UCLA

UCLA Electronic Theses and Dissertations

Title

Identification and Characterization of Targets for Splicing Factors, Polypyrimidine Tract Binding Proteins

Permalink

https://escholarship.org/uc/item/3wb8h9gt

Author

Han, A Reum

Publication Date

2013

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Identification and Characterization of Targets for Splicing Factors, Polypyrimidine Tract Binding Proteins

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Biomedical Engineering

by

A Reum Han

2013

ABSTRACT OF THE DISSERTATION

Identification and Characterization of Targets for Splicing Factors,

Polypyrimidine Tract Binding Proteins

by

A Reum Han

Doctor of Philosophy in Biomedical Engineering

University of California, Los Angeles, 2013

Professor Douglas L. Black, Chair

Splicing is an essential step in gene regulation to produce mature mRNA from a messenger RNA precursor. Through splicing, exons are assembled to generate a mature mRNA that can be translated into a desired protein by the cell. In mammals, splicing is extensively regulated by splicing factors, resulting in diverse exon arrangements and alternative splicing patterns. To know how splicing factors control genes through alternative splicing, it is critical to understand how they recognize their RNA and exon targets. In this study, we investigated targets of a family of strong splicing regulators, Polypyrimidine tract-binding (PTB) proteins. PTB proteins bind to RNA using their RNA binding domains (RBDs) and regulate splicing of exons. Among PTB proteins, PTBP1 is the most abundant and well-studied gene. Each RBD of PTBP1 binds to short, degenerate pyrimidine sequences allowing PTBP1 regulate a wide variety of targets. This variation makes it difficult to evaluate PTBP1 binding events and thus to identify PTBP1 target exons. We developed computational models that predict the binding and splicing targets of PTBP1. Models identify many previously unrecognized PTBP1 binding sites and novel exon targets. Encouraged by predictive PTBP1 models, we expanded these models to PTBP2, a neuronal paralog of PTBP1. PTBP1 is expressed in non-neuronal and neuronal progenitor cells down-regulating PTBP2. Upon neuronal differentiation, expression of PTBP1 decreases, and PTBP2 is up regulated. To understand why cells require switching between such similar genes, we compared targets for PTBP1 and PTBP2 on a global scale. We were able to assess redundancy and divergence of their binding and splicing codes. With the advance of highthrouput technologies, there are on going efforts to profile PTB-regulated transcriptomes in different tissues and cellular contexts. Integration of multiple datasets will highlight and prioritize direct targets of PTB proteins for future study. To this end, we integrated in vivo targets of PTBP2 from mouse mutant studies and identified potential direct targets. Our approaches can be applied to other splicing factors, and would be especially useful for studying multi-RBD proteins. We expect this study provides a general framework for the analysis of splicing factor binding and identification of functional targets.

The dissertation of A Reum Han is approved.

Ren Sun

Matteo Pellegrini

Christopher J. Lee

Douglas L. Black, Committee Chair

University of California, Los Angeles

2013

This dissertation is dedicated to my family, friends, and mentors, who supported me during my journey to Ph.D.

This happy journey was impossible without their love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I thank committee members, colleagues, and all members of the Black laboratory for thoughtful suggestions on the experiments, the analysis and writing.

Chapter 2 is a manuscript version of an article that has been submitted for publication: Areum Han, Peter Stoilov, Yu Zhou, Xiang-Dong Fu, and Douglas L. Black (2013). *De Novo* Prediction of PTBP1 Binding and Splicing Targets Reveals Unexpected Features of Its RNA Recognition and Function. *Submitted.*

Chapter 3 is a partial manuscript version of an article that has been in preparation for publication: Areum Han, Niroshika Keppetipola, and Douglas L. Black (2013) PTBP1 and its neuronal paralog PTBP2 have redundant binding codes but divergent splicing codes. *In preparation*.

Chapter 4 is a partial manuscript version of an article that has been in preparation for publication: Qin Li, Chia-ho Lin, Peter Stoilov, Lily Shiue, Areum Han, Manuel Ares Jr., Xiang-Dong Fu, and Douglas L. Black (2013), The splicing regulator PTBP2 maintains embryonic splicing patterns required for neuronal maturation. *In preparation*.

This work was supported by NIH R01 GM49662 and an associated ARRA supplement to D.L. Black. D.L. Black is an investigator of the Howard Hughes Medical Institute.

VITA

2006 - 2013	Graduate Student (Biomedical engineering) University of California, Los Angeles, U.S.A.
2005 - 2006	Researcher, Korea Research Institute of Bioscience and Biotechnology, South Korea.
2003 - 2005	M.S. (Biosystems), Korea Advanced Institute of Science and Technology, South Korea.
1998 - 2002	B.E. (Biomedical engineering), Kyung Hee University, South Korea.

PUBLICATIONS AND PRESENTATIONS

Areum Han, Peter Stoilov, Yu Zhou, Xiang-Dong Fu, and Douglas L. Black (July 2012), *De novo prediction of PTB binding and splicing targets from pre-mRNA sequences*, Talk presented at the ISMB Alternative Splicing Special Interest Group meeting (AS-SIG), Long Beach, U.S.A.

Areum Han, Peter Stoilov, Yu Zhou, Xiang-Dong Fu, and Douglas L. Black (June 2011), *De novo prediction of PTB (Polypyrimidine Tract Binding protein) binding and splicing targets from pre-mRNA sequences*, Talk presented at the annual RNA society meeting, Kyoto, Japan.

Peter Stoilov, Erik Anderson, Areum Han, and Douglas L. Black (February 2009), *High throughput screening for drugs that act as alternative splicing modulators*, Interdisciplinary focus meeting on high throughput technologies for alternative splicing, Valencia, Spain.

Geetanjali Chawla, Chia-Ho Lin, Areum Han, Lily Shiue, Manuel Ares Jr., and Douglas L. Black (2009) *Sam68 regulates a set of alternatively spliced exons during neurogenesis*, Mol Cell Biol. 29:201-213.

Yi Xing, Peter Stoilov, Karen Kapur, Areum Han, Hui Jiang, Shihao Shen, Douglas L. Black, and Wing Hung Wong (2008) *MADS: A new and improved method for analysis of differential alternative splicing by exon-tiling microarrays*, RNA, 14:1470-1479.

Areum Han, Woo-Yeon Kim, and Sungmin Park (2007) *SNP2NMD: A database of human SNPs causing NMD (nonsense- mediated mRNA decay)*, Bioinformatics, 23 (3):397-399.

Areum Han, Hyo Jin Kang, Yoobok Cho, Sunghoon Lee, Young Joo Kim, and Sungsam Gong (2006) *SNP@Domain: a web resource of Single Nucleotide Polymorphisms (SNPs) within protein domain and sequences*, Nucleic Acids Research, 34, W642-W644.

Hyojin Kang, Taehui Hong, Won-Hyong Chung, Younguk Kim, Jinhee Jung, Sohyun Hwang, Areum Han, and Young Joo Kim (2006) *D2GSNP: a web server for the selection of Single Nucleotide Polymorphisms within human disease genes*, Genomics and Informatics, 4(1):45 47.

Jiyoung Lim, Young Joo Kim, Yongsook Yoon, Soon Ok Kim, HyoJin Kang, Jungsun Park, A. Reum Han, Bokghee Han, Burmseok Oh, Kyuchan Kim, Bang- won Yoon, and Kyuyoung Song (2006) *Comparative study of the linkage disequi- librium of an ENCODE region, chromosome 7p15, in Korean, Japanese, and Han Chinese samples*, Genomics, 87 (3):392-398.

Areum Han, Dongsu Kang, Taewoo Ryu, Sukhyun Moon, Kwanghyung Lee, and Doheon Lee (2005) *Extraction of developmentally important genes from micro array data*, Journal of Advanced Computational Intelligence and Intelligent Informatics, 9 (3).

Areum Han, Haeja Lee, and Sunghun Park (September 2001), *Medical information system using smart card and XML*, Poster presented at the korean society of medical and biological engineering meeting, Seoul, South Korea.

CHAPTER 1

Introduction

SPLICING

Splicing is an essential step in the gene expression pathway (Cech, 1986). A precursor messenger RNA (pre-mRNA) transcribed from a protein-coding gene contains non-coding segments called introns. Splicing removes introns and ligates coding segments, exons, together during pre-mRNA maturation.

In metazoans, pre-mRNAs are spliced with two-step transesterification processes in a step-wise manner (Black, 2003; Chen and Manley, 2009). In the first step, the 5' splice site of the pre-mRNA is cleaved, and two RNA intermediate pieces are generated (Cech, 1986). One piece is the linear first exon, and the other piece is the intron-second exon RNA lariat. The lariat is formed by a phosphodiester bond between the guanosine at the 5' splice site and 2'-hydroxyl of an adenosine at the branch point (Cech, 1986; Wahl et al., 2009). During the second step, the 3' splice site is cleaved, and two exons are ligated together, generating a spliced exon pair (Cech, 1986; Wahl et al., 2009).

The excision-ligation steps of splicing are catalyzed by a large Ribonucleoprotein (RNP) complex called the spliceosome (Wahl et al., 2009). The major spliceosome has five uridine-rich (U-rich) small nuclear RNPs (snRNPs) including U1, U2, U4/U6, U5 snRNPs and auxiliary factors (Wahl et al., 2009). The spliceosome machinery recognizes signals within pre-mRNA and proceeds splicing.

Three core signals are important for splicing activity: 5' splice sites, branch points, and 3' splice sites followed by a pyrimidine tract. These sequences are also recognized by the splicing machinery in a step-wise manner (Wahl et al., 2009). First, U1snRNA in the RNP binds the 5' splice site. Then, SF1/BBP and U2AF bind to a branchpoint, pyrimidine tract, and a downstream 3' splice site via RNA recognition motif (RRM) domains (Wahl et al., 2009). In the next step, U2snRNP binds to the branch point (Wahl et al., 2009).

Therefore, three core sequence signals affect affinity of snRNPs and associated proteins and, eventually, their splicing activity (Black, 2003; Chen and Manley, 2009). In lower organisms, core signals are highly conserved, and alternative splicing patterns are rarely observed (Ast, 2004). However, in higher organisms core signals are less conserved, leaving room for additional regulation (Ast, 2004). Thus, exon ligations can occur in different arrangements, allowing generation of alternative splicing events and mRNA isoforms.

ALTERNATIVE SPLICING

In mammalian cells, alternative splicing events are prevalent as nearly all-human genes (up to 94%) are alternative spliced (Pan et al., 2008; Wang et al., 2008). Alternative splicing generates diverse mature mRNAs from a single pre-mRNA transcript. Thus, alternative splicing allows cells to generate a dynamic transcriptome and proteome from a limited genome in response to developmental or environmental cues (Black, 2003; Chen and Manley, 2009).

Alternative splicing is regulated by trans-elements and cis-elements (Black, 2003; Chen and Manley, 2009). Protein splicing factors act as trans-acting elements and are usually expressed in a tissue-specific manner. Splicing factors often bind to their cognate binding sites near the target exon and regulate whether the exon will be included or excluded in the final mRNA (Black, 2003). Several hundred splicing factors have been reported in mammalian cells, of which the two major groups are the Serine/arginine-rich (SR) proteins and the heterogeneous nuclear ribonucleoproteins (hnRNPs) (Black, 2003). Generally, SR proteins are known as splicing enhancers that enhance the recognition of splice sites via their arginine/serine-rich (RS) C-terminal domains (Black, 2003). In contrast, the hnRNPs are known mainly as splicing silencers that bind to pre-mRNA and block core splicing signals (Black, 2003). However, many studies have reported that SR proteins can act as silencers and hnRNPs proteins can act as enhancers.

Regulatory cis-elements are mostly binding sites for splicing factors. Cis-elements are categorized into four groups depending on their location and regulation mode (Fairbrother et al., 2004; Wang et al., 2004): Exonic Splicing Enhancers (ESEs), Exonic Splicing Silencers (ESSs), Intronic Splicing Enhancers (ISEs), and Intronic Splicing Silencers (ISSs). For example, exonic splicing enhancers are located in target exons and enhance inclusion of the target exons in final mRNA. Intronic splicing silencers are located in flanking introns and repress inclusion of the target exons. Interestingly, these cis-elements are not necessarily exclusively enhancers or silencers, and one regulatory sequence can have a dual effect in regulating exons.

There are two possible explanations for the dual roles of cis-elements. Firstly, the location of the cis-element may affect its splicing regulation mode. A cis-element bound splicing factor may interact differently with the splicing machinery based on its location. Secondly, a cis-element may be recognized by different splicing factors. Most splicing factors tend to have degenerate and relatively non-specific binding sites, and therefore it is likely that multiple splicing factors share same regulatory sequences.

To understand mRNA regulation by splicing factors, it is critical to know the binding affinity and specificity of a splicing factor to its cis-elements targets. This knowledge would help our understanding of the differential roles of splicing factors, enable the study of interactions among splicing factors and spliceosome machinery, and elucidate splicing regulation mechanisms by splicing factors.

The ultimate goal of alternative splicing research will be predicting alternative splicing events and mRNA isoforms fully based on expression of splicing factors and their cis-elements. This will enable researchers to correct disease-causing exons and mal-functional mRNA isoforms.

POLYPYRIMIDINE TRACT BINDING PROTEINS (PTBP)

Polypyrimidine Tract Binding (PTB) proteins are splicing factors and members of the hnRNP family (Keppetipola et al., 2012). In mammals, four PTB genes have been reported including PTBP1 (hnRNPI), PTBP2, ROD1 (Regulator of Differentiation 1 or PTBP3), and smPTB (Keppetipola et al., 2012). PTBP1 is broadly expressed in most cells except differentiating neurons and muscle cells. Conversely, the three paralog genes (PTBP2, ROD2, and smPTB) have more restricted expression (Keppetipola et al., 2012). PTBP2 is expressed in neuronal cell types, while expression of ROD1 and smPTB have been observed in hematopoietic cells and smooth muscle, respectively (Keppetipola et al., 2012).

PTBP1 is an abundant protein and has been the most extensively studied among PTB proteins (Keppetipola et al., 2012). PTBP1 is a strong splicing factor and was originally identified by its binding to the polypyrimidine tract of adenoviral major late (AdML) pre-mRNA substrate (Garcia-Blanco et al., 1989). Splicing repression of neuronal or muscle specific exons by PTBP1 have been extensively studied. Those exons include exons in alpha and beta tropomyosin, the N1 exon of the c-src pre-mRNA, exon 9 of GABA A receptor-gamma2 subunit, exon3b of FGF-R2, exon SM of alpha-actinin, and exon 4 of calcitonin/CGRP pre-mRNA (Keppetipola et al., 2012; Valcarcel and Gebauer, 1997).

PTBP1 has a nuclear localization domain and four RNA recognition modules (RRMs), or RNA binding domains (RBDs) (Oberstrass et al., 2005). Using these RRMs or RBDs, PTBP1 can bind and regulate RNA. Each RBD interacts with a short pyrimidine element, allowing the full protein to recognize a wide variety of pyrimidine-rich RNA sequences (Oberstrass et al., 2005). PTBP1 localizes predominantly to the nucleus where it regulates RNA processing events such as splicing and polyadenlylation (Keppetipola et al., 2012). Under certain conditions such as cell adhesion, PTBP1 can shuttle to the cytoplasm where it can regulate mRNA localization, translation, and decay (Keppetipola et al., 2012).

PTBP1 has been known mainly as a strong splicing repressor. Two types of repression mechanisms have been revealed in individual exon studies (Spellman and Smith, 2006; Valcarcel and Gebauer, 1997). According to one model, PTBP1 binds to the polypyrimidine tract near a 3' splice site and competes out U2AF binding. The failure of U2AF binding to the polypyrimidine tract inhibits initial spliceosome assembly on the target exon and splicing of the target exon. However, this model is not sufficient to explain other cases. In minigene studies and other PTBP1 repressed exons, PTBP1 binds to intronic regions other than the polypyrimidine tract but is still able to repress splicing of the target exon. Other model proposes that PTBP1 binding does not affect initial spliceosome assembly but rather inhibits later steps (Izquierdo et al., 2005; Sharma et al., 2008). According to these studies, PTBP1 binds to the target exon or nearby flanking introns and inhibits crosstalk between U1 snRNP and U2AF. This blocks exon or intron definition, both of which are necessary for the target exon splicing. In the case of c-src exon N1, PTBP1 binds to stem loop 4 of U1snRNP and inhibits interaction between U1snRNP and downstream U2AF, which in turn inhibits the transition from exon definition to intron definition (Sharma et al., 2008).

