Y, Hautbergue GM, Tintaru AM, Skrisovska L, Golovanov AP, Stevenin J, Lian LY, Wilson SA, Allain FH: **Molecular basis of RNA recognition and TAP binding by the SR proteins SRp20 and 9G8.** *EMBO J* 2006, **25:**5126-5137.

62. Sickmier EA, Frato KE, Shen H, Paranawithana SR, Green MR, Kielkopf CL: **Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65.** *Mol Cell* 2006, **23:**49-59.

63. Dominguez C, Allain FH: **NMR structure of the three quasi RNA recognition motifs (qRRMs) of human hnRNP F and interaction studies with Bcl-x G-tract RNA: a novel mode of RNA recognition.** *Nucleic Acids Res* 2006, **34:**3634-3645.

64. Dominguez C, Fisette JF, Chabot B, Allain FH: **Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs.** *Nat Struct Mol Biol* 2010, **17:**853-861.

65. Burd CG, Matunis EL, Dreyfuss G: **The multiple RNA-binding domains of the mRNA poly(A)-binding protein have different RNA-binding activities.** *Mol Cell Biol* 1991, **11:**3419-3424.

66. Safaee N, Kozlov G, Noronha AM, Xie J, Wilds CJ, Gehring K: **Interdomain allostery promotes assembly of the poly(A) mRNA complex with PABP and eIF4G.** *Mol Cell* 2012, **48:**375-386.

67. Bjorklund AK, Ekman D, Elofsson A: **Expansion of protein domain repeats.** *PLoS Comput Biol* 2006, **2:**e114.

68. Xu H, Li Z, Li M, Wang L, Hong Y: **Boule is present in fish and bisexually expressed in adult and embryonic germ cells of medaka.** *PLoS One* 2009, **4:**e6097.

69. Eirin-Lopez JM, Ausio J: **Boule and the Evolutionary Origin of Metazoan Gametogenesis: A Grandpa's Tale.** *Int J Evol Biol* 2011, **2011:**972457.

70. Li M, Shen Q, Xu H, Wong FM, Cui J, Li Z, Hong N, Wang L, Zhao H, Ma B, Hong Y: **Differential conservation and divergence of fertility genes boule and dazl in the rainbow trout.** *PLoS One* 2011, **6:**e15910.

71. Kimura M: **The neutral theory of molecular evolution and the world view of the neutralists.** *Genome* 1989, **31:**24-31.

72. Zhao Z, Fu YX, Hewett-Emmett D, Boerwinkle E: **Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution.** *Gene* 2003, **312:**207-213.

73. Goss DJ, Kleiman FE: **Poly(A) binding proteins: are they all created equal?** *Wiley Interdiscip Rev RNA* 2013, **4:**167-179.

74. Huang da W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res* 2009, **37:**1-13.

75. Zimmer M, Ebert BL, Neil C, Brenner K, Papaioannou I, Melas A, Tolliday N, Lamb J, Pantopoulos K, Golub T, Iliopoulos O: **Small-molecule inhibitors of HIF-2a translation link its 5'UTR iron-responsive element to oxygen sensing.** *Mol Cell* 2008, **32:**838-848.

76. Barreau C, Paillard L, Osborne HB: **AU-rich elements and associated factors: are there unifying principles?** *Nucleic Acids Res* 2005, **33:**7138-7150.

77. Matoulkova E, Michalova E, Vojtesek B, Hrstka R: **The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells.** *RNA Biol* 2012, **9:**563-576.

78. Shen H, Green MR: **RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans.** *Genes Dev* 2006, **20:**1755-1765.

79. Kato M, Han TW, Xie S, Shi K, Du X, Wu LC, Mirzaei H, Goldsmith EJ, Longgood J, Pei J, et al: **Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels.** *Cell* 2012, **149:**753-767.

80. Thomas EE: **Short, local duplications in eukaryotic genomes.** *Curr Opin Genet Dev* 2005, **15:**640-644.

81. Jurka J, Kapitonov VV, Kohany O, Jurka MV: **Repetitive sequences in complex genomes: structure and evolution.** *Annu Rev Genomics Hum Genet* 2007, **8:**241-259.

82. O'Brien KP, Remm M, Sonnhammer EL: **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucleic Acids Res* 2005, **33:**D476-480.