Though PTBP1 is a well-studied splicing factor, its general binding and splicing codes are not yet well modeled. The binding and splicing code for PTBP1 is not likely to be simple as PTBP1 tends to bind diverse pyrimidine sequences and has multiple regulation rules. Development of new methodology is needed to derive a binding and splicing code for PTBP1. In the future, new methodology can be applied to other multi-RRM domain proteins to assess binding site degeneracy and multifactorial splicing regulation.

PTBP1 and its neuronal paralog, PTBP2, are very similar proteins with over 70% amino acid sequence identity (Keppetipola et al., 2012). Additionally, there are only two amino acid differences among their RNA interface (Keppetipola et al., 2012). It has been reported that PTBP1 and PTBP2 have similar affinity to a pyrimidine-rich sequence within the upstream intron of the c-src-N1 exon (Keppetipola et al., unpublished). For pyrimidine sequences PTBP1 and PTBP2 seem to have similar RNA binding affinities. However, PTBP1 and PTBP2 are expressed mutually exclusively in many cell lines and tissues (Boutz et al., 2007b). Expression of PTBP2 is tightly down-regulated in the presence of PTBP1. In many PTBP1 knock down studies, up-regulation of PTBP2 was observed. Previous studies revealed that PTBP1 enhances nonsense-mediated decay (NMD) of PTBP2 transcripts by repressing inclusion of an alternative exon (Boutz et al., 2007b). Thus, there is a direct down-regulation of PTBP2 transcripts by PTBP1. Additional regulation at the translational level has also been observed (Boutz et al., 2007b). PTBP2 transcripts, which are not NMD subjects, are not efficiently translated in the presence of PTBP1. This tight regulation between these two paralogs might have evolved so that each protein regulates a different splicing network. It will be interesting to study whether PTBP1 and PTBP2 have similar or divergent binding and splicing codes to recognize their splicing targets.

TECHNIQUES TO STUDY BINDING TARGETS OF SPLICING FACTORS

Most splicing factors regulate pre-mRNA splicing by binding to RNA sequences through their RNA recognition motifs (RRMs) or RNA binding domains (RBDs). Biochemical methods have identified RNA sequences bound by splicing factors. Early approaches identified single or few RNA targets from RNA-splicing factor complexes. The complexes were assembled in vitro or from cells and isolated using a splicing factor specific antibody. Then, RNA in the complex was precipitated, cloned, and identified by sequencing.

Later, RNA pools have been used to identify multiple RNA targets. The Systematic Evolution of Ligands by Exponential Enrichment (SELEX) method has been a representative technique (Smith, 1998). SELEX starts from a random RNA pool and iteratively selects a splicing factor bound RNA pool. After certain number of cycles, the selected RNA pool is enriched with target RNAs, which are then cloned and sequenced. In this traditional SELEX, cloning and sequencing were the major limiting steps in terms of cost and labor. To overcome this, advanced SELEX coupled with a highrouput technique was proposed (Ray et al., 2009; Reid et al., 2009). Advanced SELEX methods such as RNAcompete and next generation SELEX systemically compare the starting RNA pool and the protein-bound RNA pool using microarrays. However, SELEX based approaches have some intrinsic limitations. Firstly, RNA-splicing factor complexes are not stable enough to endure rigorous biochemical conditions. Thus, it was difficult to remove indirect target RNAs. Secondly, SELEX is an in vitro technique and does not identify physiological RNA targets in cells.

To identify specific and in vivo RNA targets of splicing factors, CLIP (Cross-Linking and Immuno Precipitation) techniques have been developed (Licatalosi et al., 2008; Ule et al., 2005; Ule et al., 2003). Cells are exposed to ultraviolet light, which crosslinks RNA to the amino acids of splicing factors when they within a range of angstroms in vivo. The covalent bond that forms allow researchers to perform rigorous downstream procedures including RNA digestion, stringent washing, transferring complexes, and boiling in SDS (Sodium Dodecyl Sulfate) (Ule et al., 2005). After isolation of the protein-RNA complex, RNA is converted to cDNA and ligated with linkers for high throughput sequencing (Ule et al., 2005). CLIP coupled with high throuput sequencing enables identification of millions of in vivo RNA targets.

CLIP-SEQ protocols have been further improved. In the original CLIP-SEQ method, the cross-link site information was missing and binding sites could not be tracked at nucleotide resolution. To obtain the nucleotide sequence where a splicing factor is cross-linked to RNA, two methods have been proposed, including PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) and iCLIP (individual-nucleotide resolution CLIP (Hafner et al., 2010; Konig et al.). In PAR-CLIP, photoreactive ribonucleoside analogs such as 4-thiouridine (4-SU) and 6-thioguanosine (6-SG) are incorporated into nascent RNA transcripts in cultured cells and cross-linked to the splicing factor (Hafner et al., 2010). When the target RNA is converted to cDNA libraries, those analogs cause mutations (in 4-SU, T-> C and in 6-SG, G-> A). Thus, the presence of mutations enables identification of the crosslinking sites. The iCLIP method was also devised to determine cross-linking sites. In traditional CLIP, many cDNAs truncate prematurely at the cross-linked nucleotide and are lost during the library preparation. The iCLIP method rescues those truncated cDNAs using circularization (Konig et al.). Sequencing the truncated position provides information on where the splicing factor cross-linked to RNA with nucleotide resolution. Coupled with new sequencing technology development, CLIP based methods are expected to provide tremendous information about RNA targets of splicing factors.

After RNA targets are identified from the previously described methods, their binding characteristics, such as binding affinity, can be studied in depth through binding assays. The Electrophoretic mobility shift assay (EMSA) and filter-binding assay are two representative methods (Smith, 1998). In both methods, target RNA are transcribed in vitro from double stranded DNA and incubated with splicing factors or nuclear extracts containing splicing factors. In the next step, RNA-protein complexes are identified and quantified. This second step differs in the two assays. EMSA identifies the RNA-protein complexes by electrophoretic mobility retardation whereas a filter-binding assay detects membrane bound RNA-protein complexes (Nagai and Mattaj, 1994). The filter-binding assay is simpler than EMSA but cannot be used to study RNA with multiple binding sites. In EMSA, heterogeneous RNA-protein complexes can be studied as they show different movement retardation compared to the free RNA on the electrophoretic gel (Nagai and Mattaj, 1994).

Binding affinity of a splicing factor to RNA can be assessed by determining dissociation constant (*Kd*) from EMSA results (Smith, 1998). A lower Kd refers higher binding affinity and vice versa. *Kd* for RNA with a single binding site is defined from the equation $Kd = [R][P]/[PR]$, where [R] is the free RNA concentration, [P] is the free protein concentration, and [PR] is the RNA-protein complex concentration (Nagai and Mattaj, 1994). When protein concentration is in excess of the substrate RNA, one can assume that the quantity of proteins is not a limiting factor. Based on this assumption, *Kd* is equal to protein concentration, [P], when 50% of input RNA is bound to the splicing factor ($[R]/[PR]=1$) (Nagai and Mattaj, 1994). In the gel, this is the protein concentration at which the intensity of the RNA-protein band is the same as the intensity of the free RNA band (Nagai and Mattaj, 1994). When there is more than one binding site in the RNA, the calculated *Kd* from previous equation is called as the 'apparent *Kd*'. And the multiple binding properties can be analyzed in detail (Nagai and Mattaj, 1994).

With the development of biochemical methods, advances in computational methods have allowed characterization of RNA targets of splicing factors. Computational efforts have focused on motif discovery from multiple RNA targets. Different motif discovery methods include detecting over-represented words and inferring a positional frequency matrix of nucleotides. Interestingly, few splicing factors have reported to have well-defined binding sites. For example, a splicing factor, RbFox-1, was shown to bind (U)GCAUG elements in upstream flanking introns, repressing the inclusion of muscle specific exons (Jin et al., 2003). In most cases, motifs for splicing factors are relatively less specific (Gabut et al., 2008). Most splicing factors have short and degenerate motifs that are usually described as 'certain nucleotide enriched elements'. The low specificity in target RNA recognition might allow splicing factors to regulate multiple targets in combinatorial fashion. Low specificity might not be a result of degeneration, but a result of evolution to regulate transcriptome dynamically.

Because of the low specificity in splicing factor motifs, it has been challenging to model binding targets of splicing factors. As millions of RNA targets of splicing factors become available, development of statistical methods to model binding sites are important. Successful methodology will enable researchers to search binding sites and studies on binding properties of splicing factors.

BINDING TARGETS OF PTBP

PTBP1 has four RNA recognition modules (RRMs) and was originally identified by its binding to the polypyrimidine tract of pre-mRNA substrates (Garcia-Blanco et al., 1989). Researchers have been tried to identify and characterize binding targets of PTBP1 using different biochemical approaches.

One group performed a SELEX experiment with iterative selections of PTBP1 bound RNA in vitro (Singh et al., 1995). The majority of PTB1-selected clones contained the consensus sequence, (U/G)C(A/Y)GCCUG(Y/G)UGCYYYYCYYYYG(Y/G)CCC, where Y indicates pyrimidines (C or U) and N refers to any nucleotide. Later, two groups applied advanced SELEX methods with microarray (Ray et al., 2009; Reid et al., 2009). Ray *et al.* designed a substrate pool containing unstructured and stem-loop structured RNA and incubated it with PTBP1. The PTBP1 bound RNAs were recovered and labeled with Cy3 fluorescence dye. Simultaneously, initial RNA pools were labeled with Cy5 fluorescence dye. These were co-hybridized onto a microarray and differential fluorescence intensities from the array were measured as binding preferences. From a set of sequences with high PTBP1 binding preferences, ‘CUNUC' was derived as a PTBP1 consensus motif. Using this motif and 7-mer based binding preferences, they also tried to predict PTBP1 binding affinity to sequences. However, the correlation between predicted and measured binding preferences was low (0.3~0.39) (Ray et al., 2009). This result suggests that the short PTBP1 motif itself is not enough to predict binding affinity of PTBP1.

Reid *et al.* designed a different RNA pool from pre-mRNA sequences (Reid et al., 2009). This RNA pool was incubated with nuclear extract from Hela cells, where endogenous expression of PTBP1 is high. Then, PTBP1 bound RNA was co-immunoprecipitated with a PTBP1 antibody. The PTBP1 bound and unbound RNA pools were isolated and co-hybridized onto a two-color microarray, as before. Compared to previous studies, this study used endogenous PTBP1 in nuclear extract. It was possible that recovered RNA came from complexes and thus are indirect targets of PTBP1. Several motifs with consensus sequences such as CUCUC, UUUCU, and CUG were reported. Interestingly, they found a G-interrupted pyrimidine motif that could be another motif of PTBP1 or that of PTBP1 interactors (Reid et al., 2009).

Recently, PTBP1-CLIPSEQ was performed to map transcriptome-wide PTBP1-RNA interactions in Hela cells (Xue et al., 2009). A hexamer, ‘UUCUCU’, was most enriched within PTB-CLIP clusters. A consensus, sequence ‘UYUYU’, was then derived from the top twenty enriched hexamers (Xue et al., 2009).

In summary, several PTBP1-RNA binding studies previously performed all agreed that PTBP1 binds to pyrimidine rich sequences. Short and degenerate pyrimidine words were derived as a PTBP1 motif. However, the motif has not been able so far to predict affinity of PTBP1 to RNA molecules. This highlights the need for new methodology to model PTBP1 binding sites. To build a predictive PTBP1 binding model, we need to understand the structural properties of the PTBP1-RNA interaction.

Previously, structural studies revealed how PTBP1 binds to pyrimidine sequences (Lamichhane et al., 2010; Oberstrass et al., 2005). Each RBD of PTBP1 was incubated with an RNA oligonucleotide, 5’ CUCUCU 3’. Interactions between PTBP1 and RNA were studied in depth using NMR (Nuclear Magnetic Resonance) spectroscopy (Oberstrass et al., 2005). The study showed that each RBD can bind to RNA molecules. RBD1 can bind YCU, RBD2 can bind CUNN, RBD3 can bind YCUNN, and RBD4 can bind YCN. Additionally, the C terminus containing RBD3 and 4 and their domain linker bound to two RNA molecules (two CUCUCU sequences), while isolated RBD1 or RBD2 could bind one RNA molecule each. RBD3 and

RBD4 interact physically and bind to long RNA in antiparallel orientation. This can induce RNA looping, which was further studied with FRET (Fluorescence Resonance Energy Transfer), NMR, and in vivo splicing experiments (Lamichhane et al., 2010). Their results confirm that RBD3 and RBD4 can bind two distal pyrimidine elements and bring them close together by looping of the RNA. Interestingly, looping requires a certain length of gap sequence (15 nucleotides or more) between two distal elements. In a PTBP1 target RNA, there may be more gaps between binding sites because the two linkers between RBD1 and RBD2 and between RBD2 and RBD3 are flexible and do not participate in RNA binding.

Insights from previous binding and structural studies are keys to build a predictive PTBP1 binding model. First, the observation of G-interrupted pyrimidines as a second motif and in the lower cycle of SELEX suggests that PTB may bind to G-interrupted pyrimidine sequences with lower affinity. Thus, a model of PTB binding may need to count G-interrupted pyrimidines as potential low affinity sites. Secondly, the structural studies pointed out that while binding sites for each RBD are very similar to pyrimidines flexible linkers between RBD1 and RBD2, and between RBD2 and RBD3, might allow various length gaps between binding elements. RBD3 and RBD4, especially, need at least 15 nucleotide gaps between binding sites to form a loop. Therefore, the model needs to handle flexible gaps between potential binding sites. A model that grasps those properties will be able to predict overall affinity as well as potential binding sites of PTBP1.

TECHNIQUES TO STUDY EXON TARGETS OF SPLICING FACTORS

In previous sections, we discussed the physical targets of splicing factors, RNA. Another important target of splicing factors is the alternatively spliced exon. Splicing factor dependent alternative exons have been identified by assaying mRNA changes while manipulating the expression levels of splicing factors. In cells, researchers have successfully down-regulated splicing factors using small interfering RNA (siRNA) or short hairpin RNA (shRNA). Researchers have also up-regulated splicing factors by introducing expression vectors encoding splicing factors of interests. In vivo, knock-out or knock-in mouse strains for splicing factors have been created to reduce or induce splicing factor expression. From those cells or tissues, mRNA can be isolated and converted to cDNA. Then, cDNA can be subject to highthrouput assays such as microarrays or next-generation sequencing. These high throughput methods enable researchers to monitor transcriptome-wide changes in alternative splicing events.

The splicing microarray is a hybridization-based approach (Modrek and Lee, 2002). Probes for splicing microarrays cover exonic regions and may span exon junctions. The cDNA from samples of interest are labeled with different colored fluorescence and are incubated together on a microarray chip. Differential ratios of the fluorescence are measured and alternative splicing patterns are analyzed. The exon inclusion levels are calculated by the ratio of signal from probes representing exon inclusion events compared to that from probes representing exon skipping events.

On the other hand, RNA-seq is a sequencing based approach (Mortazavi et al., 2008) (Wang et al., 2009). RNA is fragmented and converted to cDNA. Adaptors are then ligated to the cDNA to prepare RNA-seq libraries. During the library preparation, additional selection steps can be introduced to enrich for a specific RNA population. For example, poly-A (Adenosine) selection enriches for polyadenylated mRNA in the libraries. Researchers have also devised strand specific library preparation to preserve RNA polarity (Parkhomchuk et al., 2009). Standard RNA-seq library preparation ligates double stranded sequencing primers to the double-stranded cDNA. Thus, information about which strand was present in the starting RNA is lost. To conserve RNA polarity, two types of approaches have been introduced (Levin et al., 2010). One approach is to ligate different adaptors to the 5' and 3' ends of the RNA. This approach is straightforward but requires handling the RNA in several reactions and increases the risk of RNA degradation. A second approach introduces chemical modification in RNA or the second-strand of the cDNA. In the RNA, bisulfate treatment converts C to U. Searching for these mutations after sequencing can retrieve original strand information. To mark the second-strand of cDNA, dUTP instead of dTTP can be incorporated during second-strand cDNA synthesis. Later, the second strand 'U' is eliminated by uracil- specific excision, and only first strand cDNA can be sequenced.

In some aspects, splicing microarrays and RNA-seq complement each other in detecting alternative splicing events. Generally, RNA-seq is more quantifiable than microarray because the amount of hybridized cDNA on chips is not linear to the amount of cDNAs, especially when the concentration of cDNA is high. Microarray probes can become saturated for highly expressed transcripts and unable to detect differential expression between two samples. However, microarrays are more sensitive in detecting alternative splicing events from lower expressed genes. In RNA-seq, abundant transcripts can take over the sequencing capacity and compete out lowly expressed transcripts. In view of data analysis, both high throughput methods are highly dependent on well-annotated alternative splicing databases. However, RNA-seq data have more potential to detect novel alternative splicing events by raw sequence assembly.

Though high throuput assays allow researchers to identify multiple exons regulated by splicing factors, false signals are inherent in such methods. For example, cross hybridization can cause false signals in microarray-based experiments. In the case of RNA-seq, bias in RNA fragmentation or faulty mapping can lead to missed exons or false quantification. Therefore, candidate exons from highthrouput experiments can be further confirmed using an additional validation method such as RT-PCR (Reverse Transcription-Polymerase Chain Reaction) (Modrek and Lee, 2002).

RT-PCR has identified alternative splicing events at the individual exon level with great accuracy. First, specific primers are designed against flanking exons of target exons. Then, cDNA fragments between the primers are amplified by PCR and radioactively labeled. Products from alternative splicing events have different molecular sizes and can be discriminated by electrophoresis.

From highthrouput assays and following RT-PCR validation, researchers have compiled lists of splicing factor dependent alternative exons. Once a list is complied, computational analysis can be applied to characterize their biological meanings. Gene ontology or pathway analysis has been used to assess enrichment of those exons within genes in certain categories of biological function or pathways. In particular, searching for binding motifs has been an important and necessary analysis step. In mammalian cells, several hundred splicing factors exist, and it is possible that target exons are also indirect targets of multiple splicing factors (Black, 2003; Blencowe, 2006). Flanking introns and exon sequences have been subjects of binding motif analyses. When the binding motifs or sites are found from binding assay results, they can be searched or mapped. Exons with known binding motif/sites nearby are likely direct targets of the splicing factor of interest. Locations of the binding motif/sites are important clues for mechanisms of the splicing factor.

On the other hand, novel binding motifs or sequential signatures can be discovered. Newly identified motifs can be confirmed further by additional binding assays. When the discovered motif is a known binding motif of other splicing factors, the discovery can lead to studies of synergetic or antagonistic activities between splicing factors.

In previous studies, it has been shown that positions of binding sites affect the regulation modes of splicing factors. Recently, binding and exon target information have been integrated and visualized as RNA maps (Llorian et al., 2010; Ule et al., 2006; Wang et al., 2012; Xue et al., 2009; Yeo et al., 2009). In an RNA map, density distributions of binding sites from CLIP-seq or known motifs are plotted along the target exons and their flanking introns. Interestingly, the most common pattern observed in several RNA maps so far is one where splicing factors binding sites within the upstream intron leads to repression of the target exon. In contrast, when binding sites are positioned in the downstream intron, splicing of the target exon tends to be enhanced. Many splicing models were derived from RNA maps to reveal new splicing mechanisms.

Multiple features and cis-elements beyond binding signals may also affect alternative splicing events. A step further from two-dimensional RNA maps and simple rules, researchers have attempted to derive splicing models/codes from multiple RNA feature collections. One study aimed to model exon targets of a neuronal specific splicing factor, Nova (Neuro-Oncological Ventral Antigen) (Zhang et al., 2010). Multiple RNA features including Nova-CLIP data, Nova splicing microarray data, known binding motifs, and evolutionary signatures were integrated in a probabilistic framework. Using a Bayesian network, they identified ~700 candidate exons that may directly regulated by Nova. Another study assembled a splicing code from RNA features including various motifs, secondary structure, and transcript structure characteristics. With hundreds of RNA features, the model built a weighted combination of decision trees and was able to predict tissue-specific alternative splicing events (Barash et al., 2010).

In mammalian cells, alternative splicing is regulated by several hundred splicing factors. Ultimately, it will be ideal to have splicing models for each splicing factor and assemble a complete splicing network model. The complete splicing network model or splicing code will aid in the understanding of mechanisms of alternative splicing and prediction of transcriptome dynamics.

EXON TARGETS OF PTBP

Using RNA interference technology, PTBP1 and PTBP2 have been successfully down-regulated in mammalian cells, including mouse N2A (neuroblastoma) and human Hela cells (Boutz et al., 2007b; Llorian et al., 2010; Xue et al., 2009). Microarray experiments with PTB protein knockdown identified target exons of PTB proteins (Boutz et al., 2007b). One study down-regulated PTBP1, PTBP2, and both PTBP1 and PTBP2 in mouse N2A cells and assayed differential alternative splicing events using microarrays. Interestingly, target exons of PTBP1 and PTBP2 overlap significantly, but some exons respond differently to loss of PTBP1 or PTBP2. This suggests that there are exons that are regulated differently by two PTB proteins or are more sensitive to one PTB protein than the other (Boutz et al., 2007b). Identified exons were further validated with RT-PCR and mapped to the orthologous human exons (Xue et al., 2009). Using these exons and previously known human exons, one group complied a list of 55 PTBP1 regulated exons, including 41 repressed and 14 enhanced exons. Then, PTBP1 CLIP cluster tags were overlaid to those exons and their flanking introns to generate a PTBP1 RNA map. The map suggested that PTBP1 repressed exons tend to have PTBP1 CLIP tags near both side of the alternative exons, whereas PTBP1 enhanced exons tend to have PTBP1 CLIP tags in the 5' splice site of the downstream flanking constitutive exons (Xue et al., 2009).

Another study down-regulated both PTBP1 and PTBP2 in human Hela cells and characterized sequence features of differentially spliced exons (Llorian et al., 2010). Since this study is performed with double knock down of PTBP1 and PTBP2, it is possible that some characteristics are from PTBP2 dependent exons. They observed that PTBP1 and/or PTBP2-repressed exons showed enrichment of PTBP1 binding motifs within the exons and in the flanking upstream intron, but not in the downstream intron. In contrast, PTBP1 and/or PTBP2 enhanced exons were associated with enrichment of PTBP1 binding motifs in the downstream intron (Llorian et al., 2010).

These two RNA maps agreed to an extent. Both RNA maps suggest that PTBP1 repressed and PTBP1 and/or PTBP2 repressed exons have enrichment of PTBP1 CLIP tags in the upstream and downstream introns (Llorian et al., 2010; Xue et al., 2009). However, PTBP1 and/or PTBP2 repressed exons showed less enrichment of PTBP1 CLIP cluster tags in the downstream introns compared to upstream introns. It is possible that PTBP2 dependent exons do not require binding sites within the downstream introns (Llorian et al., 2010). These PTB RNA maps also show distinct patterns especially for enhanced exons. PTBP1 and/or PTBP2 enhanced exons shows enrichment of binding signals in the downstream introns. This pattern is also observed for other splicing factors. Interestingly, in the other map, PTBP1 enhanced exons show enrichment of the binding signals near the 3' splice sites of downstream constitutive exons. This suggests that PTBP1 inhibits recognition of the competitive constitutive 3' splice site and enhances selection of 3' splice sites of the upstream alternative exon instead (Xue et al., 2009). Both observations were confirmed with additional minigene studies revealing new splicing enhancement models by PTBP1. Additionally, PTBP1 and/or PTBP2 repressed cassettes and alternative 3' splice site exons tend to have longer AG dinucleotide exclusion zones upstream of the 3’ splice site, indicating that they have distant branch-point locations. Furthermore, motifs such as UGCU and UGCAUG are present in downstream introns of the PTBP1 and/or PTBP2 repressed exons. The UGCU could be binding sites for PTBP1, while the UGCAUG could be binding site for other co-regulators such as Rbfox proteins (Llorian et al., 2010).

In summary, multiple exons regulated by PTB proteins have been identified and their associated sequential features studied. However, so far there is no unified model, which can assemble those sequence features and predict PTB target exons solely by their pre-mRNA sequences. Development of such a PTB splicing model will enable researchers to search novel targets of PTB proteins and aid the understanding of underlying splicing mechanisms for each individual exon. Microarray experiments suggest that two paralogues, PTBP1 and PTBP2 have some divergent exon targets. Splicing codes, which can differentiate sensitivity of exons for each protein will uncover differential roles of PTBP1 and PTBP2.

Transgenic mice are good tools to study roles of PTB proteins during mammalian development and tissue differentiation (Hedrich Hans, 2004). There are ongoing efforts to create PTB protein knock out/in mice and investigate the resulting phenotypes (Licatalosi et al., 2012; Schuelke, 2000; Shibayama et al., 2009). Identification and characterization of binding and splicing targets of PTB proteins in such in vivo settings will help to find the underlying molecular mechanisms.

CHAPTER 2

De Novo Prediction of PTBP1 Binding and Splicing Targets Reveals Unexpected Features of Its RNA Recognition and Function.

INTRODUCTION

Alternative splicing of pre-mRNA commonly determines the protein output of mammalian genes, with most human genes generating multiple mRNA and protein products (Pan et al., 2008; Wang et al., 2008). A typical alternative exon is affected by multiple pre-mRNA binding proteins that may either enhance or repress splicing (Black, 2003). The expression and activity of these splicing regulatory proteins can vary with development, cell type, or cellular stimulus (Gabut et al., 2008). This complex combinatorial regulation can be seen in the conserved sequences within and surrounding alternative exons, which generally contain the binding sites for many different regulators. These sequences make up what is sometimes called the splicing code as they determine where and when the exon is spliced into an mRNA (Barash et al., 2010; Blencowe, 2006; Matlin et al., 2005; Wang and Burge, 2008). However, this regulation is currently difficult to predict, in part due to our incomplete understanding of RNA recognition by the splicing regulators and their mechanisms of action.

Splicing regulatory proteins commonly contain multiple RRM or other RNA binding domains, with each domain recognizing a short element of a few nucleotides (Auweter et al., 2006; Black, 2003). Subtle variation in the optimal binding element of each domain and flexible peptide linkers between them allow for significant degeneracy within high affinity binding sites. Although the short sequence motifs that are common to a set of binding sites are readily identified, these likely constitute only a portion of a full high affinity site. To rank binding sites and assess their finer structures, we need an approach to search for and score clusters of these short motifs.

Whole-transcriptome methods for identifying the binding sites of individual proteins *in vivo* are yielding large amounts of information about their RNA recognition properties (Darnell, 2010; Hafner et al., 2010; Ule et al., 2005). These data can be overlapped with functional data on splicing to identify possible direct target exons (Licatalosi et al., 2012; Wang et al., 2012; Xue et al., 2009; Yeo et al., 2009). However, these data are generated one tissue or cell type at a time. It would be extremely useful to be able to assess potential binding across the complete transcriptome and to predict exon targets in tissues that have not yet been subjected to experimental analysis.

The Polypyrimidine tract binding protein 1 (PTBP1) is a widely studied splicing regulatory protein (Keppetipola et al., 2012; Spellman and Smith, 2006; Valcarcel and Gebauer, 1997). PTBP1 is known to repress the splicing of a large number of exons by binding in their adjacent introns or within the exons themselves. PTBP1 is down regulated in differentiating neurons and muscle cells to allow inclusion of PTBP1 repressed exons during development of these tissues (Boutz et al., 2007a; Boutz et al., 2007b; Llorian et al., 2010). In neurons the loss of PTBP1 is accompanied by the up-regulation of the homologous protein PTBP2 (Boutz et al., 2007a; Keppetipola et al., 2012; Makeyev et al., 2007). PTBP2 has similar binding properties to PTBP1 and represses some of the same exons (Zheng et al., 2012). Other exons are more sensitive to PTBP1 than PTBP2 and are induced to splice when PTBP2 replaces PTBP1 in early neurons (Tang et al., 2011). How these two proteins differ in their regulatory properties is not known.

Experiments with model substrates indicate that a single high affinity PTBP1 binding site placed upstream of an exon, or within it, can repress splicing (Amir-Ahmady et al., 2005). However, strong repression of an efficiently spliced exon requires an additional binding site either within the exon or downstream from an exon with an upstream high affinity site (Amir-Ahmady et al., 2005). PTBP1 is also known to enhance the splicing of certain exons (Boutz et al., 2007b; Llorian et al., 2010; Xue et al., 2009). The properties of these exons and how they differ from those that are repressed by PTBP1 are unclear, with different studies coming to different conclusions.

PTBP1 contains four RRM domains that recognize short pyrimidine elements (Oberstrass et al., 2005). Flexible linkers separate RRM domains one and two, and domains two and three. RRM domains three and four interact through a hydrophobic interface that position their RNA binding surfaces on opposite faces of the two-domain structure. This orientation requires that the RNA elements interacting with the structure be separated by an RNA loop (Lamichhane et al., 2010). The structure of each of the PTBP1 RRM domains has been solved in complex with the hexanucleotide, CUCUCU (Oberstrass et al., 2005). These structures show each domain binding a nucleotide triplet with some additional contacts, and making similar base specific interactions with CU or UC dinucleotides. Other sequences can likely make different base specific contacts, and the optimal elements for each domain are not known. Moreover, the flexible linkers separating some of the RRM domains and the requirement for a gap between elements simultaneously bound to domains three and four allow for substantial degeneracy in PTBP1 binding sites. This degeneracy and the lack of understanding of the sequence features that contribute to binding affinity have made it difficult to identify PTBP1 binding sites based on sequence alone, and to assess which sequences surrounding an exon might contribute to PTBP1 regulation.

In this study we sought to understand the sequence features that determine RNA binding by PTBP1 and to examine how they are combined in exons that are targeted by the protein. We first developed a statistical model of PTBP1 binding sites that identifies new features of RNA recognition by the protein. This binding model was then applied to the assessment of exon regulation by PTBP1 across the transcriptome.

RESULTS

G containing triplets contribute to PTBP1 binding

To examine the interactions of PTBP1 across many binding sites, we used a set of PTBP1-bound sequences identified by crosslinking immunoprecipitation (CLIP) (Xue et al., 2009). 48,604 CLIP clusters from the human transcriptome were extracted and used to train a Hidden Markov Model (HMM) of PTBP1 binding sites as shown in Figure 1A (Durbin, 1998; Rabiner, 1989). This model defined two states, PTBP1-binding and non PTBP1-binding, and scored nucleotide triplets for their probability of being found in each of the two states. We chose triplets because each of the PTBP1 RRMs makes base specific contacts with at least three nucleotides (Oberstrass et al., 2005). We identified 20 triplets that increased the probability of PTBP1 binding (Figure 1A). The remaining 44 triplets decreased the probability of binding. The 20 triplets predictive of PTBP1 binding included the expected pyrimidine sequences with the top scoring triplet UCU showing the alternating C and U nucleotides seen in many characterized PTBP1 binding sites. All triplets containing only pyrimidines increase the probability of PTBP1 binding (Figure 1B).

Interestingly, some triplets containing G residues were also predictive of PTBP1 binding (Figure 1B). These triplets often contain U residues as the other nucleotides, and some increase the probability of PTBP1 binding substantially (e.g. UGU). In contrast, triplets containing A residues, even if the other two nucleotides are pyrimidines, all decrease the probability of PTBP1 binding. These results indicate that PTBP1 is not strictly pyrimidine specific. At least one of its RRM domains can presumably make specific contacts with G residues. On the other hand, A residues are less tolerated in PTBP1 binding sites and are likely to inhibit binding.

We next tested how well the HMM was able to predict PTBP1 binding. We applied the trained model to a set of 100,000 random 69 nucleotide sequences. This length allows for one hexanucleotide binding site for each of the four RRMs with 15 nucleotide gaps, the minimum gap required for simultaneous binding by RRMs 3 and 4 (Lamichhane et al., 2010). These sequences generated a distribution of scores that was used to normalize the experimental scores, where the average score for random sequence is set to zero, and the z-score defined as the deviation from the average as shown in Supplemental Figure 1A (Durbin, 1998). Thus a sequence with a z-score of 2.74 is 2.74 standard deviations from the average (p-value = 0.005), and is predicted to be a significantly stronger binder than the average sequence (500 of the 100,000 random sequences are expected to have scores equal or greater than this sequence). A negative z-score is predicted to bind less well than the average sequence. We isolated thirteen sequences from the mouse transcriptome that exhibited a range of scores from -2.62 to +4.40 (Figure 2A). These were transcribed in vitro and subjected to electrophoretic mobility shift assay to measure binding to recombinant PTBP1 (Figure 2B; Supplemental Figure 1B). Sequences yielding negative scores all failed to bind PTBP1 within the protein concentration range tested, with the exception of probe 4, which bound weakly, below the level that would allow measurement of an affinity constant. Positive scoring sequences all yielded PTBP1 bound complexes that were assayable by gel shift to derive apparent binding affinities. The apparent Kds of these RNAs showed a very strong negative correlation with their binding score from the model (Pearson correlation coefficient = -0.9), where a higher score predicts a lower Kd and hence a higher affinity (Figure 2A). Two sequences (probes 9 and 11) showed variable and somewhat weaker binding that was off the fitted curve relating z-score to Kd. These may have secondary structures that reduce their apparent Kd. The data show that HMM scoring based on triplet frequencies can accurately predict the observed binding affinities across a wide range of Kd values (from ~250nM to 1nM). Probe 6 yields a z-score of 0.82 and binds with a Kd of 257nM, whereas probe 10 scores 2.74 in the model and binds with a Kd of 73nM (Figure 2B). These sequences include G containing triplets that contribute to the binding scores.