83. Oltean S, Bates DO: **Hallmarks of alternative splicing in cancer.** *Oncogene* 2013.

84. Zhang J, Manley JL: **Misregulation of pre-mRNA alternative splicing in cancer.** *Cancer discovery* 2013, **3:**1228-1237.

85. Naor D, Nedvetzki S, Golan I, Melnik L, Faitelson Y: **CD44 in cancer.** *Crit Rev Clin Lab Sci* 2002, **39:**527-579.

86. Ishimoto T, Nagano O, Yae T, Tamada M, Motohara T, Oshima H, Oshima M, Ikeda T, Asaba R, Yagi H, et al: **CD44 variant regulates redox status in cancer cells by stabilizing the xCT subunit of system xc(-) and thereby promotes tumor growth.** *Cancer Cell* 2011, **19:**387-400.

87. Yae T, Tsuchihashi K, Ishimoto T, Motohara T, Yoshikawa M, Yoshida GJ, Wada T, Masuko T, Mogushi K, Tanaka H, et al: **Alternative splicing of CD44 mRNA by ESRP1 enhances lung colonization of metastatic cancer cell.** *Nat Commun* 2012, **3:**883.

88. Brown RL, Reinke LM, Damerow MS, Perez D, Chodosh LA, Yang J, Cheng C: **CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression.** *The Journal of clinical investigation* 2011, **121:**1064-1074.

89. Adams JM, Cory S: **The Bcl-2 apoptotic switch in cancer development and therapy.** *Oncogene* 2007, **26:**1324-1337.

90. Matera AG, Wang Z: **A day in the life of the spliceosome.** *Nature reviews Molecular cell biology* 2014, **15:**108-121.

91. Choudhury R, Roy SG, Tsai YS, Tripathy A, Graves LM, Wang Z: **The splicing activator DAZAP1 integrates splicing control into MEK/Erk-regulated cell proliferation and migration.** *Nature communications* 2014, **5:**3078.

92. Al-Ayoubi AM, Zheng H, Liu Y, Bai T, Eblen ST: **Mitogen-activated protein kinase phosphorylation of splicing factor 45 (SPF45) regulates SPF45 alternative splicing site utilization, proliferation, and cell adhesion.** *Molecular and cellular biology* 2012, **32:**2880-2893.

93. David CJ, Chen M, Assanah M, Canoll P, Manley JL: **HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer.** *Nature* 2010, **463:**364-368.

94. Kornblihtt AR, Schor IE, Allo M, Dujardin G, Petrillo E, Munoz MJ: **Alternative splicing: a pivotal step between eukaryotic transcription and translation.** *Nature reviews Molecular cell biology* 2013, **14:**153-165.

95. Liu J, Lee W, Jiang Z, Chen Z, Jhunjhunwala S, Haverty PM, Gnad F, Guan Y, Gilbert HN, Stinson J, et al: **Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events.** *Genome Res* 2012, **22:**2315-2327.

96. Eswaran J, Horvath A, Godbole S, Reddy SD, Mudvari P, Ohshiro K, Cyanam D, Nair S, Fuqua SA, Polyak K, et al: **RNA sequencing of cancer reveals novel splicing alterations.** *Sci Rep* 2013, **3:**1689.

97. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, et al: **Mutational landscape and significance across 12 major cancer types.** *Nature* 2013, **502:**333-339.

98. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G: **Discovery and saturation analysis of cancer genes across 21 tumour types.** *Nature* 2014, **505:**495-501.

99. Kervestin S, Jacobson A: **NMD: a multifaceted response to premature translational termination.** *Nature reviews Molecular cell biology* 2012, **13:**700-712.

100. Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks.** *BMC Bioinformatics* 2003, **4:**2.

101. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4:**44-57.

102. Lefave CV, Squatrito M, Vorlova S, Rocco GL, Brennan CW, Holland EC, Pan YX, Cartegni L: **Splicing factor hnRNPH drives an oncogenic splicing switch in gliomas.** *The EMBO journal* 2011, **30:**4084-4097.

103. Stark M, Bram EE, Akerman M, Mandel-Gutfreund Y, Assaraf YG: **Heterogeneous nuclear ribonucleoprotein H1/H2-dependent unsplicing of thymidine phosphorylase results in anticancer drug resistance.** *The Journal of biological chemistry* 2011, **286:**3741-3754.

104. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al: **Supervised risk predictor of breast cancer based on intrinsic subtypes.** *J Clin Oncol* 2009, **27:**1160-1167.

105. Prat A, Perou CM: **Deconstructing the molecular portraits of breast cancer.** *Molecular oncology* 2011, **5:**5-23.

106. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, et al: **A gene-expression signature as a predictor of survival in breast cancer.** *The New England journal of medicine* 2002, **347:**1999-2009.

107. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, et al: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415:**530-536.

108. Bednarek AK, Laflin KJ, Daniel RL, Liao Q, Hawkins KA, Aldaz CM: **WWOX, a novel WW domain-containing protein mapping to human chromosome 16q23.3-24.1, a region frequently affected in breast cancer.** *Cancer Res* 2000, **60:**2140-2145.

109. Tang X, Jin R, Qu G, Wang X, Li Z, Yuan Z, Zhao C, Siwko S, Shi T, Wang P, et al: **GPR116, an adhesion G-protein-coupled receptor, promotes breast cancer metastasis via the Galphaq-p63RhoGEF-Rho GTPase pathway.** *Cancer Res* 2013, **73:**6206-6218.

110. Stulberg J, Kamel-Reid S, Chun K, Tokunaga J, Wells RA: **Molecular analysis of a new variant of the CBF beta-MYH11 gene fusion.** *Leuk Lymphoma* 2002, **43:**2021-2026.

111. Costello R, Sainty D, Lecine P, Cusenier A, Mozziconacci MJ, Arnoulet C, Maraninchi D, Gastaut JA, Imbert J, Lafage-Pochitaloff M, Gabert J: **Detection of CBFbeta/MYH11 fusion transcripts in acute myeloid leukemia: heterogeneity of cytological and molecular characteristics.** *Leukemia* 1997, **11:**644-650.

112. Langer W, Sohler F, Leder G, Beckmann G, Seidel H, Grone J, Hummel M, Sommer A: **Exon array analysis using re-defined probe sets results in reliable identification of alternatively spliced genes in non-small cell lung cancer.** *BMC Genomics* 2010, **11:**676.

113. Zong FY, Fu X, Wei WJ, Luo YG, Heiner M, Cao LJ, Fang Z, Fang R, Lu D, Ji H, Hui J: **The RNA-binding protein QKI suppresses cancer-associated aberrant splicing.** *PLoS genetics* 2014, **10:**e1004289.

114. Bechara EG, Sebestyen E, Bernardis I, Eyras E, Valcarcel J: **RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation.** *Molecular cell* 2013, **52:**720-733.

115. Rybak JN, Roesli C, Kaspar M, Villa A, Neri D: **The extra-domain A of fibronectin is a vascular marker of solid tumors and metastases.** *Cancer Res* 2007, **67:**10948-10957.

116. Wang Y, Cheong CG, Hall TM, Wang Z: **Engineering splicing factors with designed specificities.** *Nat Methods* 2009, **6:**825-830.

117. Mercatante DR, Bortner CD, Cidlowski JA, Kole R: **Modification of alternative splicing of Bcl-x pre-mRNA in prostate and breast cancer cells. analysis of apoptosis and cell death.** *The Journal of biological chemistry* 2001, **276:**16411-16417.

118. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D: **Identification of genes periodically expressed in the human cell cycle and their expression in tumors.** *Mol Biol Cell* 2002, **13:**1977-2000.

119. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, et al: **MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.** *Nucleic Acids Res* 2010, **38:**e178.

120. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC bioinformatics* 2011, **12:**323.

121. Bullard JH, Purdom E, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.** *BMC bioinformatics* 2010, **11:**94.

122. Katz Y, Wang ET, Airoldi EM, Burge CB: **Analysis and design of RNA sequencing experiments for identifying isoform regulation.** *Nat Methods* 2010, **7:**1009-1015.

123. Snel B, Lehmann G, Bork P, Huynen MA: **STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene.** *Nucleic Acids Res* 2000, **28:**3442-3444.

124. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ: **STRING v9.1: protein-protein interaction networks, with increased coverage and integration.** *Nucleic Acids Res* 2013, **41:**D808-815.

125. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.

126. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15:**1034-1050.

127. Fairbrother WG, Yeh RF, Sharp PA, Burge CB: **Predictive identification of exonic splicing enhancers in human genes.** *Science* 2002, **297:**1007-1013.

128. Bonomi S, Gallo S, Catillo M, Pignataro D, Biamonti G, Ghigna C: **Oncogenic alternative splicing switches: role in cancer progression and prospects for therapy.** *International journal of cell biology* 2013, **2013:**962038.

129. Fu Y, Masuda A, Ito M, Shinmi J, Ohno K: **AG-dependent 3'-splice sites are predisposed to aberrant splicing due to a mutation at the first nucleotide of an exon.** *Nucleic Acids Res* 2011, **39:**4396-4404.

130. Shao C, Yang B, Wu T, Huang J, Tang P, Zhou Y, Zhou J, Qiu J, Jiang L, Li H, et al: **Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome.** *Nat Struct Mol Biol* 2014, **21:**997-1005.

131. Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, Pellagatti A, Wainscoat JS, Hellstrom-Lindberg E, Gambacorti-Passerini C, et al: **Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts.** *N Engl J Med* 2011, **365:**1384-1395.

132. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, et al: **Frequent pathway mutations of splicing machinery in myelodysplasia.** *Nature* 2011, **478:**64-69.

133. Gatto F, Nookaew I, Nielsen J: **Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma.** *Proc Natl Acad Sci U S A* 2014, **111:**E866-875.

134. Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, Fisher R, McGranahan N, Matthews N, Santos CR, et al: **Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing.** *Nat Genet* 2014, **46:**225-233.

135. Dalgliesh GL, Furge K, Greenman C, Chen L, Bignell G, Butler A, Davies H, Edkins S, Hardy C, Latimer C, et al: **Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes.** *Nature* 2010, **463:**360-363.

136. Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, Davies H, Jones D, Lin ML, Teague J, et al: **Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma.** *Nature* 2011, **469:**539-542.

137. Guo G, Gui Y, Gao S, Tang A, Hu X, Huang Y, Jia W, Li Z, He M, Sun L, et al: **Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma.** *Nat Genet* 2012, **44:**17-19.

138. Simon JM, Hacker KE, Singh D, Brannon AR, Parker JS, Weiser M, Ho TH, Kuan PF, Jonasch E, Furey TS, et al: **Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects.** *Genome Res* 2014, **24:**241-250.

139. Boutz PL, Bhutkar A, Sharp PA: **Detained introns are a novel, widespread class of post-transcriptionally spliced introns.** *Genes Dev* 2015, **29:**63-80.

140. Cancer Genome Atlas Research N: **Comprehensive molecular characterization of clear cell renal cell carcinoma.** *Nature* 2013, **499:**43-49.

141. Mayr C, Bartel DP: **Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.** *Cell* 2009, **138:**673-684.

142. Lin Y, Li Z, Ozsolak F, Kim SW, Arango-Argoty G, Liu TT, Tenenbaum SA, Bailey T, Monaghan AP, Milos PM, John B: **An in-depth map of polyadenylation sites in cancer.** *Nucleic Acids Res* 2012, **40:**8460-8471.

143. Calin GA, Croce CM: **MicroRNA signatures in human cancers.** *Nat Rev Cancer* 2006, **6:**857-866.

144. Esteller M: **Non-coding RNAs in human disease.** *Nat Rev Genet* 2011, **12:**861-874.

145. Wang KC, Chang HY: **Molecular mechanisms of long noncoding RNAs.** *Mol Cell* 2011, **43:**904-914.

146. Yang YC, Di C, Hu B, Zhou M, Liu Y, Song N, Li Y, Umetsu J, Lu Z: **CLIPdb: a CLIP-seq database for protein-RNA interactions.** *BMC Genomics* 2015, **16:**51.

147. Singh D, Orellana CF, Hu Y, Jones CD, Liu Y, Chiang DY, Liu J, Prins JF: **FDM: a graph-based statistical method to detect differential transcription using RNA-seq data.** *Bioinformatics* 2011, **27:**2633-2640.

148. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C: **EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments.** *Bioinformatics* 2013, **29:**1035-1043.