Placement of PTBP1 binding sites adjacent to target exons

We next examined known PTBP1 target exons for the location of predicted PTBP1 binding sites. Using a set of exons showing altered splicing after Ptbp1 knockdown, we defined four groups of exons (Boutz et al., 2007b; Xing et al., 2008). These included 68 PTBP1-repressed exons whose splicing increases after Ptbp1 knockdown, 37 PTBP1-enhanced exons whose splicing decreases after knockdown, 69 control (PTBP1-non regulated) exons that are not affected by Ptbp1 depletion but are known to be alternatively spliced, and 1,000 constitutive exons. We determined the density of predicted PTBP1 binding states within a 24-nucleotide window sliding along the exon region. We also examined the sequence encompassing the adjacent constitutive exons (Figure 3A). As expected, the control and constitutive exon sets did not exhibit high probabilities of PTBP1 binding except in the polypyrimidine tract of the 3' splice site. On the other hand, the introns upstream of PTBP1 repressed exons show enrichment of potential PTBP1 binding sites starting from 250 nucleotides upstream of the exon. Relative to the control exons, exons repressed by PTBP1 also exhibited substantial enrichment of PTBP1 binding sites within the exon itself, and within the first 100 nucleotides of the downstream intron. The PTBP1-enhanced exon set also have enrichment of PTBP1 binding sites within the downstream intron relative to control exons, although the distribution of binding sites across this region was different between the repressed and enhanced exon sets (Figure 3A). Some PTBP1 enhanced exons have PTBP1 binding sites upstream of the exon, although these are not as strongly enriched as in the PTBP1 repressed set. PTBP1 enhanced exons show no enrichment of binding sites within the exons themselves. These results are generally consistent with the known placement of PTBP1 binding sites in PTBP1 target exons (Amir-Ahmady et al., 2005; Llorian et al., 2010; Xue et al., 2009).

In an earlier study, it was noted that exons adjacent to PTBP1 enhanced exons sometimes have PTBP1 CLIP tags nearby and that PTBP1 binding at these positions can enhance the splicing of a central minigene exon (Xue et al., 2009). Examining the predicted binding sites near the exons adjacent to the PTBP1 enhanced exons; we find a slight enrichment of sites adjacent to the 3’ splice site of the downstream exons (Figure 3A). This enrichment is small when examining these exons as a group. However, this does not rule out individual introns within the set exhibiting this characteristic.

The location of PTBP1 binding sites within its known target exons is variable. To visualize the heterogeneity between individual exons, we created heat maps of the binding scores upstream, within, and downstream of each exon in the PTBP1 target set (Figure 3B). We found that 60% of PTBP1 repressed exons are predicted to have strong binding sites within the upstream intron. Most of these exons also have strong binding sites within either the exon or the downstream intron. However, other patterns are also seen. Some repressed exons score highly for PTBP1 binding only within the exon or in both the exon and the downstream intron. About half of PTBP1 enhanced exons have strong PTBP1 binding sites downstream (Figure 3B). These can co-occur with upstream intron-binding sites, but rarely with exon binding sites. Interestingly, there are exons enhanced by PTBP1 with strong upstream binding in the absence of other sites. These data demonstrate the heterogeneity in the positions of PTBP1 binding sites for its target exons. This heterogeneity needs to be considered for predicting PTBP1 dependent regulation.

PTBP1 repressed exons exhibited significantly higher average binding scores in both the upstream intron and in the exon itself, than either the control group of alternative exons or PTBP1 enhanced exons (Figure 3C). The average binding scores in the downstream introns were higher for both the PTBP1-repressed and PTBP1-enhanced exons than the control group (Figure 3C), although not at the same statistical significance. The variability of binding site placement within the smaller group of PTBP1-enhanced exons presumably contributes to the weaker statistical correlation of binding scores with positive regulation.

We also compared the three exon sets for other features that might contribute to their ability to be regulated by PTBP1, including exon length, flanking intron length, and 5' and 3' splice site strength. Most of these features were not statistically different among the three-exon groups. However, both PTBP1 enhanced and PTBP1 repressed exons were found to carry significantly weaker 3' splice sites than the control exon set, as measured by the Analyzer Splice Tool and shown in Figure 3C (Carmel et al., 2004; Shapiro and Senapathy, 1987).

These results indicate that PTBP1-repressed exons, and perhaps PTBP1-enhanced exons, exhibit an ensemble of sequence features that define them as PTBP1 regulated and should allow their identification by sequence alone.

Prediction of PTBP1 repressed exons

To predict new PTBP1 dependent exons, we developed a multinomial logistic regression model and trained it on the three sets of regulated exons described above (Hosmer and Lemeshow, 2000) (Figure 4A). Scores for sequence features were plotted against the percent of exons exhibiting that score that also exhibit PTBP1 dependent exon repression (Supplemental Figure 2). These plots produced distinct sigmoidal curves where most exons regulated by PTBP1 were found above or below a particular score. This strongly suggests that a logistic regression model incorporating each of these scores will be predictive of PTBP1 repression. The four features (x_1 through x_4) that correlated with PTBP1 repression were determined for each exon in the training set. These included the 3' splice site strength, and the PTBP1 binding score for each of three regions: the 250 nucleotides upstream of the exon, the exon itself, and the 100 nucleotides downstream of the exon. These intron lengths encompass the regions exhibiting enrichment of binding sites for PTBP1 dependent exons (Figure 3).

To model three exon classes (PTBP1 repressed exons, PTBP1 enhanced exons, and non-regulated exons), we used a multinomial logistic regression. We trained the model given in Figure 4A to determine the b coefficients. The PTBP1-enhanced exons are fewer in number and show more limited enrichment of PTBP1 binding sites than PTBP1-repressed exons making the prediction for these exons less accurate. We first tested models that considered just PTBP1-repressed exons relative to control groups. However, we found that these were less predictive than models that incorporated a separate group of PTBP1-enhanced exons (data not shown). Thus, including the enhanced exons as a training group improved the prediction of repressed exons.

As expected the upstream binding score was weighted most heavily in predicting PTBP1 repression (Supplemental Table 1), although binding scores in all three regions contributed to the score for PTBP1 repression. In contrast, we found that only the downstream binding score was significantly associated with PTBP1 enhancement. The upstream score generated a b coefficient close to zero making it essentially neutral in the prediction of enhanced exons. The exon binding score was subject to a negative b coefficient, indicating that exon binding reduces the probability of PTBP1 enhancement. Using these b coefficients, the trained models for repression or enhancement each yield a value of the g-function (logit) for an exon (x) given by the log of the ratio of the probability of repression or enhancement over the probability that the exon is not regulated. From this, the probability that an exon is repressed by PTBP1 can be determined from the two g-values as shown in Figure 4A.

We assessed the multinomial logistic regression model by recursively retraining on exon sets with one exon left out and then scoring the missing exon. This leave-one-out cross validation enabled assessment of the overall performance of the model (Hosmer and Lemeshow, 2000)(Supplementary Figure 3). The PTBP1 dependent exon repression logit showed good prediction, with an area under the curve value of 0.72, substantially greater than random guessing (AUC=0.5). As expected, the enhanced exon logit was not as accurate as the repression logit (AUC = 0.57), although it was better than random (Supplementary Figure 3A). Using these data, we assessed the sensitivity and specificity across the range of scores to define a decision threshold for exon repression scores (Supplementary Figure 3B). Increasing the threshold increases the specificity by eliminating many false positives, but decreases the sensitivity of the model in identifying maximum numbers of repressed exons. We sought to choose a threshold that minimized the false positive rate. We found that above a threshold score of 0.65 the false positive rate was 10% or lower (Supplementary Figure 3B).

Applying the model to 4494 alternative cassette exons, we found 243 exons (5.4%) that yielded a repression probability score greater than 0.65. The 50 top-scoring cassette exons are listed in Table 1. These included known PTBP1 target exons that were not in the training set. An exon of Gabrg2 yields a probability score of 0.92. Although we could not confirm its repression in N2A cells because of low expression of the transcript, the orthologous exon in rat is a well-characterized PTBP1 repression target (Ashiya and Grabowski, 1997). Exon 2 of Ptbp3 (Rod1), another known PTBP1 target (Spellman et al., 2007), yielded a repression probability score of 0.89 and was clearly confirmed by RT/PCR after Ptbp1 knockdown (Figure 4B). We performed additional RT-PCR validation in triplicate of a series of high and low scoring exons from transcripts expressed in N2A cells (Figure 5 & Supplementary Figure 4). Seven of ten exons scoring above 0.65 were de-repressed after Ptbp1 knockdown in N2A cells, yielding a validation rate of 70%. The actual false positive rate is difficult to estimate because exons with high repression scores that are not affected by Ptbp1 depletion in N2A cells might be regulated by PTBP1 in other cells. An indication that this might be occurring is that the average inclusion level (or percent spliced in value, PSI) of the putative false positives is significantly higher than the confirmed true positives in N2A cells, indicating that they will be less prone to change upon Ptbp1 depletion and be more difficult to validate (Supplemental Figure 6A and B). Thus, the true positive rate may be greater than 70%. Importantly, the high validation rate for exons scoring above 0.65 indicates that the binding model and the regulation model based upon it can identify many new PTBP1 targets that were not previously known (Table1).

High scoring exons might fail to be validated because of regulation by other proteins. Knockdown of Ptbp1 induces expression of its close homolog Ptbp2, which targets some of the same exons (Boutz et al., 2007b) (Supplementary Figure 4). To test whether PTBP2 was also targeting the predicted PTBP1 repressed exons, we knocked down Ptbp2 or both genes in N2A cells and re-assayed the exons in triplicate (Supplemental Figures 5 & 6A). Although some exons showed greater inclusion in the double knockdown compared to depletion of Ptbp1 alone, this did not validate any additional predicted PTBP1 repressed exons. We did identify some high and low scoring exons showing more complex regulation by the two PTB proteins (Supplemental Figure 5A & 5B). We expect that a comprehensive model incorporating binding data for both proteins will help dissect their redundant and divergent target exons.

We also examined a set of low scoring exons (probability score < 0.2) by RT-PCR after Ptbp1 and/or Ptbp2 depletion (Figure 5B and Supplemental Figure 5B). All of these exons (8 of 8) failed to respond to the loss of PTBP1 and are likely true negatives. Thus, PTBP1 repression scores above 0.65 and below 0.2 were highly predictive for regulation and its absence, respectively. As expected, intermediate scores were less consistent in their predictive value (Supplemental Figure 6C). Some exons in the intermediate scoring group were affected by PTBP1 and will be interesting to assess further.

The prediction of PTBP1-repressed exons was improved by treating PTBP1-enhanced exons as a separate class, but the probability scores for PTBP1 enhancement did not consistently identify new PTBP1 target exons (data not shown). This is presumably due to the smaller number of exons in the training set and their heterogeneity, with some possibly being indirect targets. These predictions will likely improve with training on larger numbers of PTBP1 enhanced exons as they are identified.

We next tested the model on a genomewide scale, by applying it to a set of 168,111 mouse internal exons and ranking them by their probability of PTBP1 repression. This analysis yielded 3824 exons (2.3%) with probability scores above 0.65 for being repressed by PTBP1. Among other activities, these exons were enriched in genes that function in calcium ion transport, cytoskeletal organization, intracellular transport, and synaptic transmission, all functions affected by previously known PTB targets (Supplementary Table 3).

To assess splicing of this large set of predicted PTB targets, we used RNA-seq to generate a large dataset of exons that change after Ptbp1 knockdown. RNA from control and PTBP1-depleted N2A cells was subjected to high density short read sequencing on the Illumina HiSeq platform using a strand specific, paired end protocol (Parkhomchuk et al., 2009). Exons whose inclusion changed between the two samples were identified by alignment to an exon database and quantification of exon inclusion using the SpliceTrap program (Wu et al., 2011). After filtering for read coverage and removing the training set, we identified 573 alternative exons whose splicing was assayable in N2A cells. These exons exhibit changes in percent exon inclusion (delta PSI) ranging from -29% to 62% upon PTBP1 depletion. The exons were binned by their PTBP1 repression probability scores and plotted for their change in PSI (Figure 6). The average changes in splicing were significantly correlated with the repression probability. Exons scoring below 0.5 distributed around zero change in PSI, but above this score the average exon inclusion is altered by PTBP1 depletion. Most notably, exons with a repression probability score above 0.65 exhibited significantly larger changes in splicing than exons with lower scores. Exons with intermediate scores and hence weaker binding sites show smaller changes in splicing than high scoring exons. Setting a threshold of a 5% change in PSI as validation, 22 of 33 exons

(67%) that scored above 0.65 for PTBP1 regulation were confirmed as PTBP1 repression targets in N2A cells. At least some of the other 11 exons are presumably PTBP1 targets in other cells.

From these results, we can now scan genomic sequence to score exons for PTBP1 regulation. Applying the logistic model genomewide, the PTBP1 repression probability scores were integrated into the UCSC genome browser. These data, displayed with the RNAseq data from N2A cells are available at our website (http://www.mimg.ucla.edu/faculty/black/ptbatweb/). A new PTBP1 repressed exon in the Kcnq2 gene is shown in Figure 6B. The logistic model thus provides a tool allowing the assessment of any exon across the transcriptome for likely PTBP1 regulation.

DISCUSSION

New Features of PTBP1 Binding sites

We have developed two computational models, one that allows accurate prediction of PTBP1 binding sites and another that predicts likelihood of PTBP1 regulation of exons across the transcriptome. These models uncovered several new features of RNA recognition by PTBP1 and the properties of its target exons. The PTBP1 binding model was based on triplets following the structures of the PTBP1 RRM domains, whose sequence specific contacts are each primarily to three nucleotides. We find that the set of triplets that increase the probability of binding includes the expected pyrimidine motifs, particularly those with alternating cytosines and uridines. However, many triplets with guanosine residues also increase binding probability. In contrast, adenosine residues have a negative effect on binding. Thus, RNA recognition by PTBP1 is not solely dependent on pyrimidine nucleotides. The recognition of G residues by PTB was unexpected, although some previously characterized PTB binding sites did contain G residues

(Reid et al., 2009; Xue et al., 2009). With this model, we can now predict PTBP1 binding affinity to any site in the transcriptome.

The base-specific contacts that PTBP1 makes with Guanosine are not yet clear. Recent studies of RNA recognition by SRSF2 (SC35) protein have shown that the element GGAG can be recognized by the same RRM as CCAG by flipping the initial two G nucleotides to the syn conformation (Daubner et al., 2012). It will be very interesting to investigate whether a similar anti to syn switch occurs in RNA bound by PTBP1, when C residues are replaced with G.

Previous characterizations of PTBP1 binding sites have focused on finding enriched short motifs within populations of bound RNAs or regulated exon sequences (Perez et al., 1997; Ray et al., 2009; Reid et al., 2009; Singh et al., 1995; Xue et al., 2009). These methods generally identify elements whose short length will allow interaction with only one RRM domain. Searching for new binding sites comprised of clusters of these short elements can identify higher affinity sites but does not consider all elements or rank them. The use of the HMM to rank all triplets allows more complex clusters of gapped short elements to be assessed for binding.

Many RNA binding proteins are similar to PTBP1 in having multiple domains that may each make different base specific contacts with RNA. The widespread generation of transcriptome-wide datasets from crosslinking-immunoprecipitation will allow the modeling of RNA recognition by almost any protein based on a large number of known binding sites. We are particularly interested in applying this approach to PTBP2 (nPTB), which should allow more precise comparison of RNA recognition by these two paralogs.

Defining PTBP1 target exons

Several PTBP1 target exons have been analyzed in detail (Garcia-Blanco et al., 1989). These exons vary in the placement and action of their PTBP1 binding sites. It is common for PTBP1-repressed exons to have a binding site upstream, often encompassing the branch point of the 3' splice site (Ashiya and Grabowski, 1997). Exons can also be repressed by PTBP1 binding within the exon (Shen et al., 2004). Other exons contain downstream binding sites that are needed in conjunction with an upstream site to achieve splicing repression (Gooding et al., 1998). Although acting as a repressor for most of its targets, PTBP1 also activates the splicing of a group of exons. There have been divergent reports about placement of PTBP1 binding sites needed to mediate PTBP1 enhancement of splicing. Some exons are stimulated by a PTBP1 site in the intron sequence immediately downstream (Llorian et al., 2010). However, in enhancing other exons, PTBP1 appears to bind near the adjacent constitutive exons (Xue et al., 2009).

The HMM based PTBP1 binding model allowed us to examine PTBP1 binding site placement across a large set of known PTBP1 target exons. Nearly all exons had predicted high affinity PTBP1 binding sites nearby. We find that more than half of PTBP1 repressed exons have high affinity binding sites upstream, and a fraction of PTBP1 enhanced exons have high affinity sites downstream. These exons fit with recent results on several other splicing regulators where the placement of the binding site determines the direction of the regulatory effect (Ule et al., 2006; Wang et al., 2012; Yeo et al., 2009). However, for PTBP1 these rules are not as generalizable. Some PTBP1 repressed exons have their strongest predicted binding site downstream or within the exon. Some PTBP1 enhanced exons have high affinity sites upstream and not downstream. This variability is reflected in the weighting factors (b factors) assigned to PTBP1 binding sites in predicting regulation of an exon by the protein.

To quantify the predictive value of the PTBP1 binding scores for PTBP1 repression, we built a logistic model for PTBP1 regulation. For exons repressed by PTBP1, binding scores for the upstream, downstream, and exon sequences all contribute to the probability of repression. Exons enhanced by PTBP1 were too few to achieve accurate predictions from the model. However, treating these as a separate exon class improves the prediction of PTBP1 repression. We find that for probability scores above 0.65 the model is strongly predictive of PTBP1 repression. Applying this criterion across the transcriptome, we identified hundreds of new PTBP1 target exons.

Alternative exons are generally regulated by multiple proteins acting in combination, and a particular exon will often be subject to both positive and negative regulation by antagonistic factors. For a model based on one factor, these other proteins will confound predictions. Exons with high PTBP1 binding scores may be counteracted by antagonistic factors in some cell types. Alternatively, synergistic factors may allow an exon with a relatively weak binding site to still recruit PTBP1. In this study, our intent was to measure the effect of PTBP1 binding alone before considering the contributions of other factors. The logistic modeling allowed the contributions of different binding site placements to PTBP1 regulation to be measured.

Several studies have used Bayesian models to dissect the regulatory properties of exons (Barash et al., 2010; Zhang et al., 2010). These models can generate accurate predictions by incorporating a wide variety of sequence, expression and conservation data. However, because so many disparate variables can be incorporated, it can be difficult to draw mechanistic conclusions from these models regarding any one protein. For example, the presence of high pyrimidine density upstream from the branch point can be predictive of exons showing neuronal specific inclusion (Barash et al., 2010; Castle et al., 2008). This is presumably in part due to many neuronal exons being regulated by PTBP1 and PTBP2. However, a subset of these exons may be regulated by other factors with pyrimidine rich binding sites. In the long term, it will be most accurate to develop predictive binding models for each protein, similar to the PTBP1 model here, and then to incorporate each of these binding models into a larger network model. Such an approach will allow the analysis of the many overlapping regulatory programs controlled by RNA binding proteins.

METHODS

Hidden markov model for PTBP1 binding affinity prediction

A Hidden Markov Model (HMM) was designed and trained by an expectation–maximization (EM) method (Baum-Welch algorithm) using published PTBP1 CLIP data (Durbin, 1998; Rabiner, 1989; Xue et al., 2009). In total, 48,604 PTBP1-CLIP cluster sequences were used to train model parameters. Average length of the CLIP cluster tags was 29nt (1st and 3rd quartiles were 16nt and 33nt, respectively). During the training step, multiple initial values were tested to avoid a local maximum problem. Trained parameters included emission probabilities for nucleotide triplets, initial probabilities and transition probabilities between states (Durbin, 1998; Rabiner, 1989). The trained model was used to score mouse exon and intron sequences. From mouse genome and annotations, we were able to retrieve 168,111 internal exons and their flanking introns using a python library, Pygr. Then, for each sequence, raw PTBP1 binding scores were calculated. We define the raw PTBP1 binding score as a log-likelihood ratio, the log ratios of the probability of the sequence given the PTBP1 binding model compared to a random model (Durbin, 1998). For the random model, we assumed uniform distribution of emission and transition probabilities. Since the raw PTBP1 binding scores were dependent on their length and location (upstream intron, exon, or downstream intron), we grouped sequences and performed score normalization for individual sequence group (Durbin, 1998). For each group, an average and a standard deviation of raw PTBP1 binding scores were calculated and used to transform the raw PTBP1 binding scores to z-scores (Supplementary Figure 1). Throughout the paper, we called the z-score as a final predicted PTBP1 binding score. Additionally, for each RNA sequence, we predicted PTBP1 binding sites using the Viterbi algorithm (Durbin, 1998; Rabiner, 1989)

Validation of PTBP1 binding model scores by binding assay

To test predicted PTBP1 binding scores, we selected thirteen mouse exon/intron RNA sequences (69 nucleotides) with various scores. In the selection, other sequence features were not considered. Target RNA were *in vitro* transcribed from dsDNA using a T7 promoter and subjected to an electrophoretic mobility shift assay (EMSA). During the transcription, radioactive α-32P UTP were incorporated into RNA to visualize probes. The RNA probes were then denatured for 2 min at 85°C and cooled down on ice immediately to prevent secondary structure formation. Binding assay was carried out as previously described with some modifications (Amir-Ahmady et al., 2005). In detail, each gel mobility shift reaction (10 μL) contained 6 μL DG buffer with recombinant human PTBP1 (20 mM Hepes-KOH ph7.9, 20% glycerol, 80 mM potassium glutamate, 0.2 mM EDTA, 0.2 mM PMSF, plus recombinant human PTBP1), 1 μL 22 mM MgCl2, 1 μL 0.5 mg/ml tRNA, 0.5 μL RNase inhibitor (20 unit, RNaseOut from invitrogen), 0.5 μL DEPC treated H2O, and 1 μL 100 nM RNA probes. At first, all reaction components excluding RNase inhibitor, tRNA, and RNA probes were mixed and incubated for 8 min at 30°C. Then RNase inhibitor and tRNA were added and mixed. The last component, RNA probes were added and incubated for an additional 15 min. The reactions were put on ice for 5 min and mixed with 1.2 µL glycerol loading dye (30% glycerol). They were separated on 8% native polyacrylamide gels with 25mM Tris-Gly running buffer in a cold room. Gels were dried and exposed to a phosphor screen. Then images were scanned using Typoon 9410 and quantified using ImageQuant TL program (GE Lifesciences). The apparent *Kd* values were estimated by fitting the data to non-linear curves using Prism software.

Logistic regression model for PTBP1 dependent exon prediction

An exon training set was compiled from previous microarray and RT-PCR experiments (Boutz et al., 2007b; Xing et al., 2008). The training set was composed with 68 PTBP1 repressed, 37 PTBP1 enhanced, and 69 non-PTBP1 regulated simple cassette exons. In this study, we only considered exons with canonical splice sites (GU-AG) in their introns. Exons were classified as a PTBP1 repressed or enhanced exon when 1) the inclusion level (PSI) of its minor isoform is greater than 5% in both control and knock-down sample and 2) the inclusion level of minor isoform was changed over 30% in *Ptbp1* knock down condition compared to the control sample. Next, we collected sequence features for each exon and their flanking locus. The features included PTBP1 binding scores, 5' and 3' splice site strengths, exon/intron lengths, and word frequencies. The PTBP1 binding scores were calculated from the PTBP1 binding model described above. The strength of splice sites was calculated by the splice-site analyzer tool (Shapiro and Senapathy, 1987). Using a whole internal mouse exon set, we normalized features and fed them into the model. The PTBP1 splicing model is based on a multinomial logistic regression framework using the following steps: 1) selection of initial variables with a moderate level of association (p-value from t-test <0.25), 2) removal of outlier exons, 3) stepwise variable selections (Hosmer and Lemeshow, 2000). We scored mouse internal exons with the trained PTBP1 splicing model and validate candidate exons with RT-PCR and RNA-seq experiments. Exons from the training set were excluded from the validation.

Validation of exon candidates by RT-PCR and RNA-seq

To test alternative splicing events for candidate exons, we assayed exon inclusion levels in cells followed by *Ptbp1*, *Ptbp2*, and both *Ptbp1* & *Ptbp2* knock down. The knockdown experiment was performed as described before with minor modification (Boutz et al., 2007b). Mouse neuroblastoma (N2A) cells were cultured in DMEM with 10% FBS and 2mM L-glutamine. At 70 to 80% confluency, cells were trypsinized and suspended in the growth medium. DNA–Lipofectamine 2k (Invitrogen) complexes were prepared and mixed with cells in a tube according to manufacturer's instructions. Tubes were incubated for 5 h with mixing in every half hour. Then cells were centrifuged and cultured in plates for 3 d. Proteins and RNA was extracted from collected cells. Protein samples were subjected to fluorescence immunoblotting to monitor knockdown efficiency of *Ptbp1* and *Ptbp2*. Total RNA was collected using Trizol (Invitrogen) according to the manufacturer's instructions. The RNA was further treated with DNase I to avoid DNA contamination. For RT-PCR (Reverse Transcription-PCR) assays, the RNA was reverse transcribed to cDNA with random hexamers using SuperScript enzyme (Invitrogen) following the manufacturer's instructions. Then PCR reactions were performed to investigate alternative splicing events of the target exon. First, forward and reverse PCR primers were designed against to flanking exons using PRIMER3 program (Rozen and Skaletsky, 2000). To label PCR products, a 5' fluorescent-labeled universal primer (5'-FAM-CGTCGCCGTCCAGCTCGACCAG-3') was added to the PCR reaction and also a universal priming site was introduced to the 5’ end of the forward primer (5’-CGTCGCCGTCCAGCTCGACCAG-Forward Primer-3’). Each PCR reaction (15 µL) was carried out with 1.5 pico mole of the forward primers and 6.75 pico mole of the reverse and universal primers (Schuelke, 2000). PCR amplification proceeded with an initial denaturation at 94°C for 4 m followed by 24 cycles of 94°C for 30 s, at a melting temperature of the reverse primer for 45 s, and 72°C for 45 s, with a final extension step at 72 °C for 10 m. The samples were mixed with 2Xformamide buffer (Formamide with 1mM EDTA pH8.0) and denatured at 95°C for 5 min. Then samples were chilled on ice and run at 8%denaturing polyacrylamide gels. Gels were directly scanned by Typoon and quantified by ImageQuant program.

RNA-seq libraries were constructed following standard protocols (Illumina TruSeq RNA Sample Prep Kit). To make strand-specific libraries, we added two extra steps to the protocol (Parkhomchuk et al., 2009). After first strand cDNA synthesis, remaining dNTPs were removed by a size selection using beads (AMPure XP). Second-strand cDNA was synthesized with dNTP mix with dUTP instead of dTTP. The reaction contained samples eluted in 50µl resuspension buffer, 2µl 5XFS buffer, 1µl 50mM MgCl2, 1µl 100mM DTT, 2µl 10mM dUTP nucleotides mix, 15µl Second Strand Buffer (Invitrogen), 0.5µl E.coli DNA Ligase (10U/µl;NEB), 0.5µl RNase H(2U/µl;Invitrogen), 2µl DNA E.coli Polymerase I(10U/µl;NEB). The reaction was incubated for 2h at 16°C. After sequencing adaptors were ligated, 1µl USER (Uracil-Specific Excision Reagent enzyme;NEB) was added to reactions to degrade the second strand cDNA with uracil. The samples were incubated for 15min at 37°C and the reaction were inactivated at 94°C for 5min. The samples were put in ice and preceded to next PCR amplification step. Average size of inserts was about 225bp and the libraries were subjected to 100bp paired-end sequencing (Illumina HiSeq2000 platform). Using SpliceTrap (Wu et al., 2011), 60-65% of reads were mapped to exon duos or trios. In total, 180M (179,511,116) and 145M (145,334,711) paired end reads were used to infer exon inclusion ratios for control and *Ptbp1* knockdown conditions, respectively.

TABLES AND FIGURES

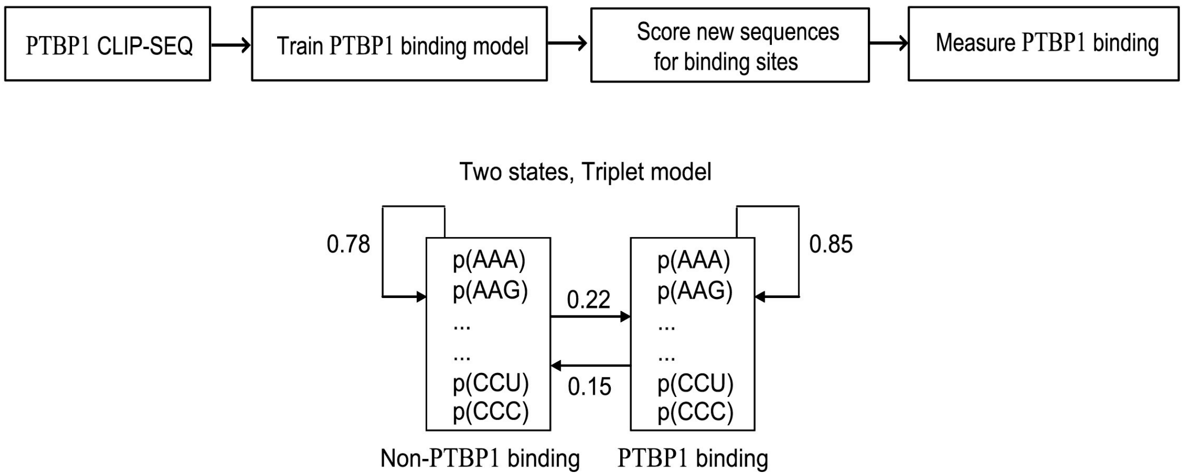


Figure 1. A. Scheme of the PTBP1 binding model. The PTBP1 binding model was trained by PTBP1 bound RNA sequences (48,604 clusters) from published PTBP1-CLIP experiments and was used to predict PTBP1 binding scores for new RNA sequences. The validity of scores was assessed by PTBP1-RNA binding assay. Diagram presents structure of the PTBP1 HMM (Hidden Markov Model) model and trained transition probabilities.

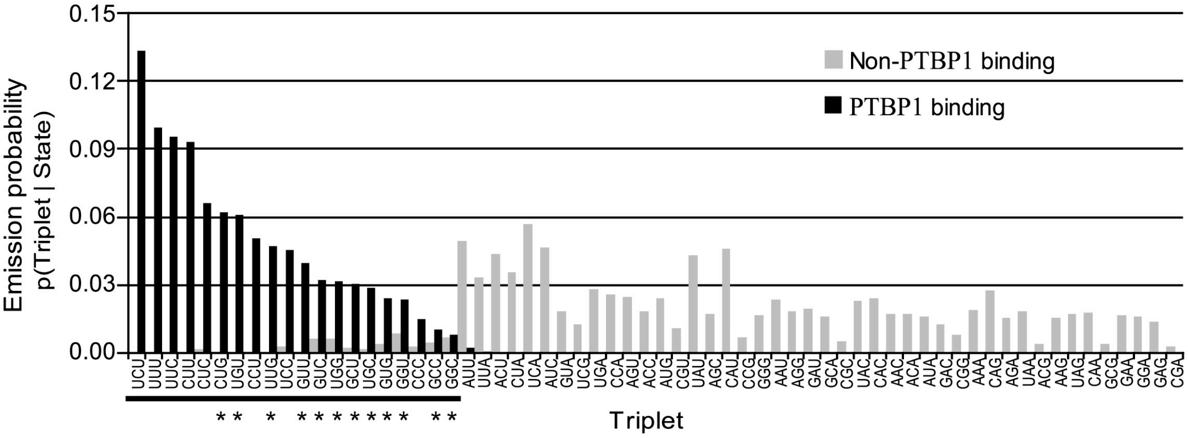
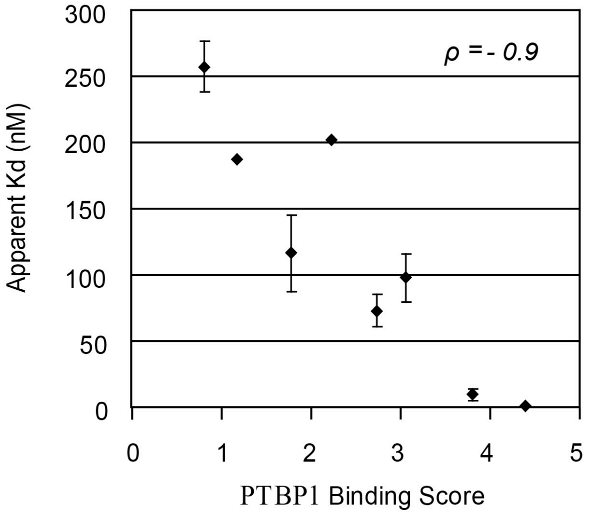


Figure 1. B. Emission probabilities to observe a certain triplet in PTBP1 binding state and non-PTBP1 binding state are plotted in black and gray bars, respectively. Asterisks indicate G containing pyrimidine triplets.



RNA Probe No.	PTBP1 Binding Score	Apparent Kd (Avg.) nM	Std. Err.
1	-2.62	No binding	-
2	-1.35	No binding	-
3	-0.55	No binding	-
4	-0.48	20% binding	-
5	-0.43	No binding	-
6	0.82	257.03	19.12
7	1.18	187.25	-
8	1.78	116.22	29.31
9	2.24	201.80	-
10	2.74	73.01	11.97
11	3.06	97.87	18.10
12	3.80	9.43	4.37
13	4.40	1.11	0.21

Figure 2. A. Validation of PTBP1 binding model and prediction examples. To validate binding scores, thirteen RNAs with various PTBP1 binding scores were transcribed *in vitro* and subjected to binding assay. Apparent *Kd* (dissociation constant) was highly negatively correlated with PTBP1 binding scores (Pearson correlation = - 0.9).

Binding score(z)	p-value	RNA sequences
-2.62	0.997	GC**GCUGC**AGGAGGAGCGCAAGGCCGAGGA**CUGCUC**GCCGAGUA**GGCUC**GCACGGCAAGUA**GGCUCC**GAG
-0.43	0.659	GGAGAAGUAGGUCAGAUGGGACCCACAG**GGCCU**AGAGGACCAAUGGGCA**UUGG**AGUACAG**GGUCC**AAAG
0.82	0.202	**GCCC**AAGGAGGUGAUAGCA**UUUUUC**AGAGA**UUG**AAAAGAAUAGUA**UUUG**A**UGCC**AAA**UCU**ACAA**UUGUG**
2.74	0.005	**GUG**AG**GGCUCUUUCUUCGUGG**GACCGUAGAUAGGUA**GCUGCUGCUGGUCUC**ACA**CUGUUCUCCCU**ACAG

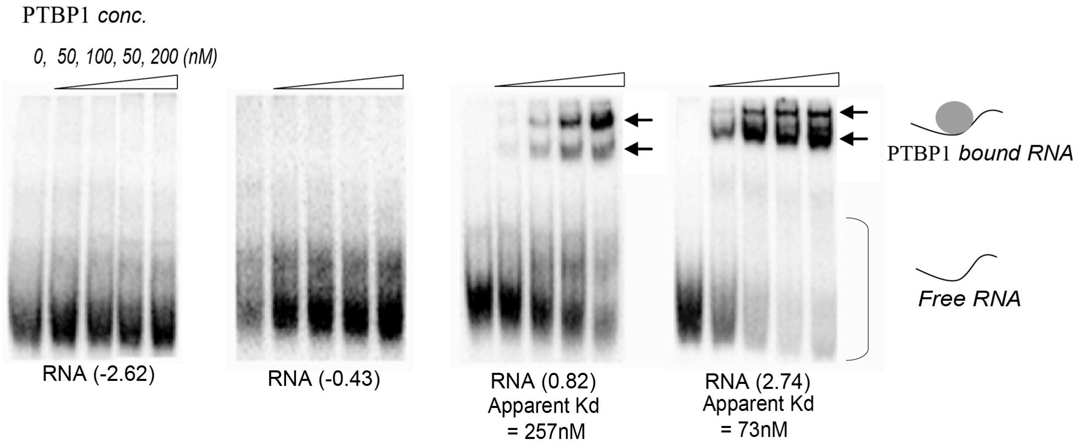


Figure 2. B. As prediction examples, four RNA sequences are presented with predicted PTBP1 binding scores. Potential PTBP1 binding sites are underlined and in bold. Experimental binding affinities were assessed by mobility retardation of RNA by PTBP1-RNA binding and compared with prediction scores. Apparent dissociation constants (*Kd*) were determined by the concentration at which half the protein was bound to RNA.

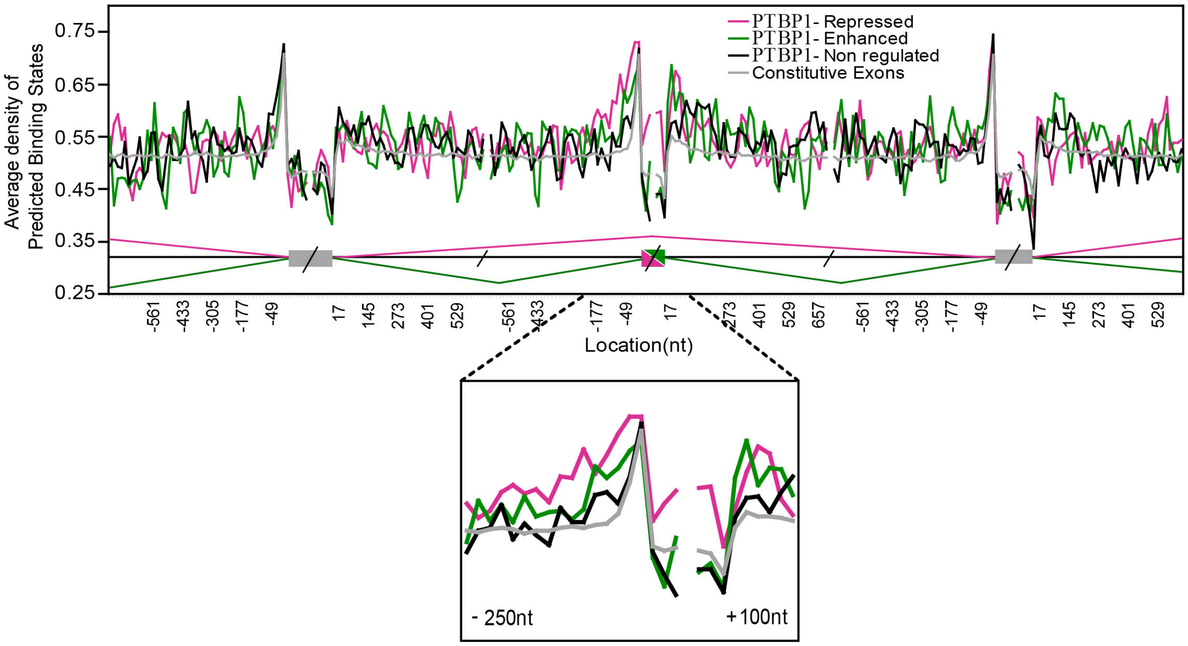


Figure 3. A. Sequence characteristics of PTBP1-dependent alternatively spliced exons. An RNA map shows enrichment of predicted PTBP1 binding sites near PTBP1-dependent exons. The Y-axis represents average density of predicted PTBP1 binding states within a 24nt window; the length of overlap between two adjacent windows was 8nt.

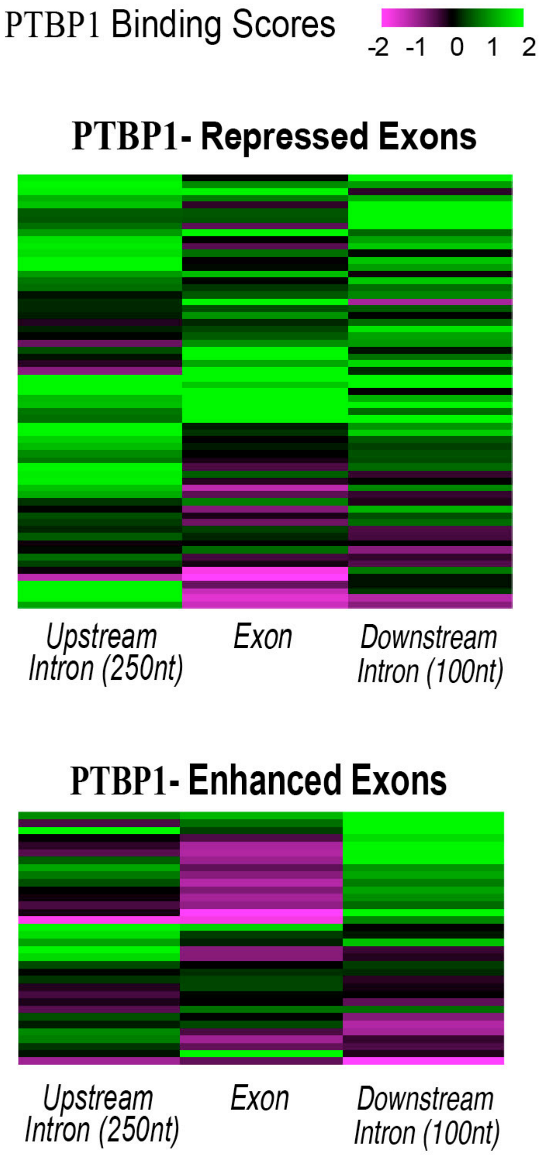


Figure 3. B. To assess PTBP1 binding signatures of individual exons, known PTBP1 regulated exons were clustered by their PTBP1 binding score profiles and visualized as heat maps. The heat maps indicate that there is wide variation in the positions of PTBP1 binding sites between individual exons.

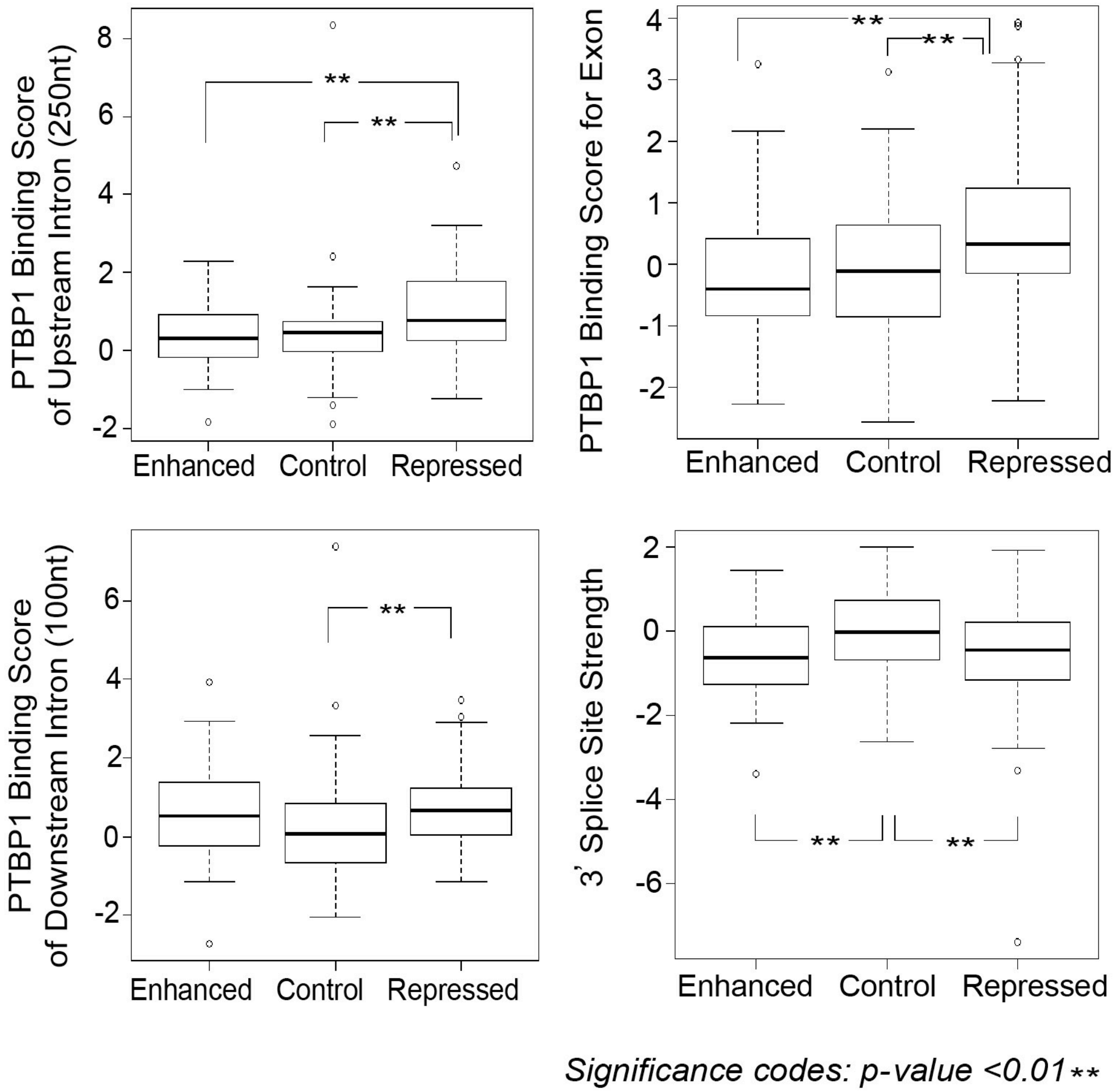


Figure 3. C. Four sequence features including the PTBP1 binding scores and 3' splice site strength show statistically significant differences between regulated and control exon groups (one-tailed Student's t-tests).

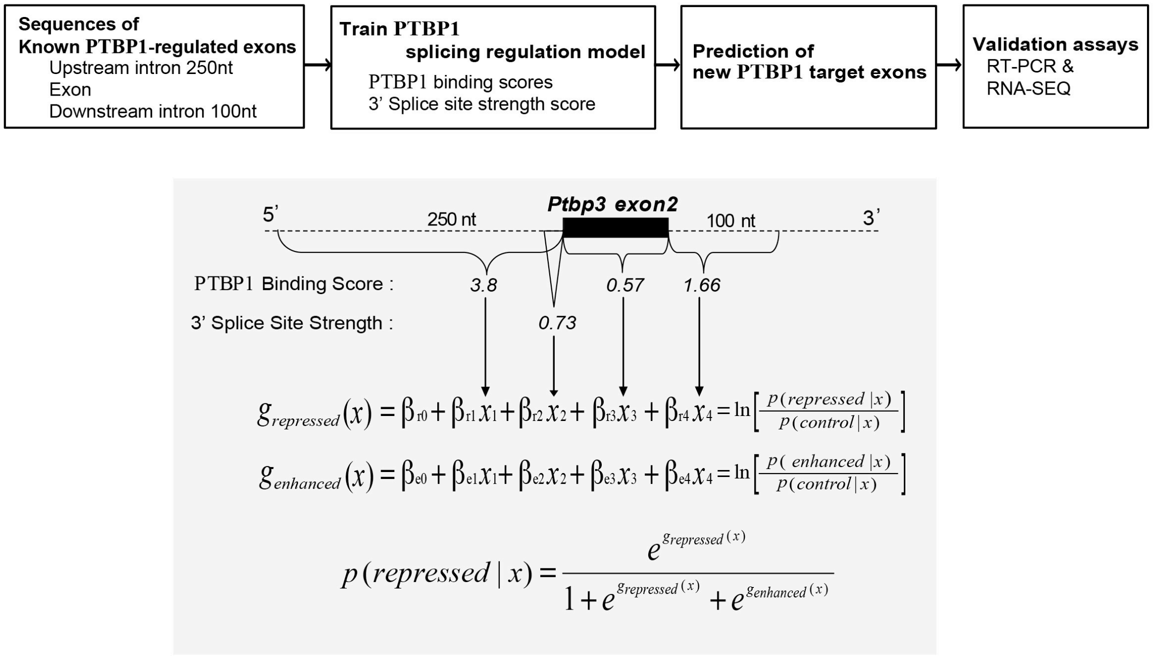

Sequences of
Known PTBP1-regulated exons
Upstream intron 250nt
Exon
Downstream intron 100nt
Train PTBP1
splicing regulation model
PTBP1 binding scores
3' Splice site strength score
Prediction of
new PTBP1 target exons
Validation assays
RT-PCR &
RNA-SEQ
5'
250 nt
Ptbp3 exon2
100 nt
3'
PTBP1 Binding Score :
3' Splice Site Strength :
3.8
0.73
0.57
1.66
$g_{repressed}(x) = \beta_{r0} + \beta_{r1}x_1 + \beta_{r2}x_2 + \beta_{r3}x_3 + \beta_{r4}x_4 = \ln\left[\frac{p(repressed\,|\,x)}{p(control\,|\,x)}\right]$
$g_{enhanced}(x) = \beta_{e0} + \beta_{e1}x_1 + \beta_{e2}x_2 + \beta_{e3}x_3 + \beta_{e4}x_4 = \ln\left[\frac{p(\,enhanced\,|\,x)}{p(control\,|\,x)}\right]$
$p(repressed\,|\,x) = \frac{e^{g_{repressed}(x)}}{1 + e^{g_{repressed}(x)} + e^{g_{enhanced}(x)}}$

Figure 4. A. Scheme of the PTBP1 splicing regulation model and its application to an exon in *Ptbp3*. The PTBP1 splicing regulation model was trained on known PTBP1-regulated and non-regulated exons and used to predict new PTBP1-dependent exons. Prediction results were compared to changes in exon inclusion (PSI) measured by RT-PCR and RNA-seq. An exon from *Ptbp3* is presented as a prediction example. From intron and exon sequences, PTBP1 binding scores and 3' splice site strength were calculated and fed into the regulation model.

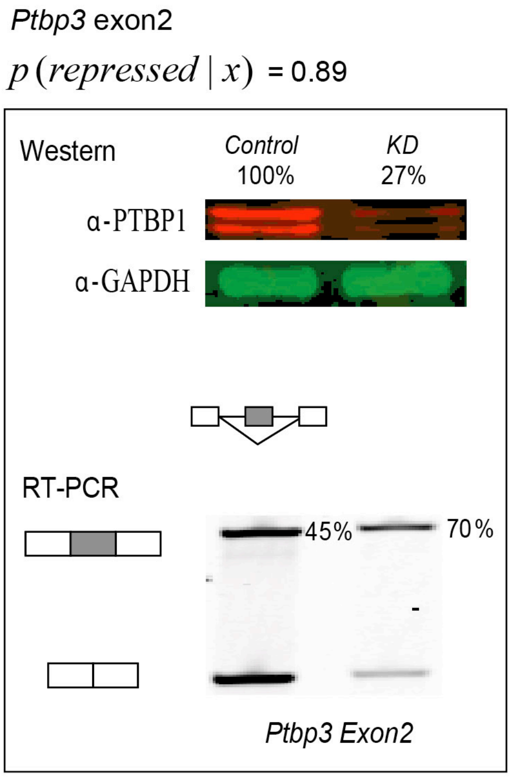


Figure 4. B. The model predicts exon 2 of *Ptbp3* as repressed by PTBP1 with high probability (0.89). *Ptbp1* knockdown in mouse neuroblastoma cells (*N2A*) confirmed de-repression of the exon (from PSI=45 to PSI=70).

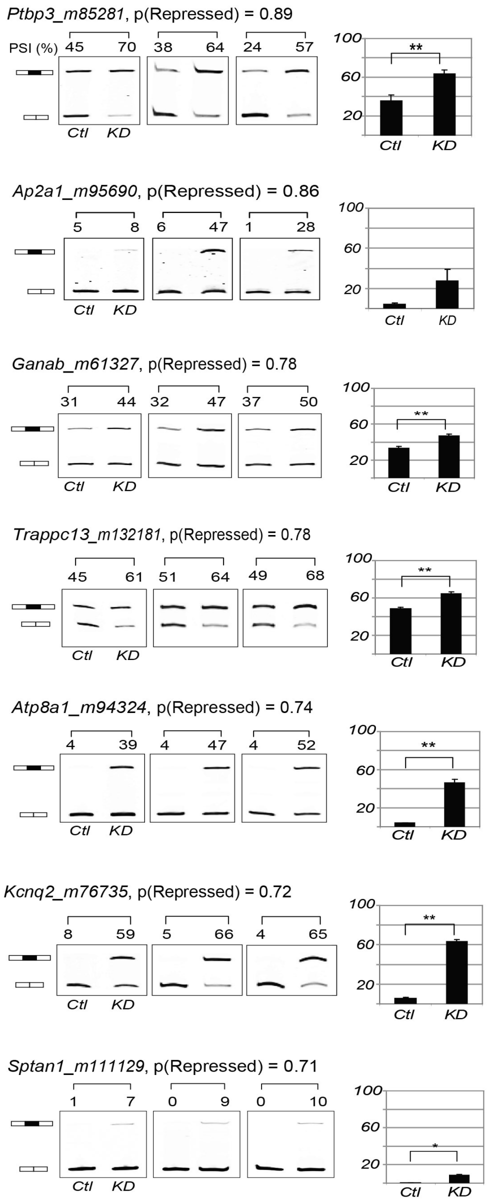


Figure 5. A. Validation of novel PTBP1-repressed exons by RT-PCR. Candidate PTBP1-repressed exons with probability greater than 0.65 were validated by RT-PCR following Ptbp1 knockdown. Data shown are averages ± standard error of PSI (Percent of Spliced In) in biological triplicates. Statistical analysis was performed using paired one-tailed Student's t-test (p-values < 0.01**, < 0.05*).

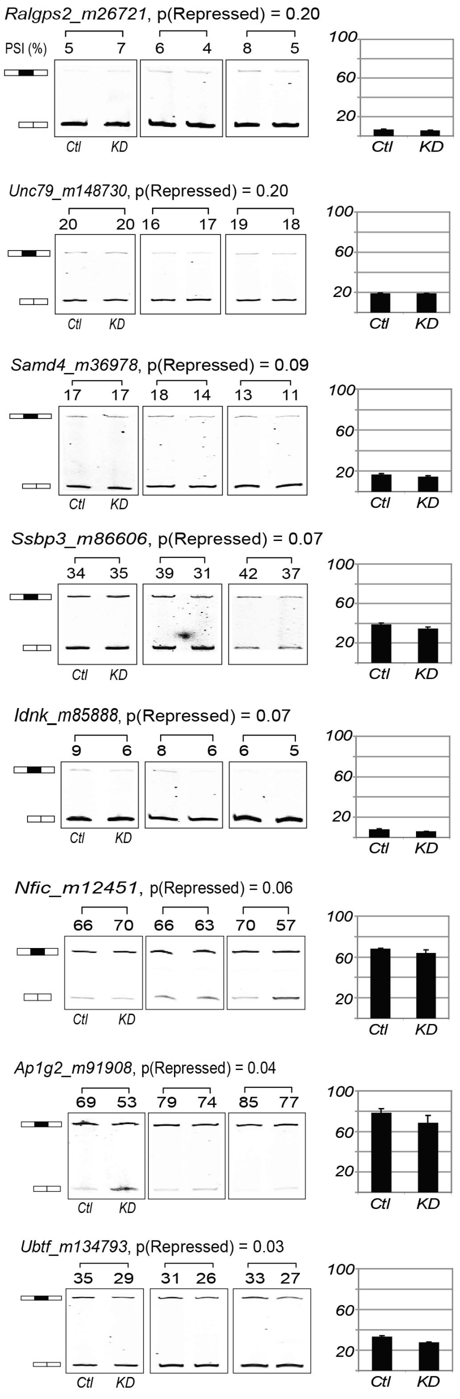


Figure 5. B. Exons with low PTBP1 repression probabilities (≤0.2) were also validated by RT-PCR following *Ptbp1* knockdown in biological triplicates.

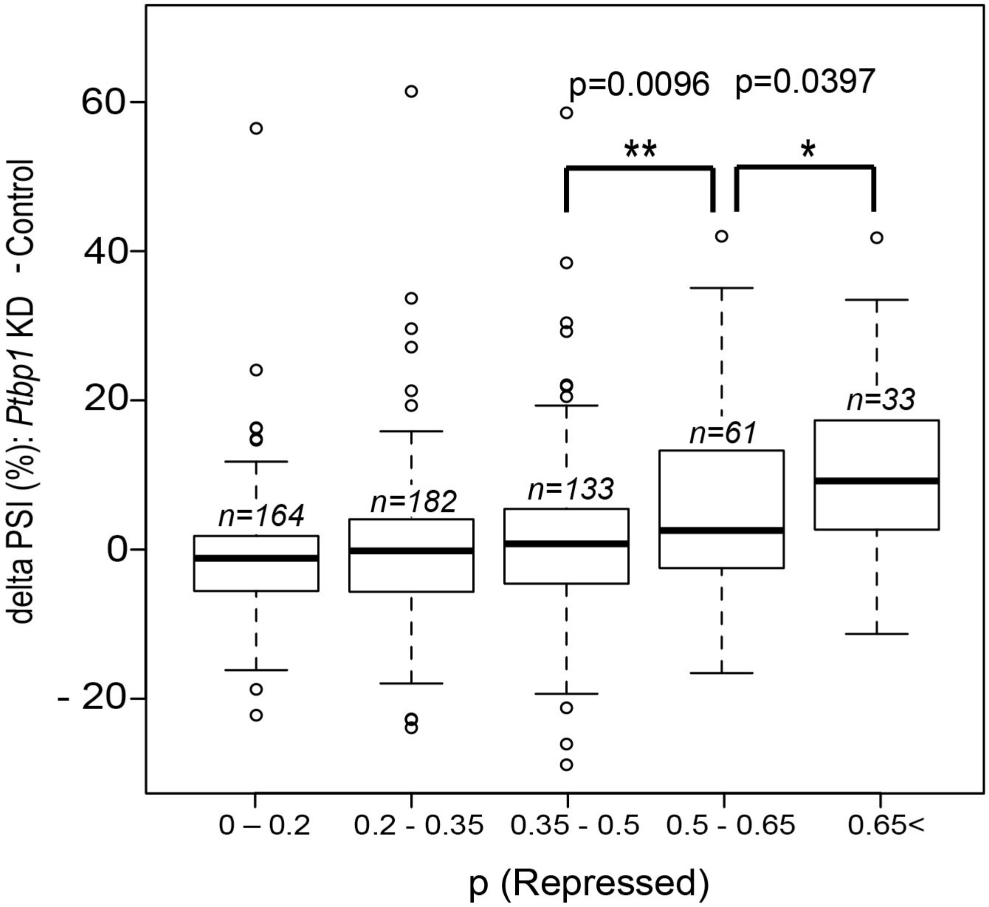


Figure 6. A. Validation of the PTBP1 splicing model using RNA-seq. After *Ptbp1* knockdown, we performed RNA-seq experiments and estimated changes in PSI (Percent of Spliced In) for 573 cassette exons. The graph shows average delta PSI values for exons, grouped by their probabilities to be repressed by PTBP1. The number of exons in the corresponding probability bin is given by n. P-values were calculated from one-tailed Student's t-test.

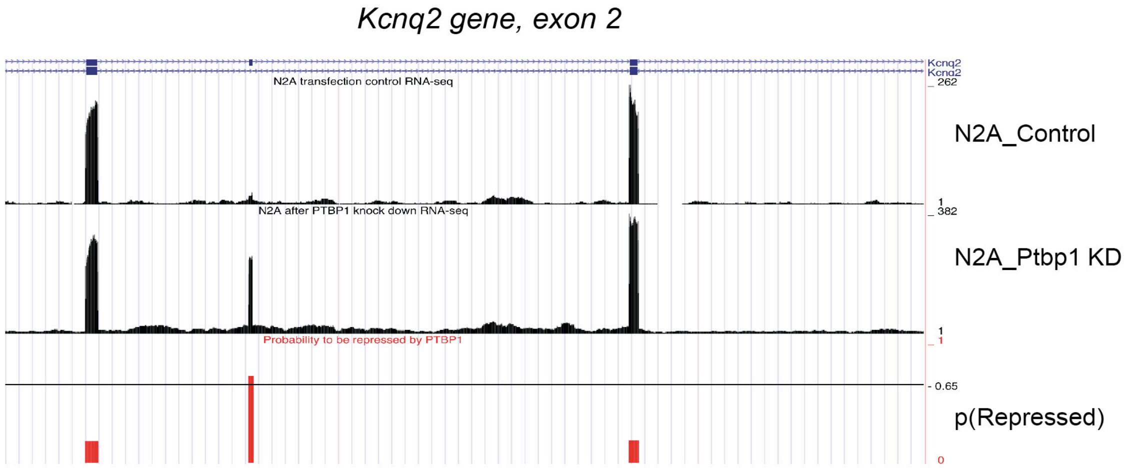

Kcnq2 gene, exon 2
Kcnq2
Kcnq2
N2A transfection control RNA-seq
262
N2A_Control
1
N2A after PTBP1 knock down RNA-seq
382
N2A_Ptbp1 KD
1
Probability to be repressed by PTBP1
1
0.65
p(Repressed)
0

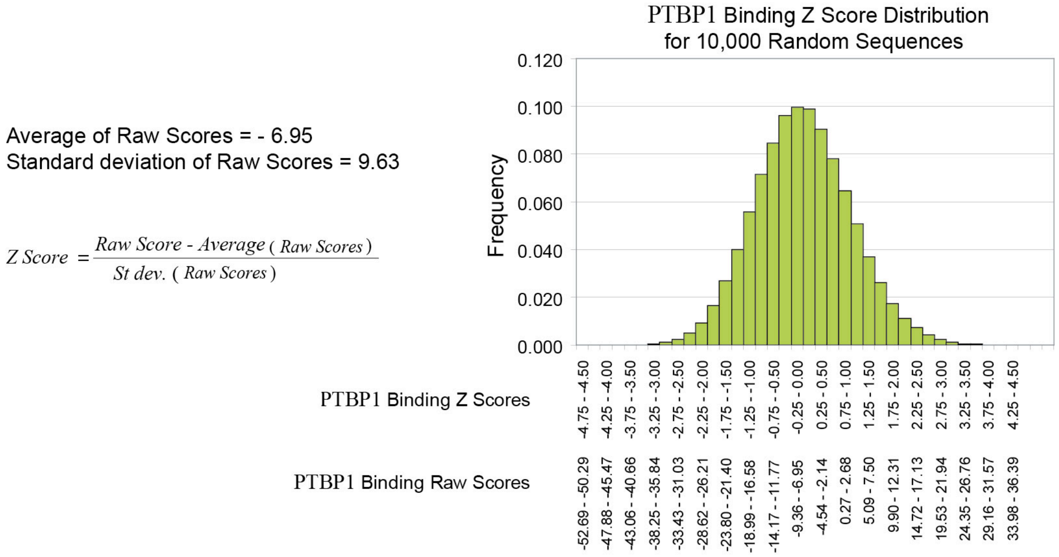
Figure 6. B. A genome browser screenshot of a novel PTBP1-regulated exon: exon 2 of the *Kcnq2* gene. For whole internal mouse exons, we created custom genome browser tracks to visualize the PTBP1 splicing model and mapped RNA seq reads.

Table 1. The PTBP1-dependent splicing model identified novel PTBP1 repressed exons.

This table lists the top 50 exons predicted to be repressed by PTBP1.

Gene Name	Gene Description	mm9 coordinates	p(Repressed)
Pax6	paired box gene 6	chr2:105523985-105524115(+)	0.99
Mbd5	methyl-CpG binding domain protein 5	chr2:49134101-49135303(+)	0.98
Arhgap24	Rho GTPase activating protein 24	chr5:102981145-102981338(+)	0.97
Tle1	transducin-like enhancer of split 1	chr4:71819247-71819451(-)	0.94
Acsl6	acyl-CoA synthetase long-chain family	chr11:54150438-54150515(+)	0.94
Ryr1	ryanodine receptor 1, skeletal muscle	chr7:29829938-29829955(-)	0.94
Ankhd1	ankyrin repeat and KH domain containing 1	chr18:36784163-36784921(+)	0.93
Slc39a14	solute carrier family 39 (zinc transporter)	chr14:70713408-70713577(-)	0.92
Gabrg2	gamma-aminobutyric acid (GABA) A receptor	chr11:41727472-41727495(-)	0.92
Itga7	integrin alpha 7	chr10:128378878-128378997(+)	0.92
Iqsec2	IQ motif and Sec7 domain 2	chrX:148615540-148615635(+)	0.91
Smarca2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	chr19:26825612-26825646(+)	0.91
Zfand3	zinc finger, AN1-type domain 3	chr17:30197755-30197795(+)	0.91
Agap2	ArfGAP with GTPase domain, ankyrin repeat and PH domain 2	chr10:126527198-126527257(+)	0.90
Ttn	Titin	chr2:76723554-76723832(-)	0.90
Ptbp3	ROD1 regulator of differentiation 1 (S. pombe)	chr4:59559021-59559054(-)	0.89
Mapk8	mitogen-activated protein kinase 8	chr14:34203859-34203930(-)	0.89
Snap91	synaptosomal-associated protein 91	chr9:86693373-86693534(-)	0.88
Fmnl1	formin-like 1	chr11:103059449-103059547(+)	0.88
Phldb1	pleckstrin homology-like domain, family B	chr9:44509029-44509169(-)	0.87
2310035C23Rik	RIKEN cDNA 2310035C23 gene	chr1:107637012-107637094(+)	0.87
Arnt	aryl hydrocarbon receptor nuclear translocator	chr3:95270715-95270759(+)	0.87
Smyd2	SET and MYND domain containing 2	chr1:191723697-191723807(-)	0.86
Ap2a1	adaptor protein complex AP-2, alpha 1 subunit	chr7:52158832-52158897(-)	0.86
Klra	killer cell lectin-like receptor, subfamily A	chr6:130329011-130329100(-)	0.86
Spag9	sperm associated antigen 9	chr11: 93942054-93942068(+)	0.86
Col4a3bp	collagen, type IV, alpha 3 binding protein	chr13:97386949-97387026(+)	0.86
Garnl3	GTPase activating RANGAP domain-like 3	chr2:32941395-32941464(-)	0.86
Dennd1a	DENN/MADD domain containing 1A	chr2:37982049-37982168(-)	0.86
Ms4a7	membrane-spanning 4-domains, subfamily A	chr19:11400297-11400353(-)	0.86
BC030307	cDNA sequence BC030307	chr10:86169981-86170089(+)	0.85
Phactr1	phosphatase and actin regulator 1	chr13:43154940-43155146(+)	0.85
R3hdm2	R3H domain containing 2	chr10:126902187-126902240(+)	0.84
Cdc14b	CDC14 cell division cycle 14B	chr13:64306579-64306725(-)	0.84
Ubqln1	ubiquilin 1	chr13:58282183-58282266(-)	0.84
Ttn	Titin	chr2:76739898-76740179(-)	0.84
Stx3	syntaxin 3	chr19:11857290-11857400(-)	0.84
Slc8a3	solute (sodium/calcium) carrier family 8	chr12: 82310340-82310458(-)	0.84
Zfp62	zinc finger protein 62	chr11:49028057-49028156(+)	0.83
Dlg1	discs, large homolog 1 (Drosophila)	chr16:31771843-31771941(+)	0.83
Nrxn2	neurexin II	chr19:6463824-6463847(+)	0.83
Klra7	killer cell lectin-like receptor, subfamily A	chr6:130179953-130180042(-)	0.83
Picalm	phosphatidylinositol binding clathrin assembly	chr7:97330729-97330878(+)	0.83
Acad8	acyl-Coenzyme A dehydrogenase family	chr9:26798168-26798277(-)	0.83
Epn1	epsin 1	chr7:5033620-5033723(+)	0.82
Grip1	glutamate receptor interacting protein 1	chr10:119422530-119422685(+)	0.82
Csmd3	CUB and Sushi multiple domains 3	chr15:47587514-47587627(-)	0.82

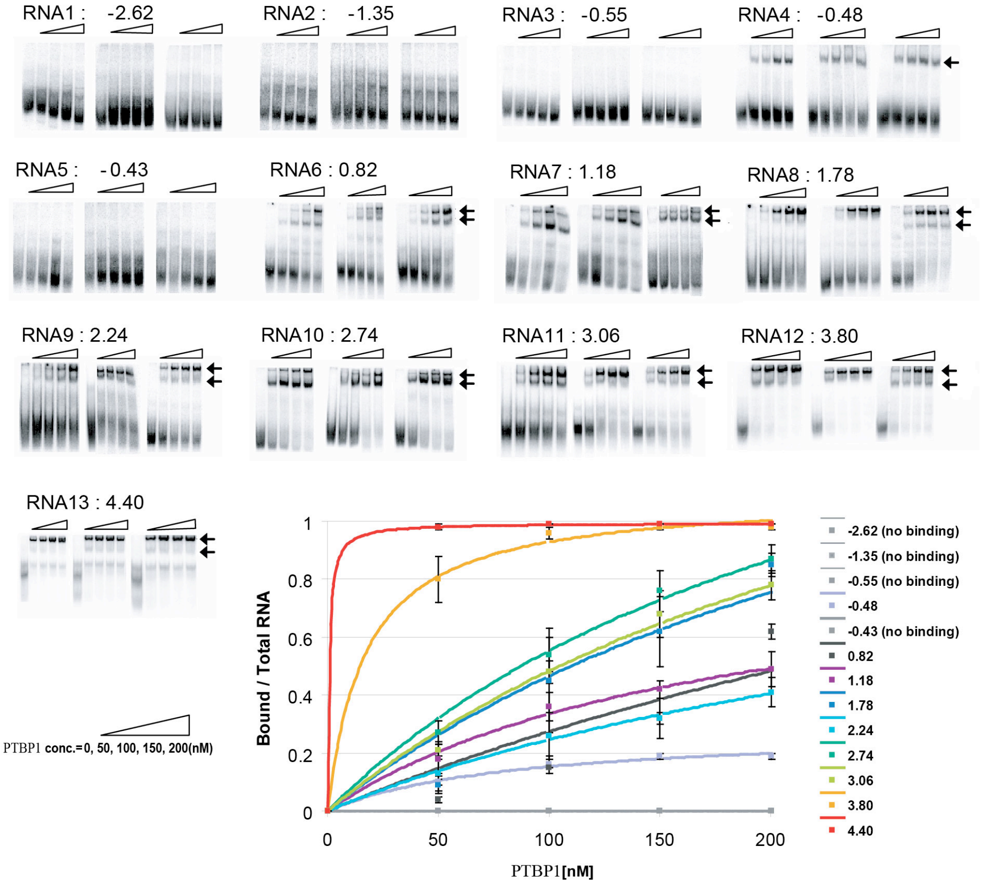
Lrrfip1	leucine rich repeat (in FLII) interacting protein 1	chr1:92990137-92990214(+)	0.82
Srsf11	serine/arginine-rich splicing factor 11	chr3:157703405-157703586(+)	0.82
Tmem209	transmembrane protein 209	chr6:30441087-30441184(-)	0.82



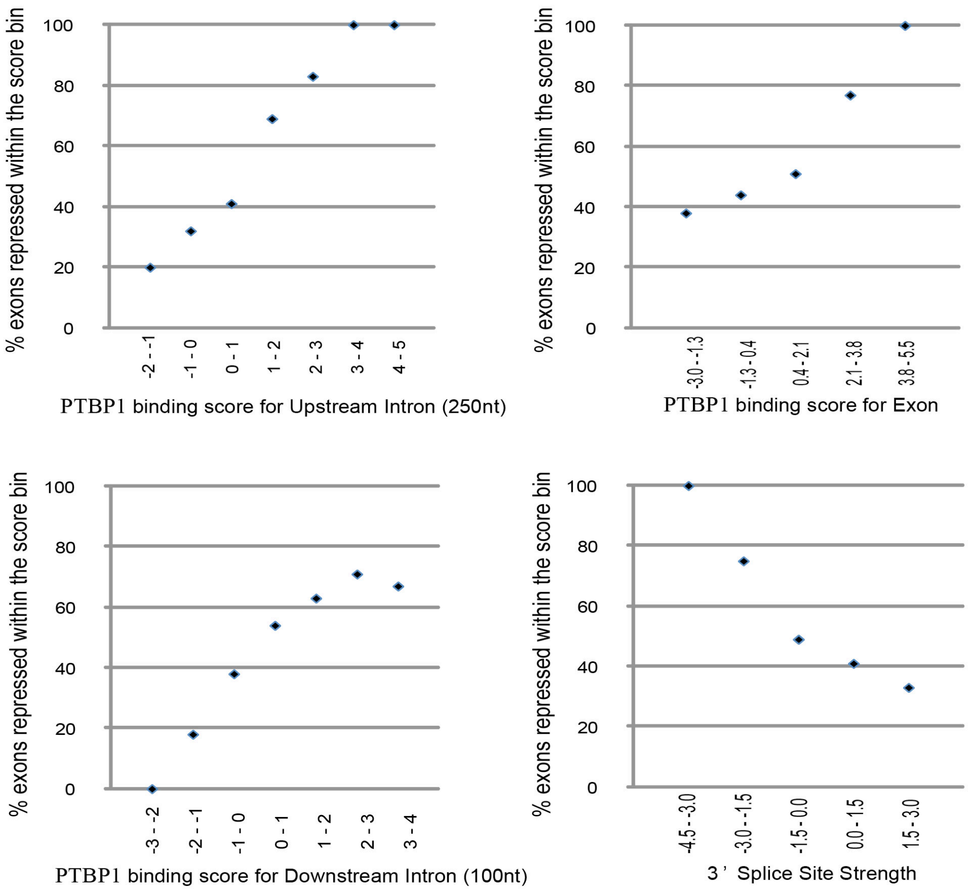
Average of Raw Scores = - 6.95
Standard deviation of Raw Scores = 9.63

$$Z\ Score\ = \frac{Raw\ Score\ \text{-}\ Average\ (\ Raw\ Scores\ )}{St\ dev.\ (\ Raw\ Scores\ )}$$

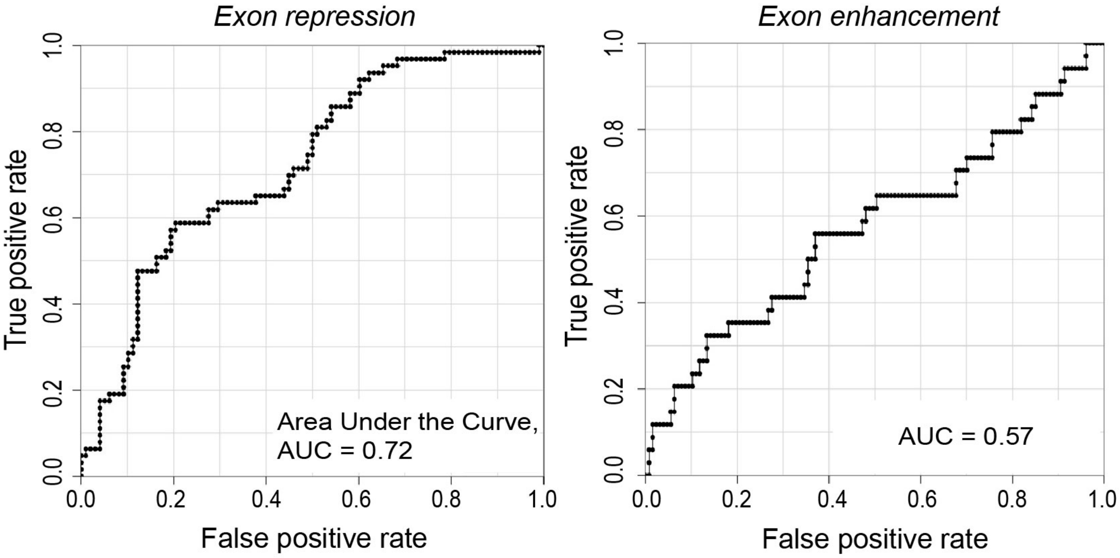
Supplementary Figure 1. A. PTBP1 binding model scores and validation assay results. A. Summary statistics and distribution of raw and normalized PTBP1 binding scores for 10,000 random sequences.



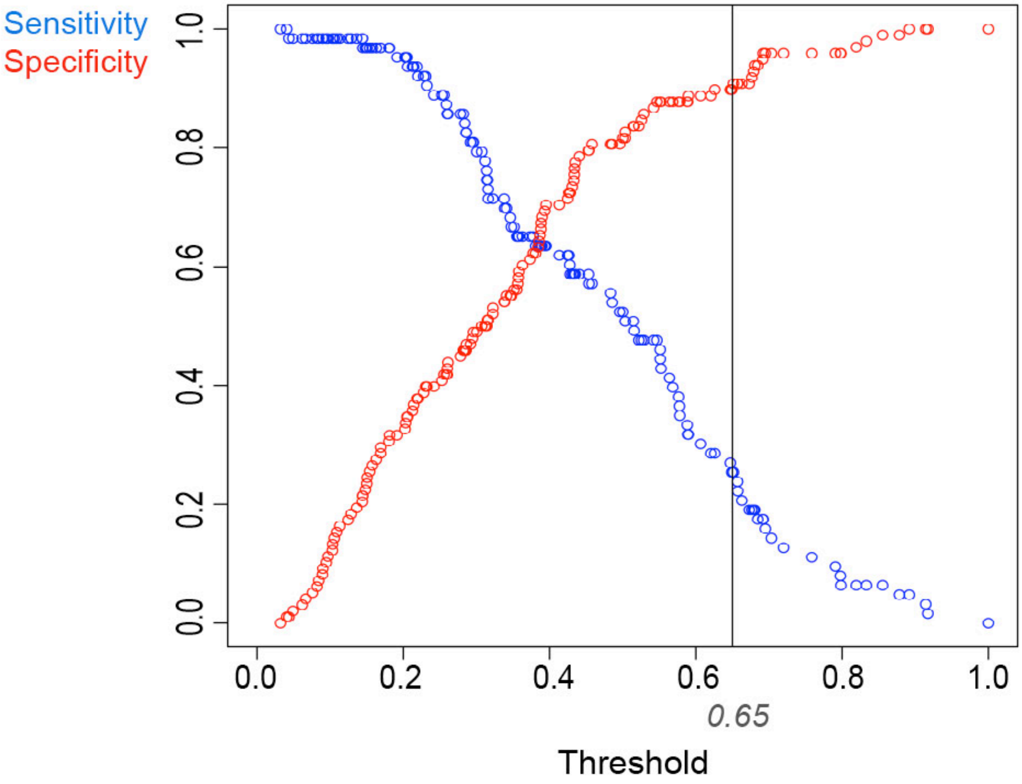
Supplementary Figure 1. B. Electrophoretic mobility shift assay results of thirteen RNAs with various PTBP1 binding scores. RNAs were in vitro transcribed, incubated with purified PTBP1 (0 to 200nM), and run out on native gels. Arrows indicate RNA-protein complexes. The graph shows the quantified fraction of PTBP1-bound RNA.



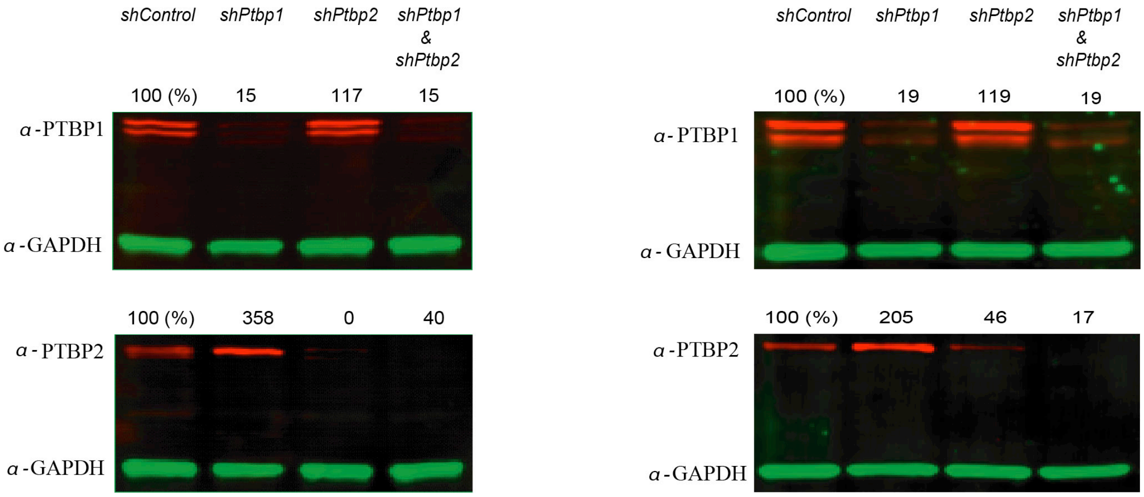
Supplementary Figure 2. Plots of the percentage of PTBP1 repressed exons in each PTBP1 binding score bin. For 68 PTBP1-repressed and 69 PTBP1 non-regulated exons, we calculated scores for sequence features and plotted the fraction of PTBP1-repressed exons in each score bin. Plots visualize relationships between exon repression by PTBP1 and individual splicing feature.



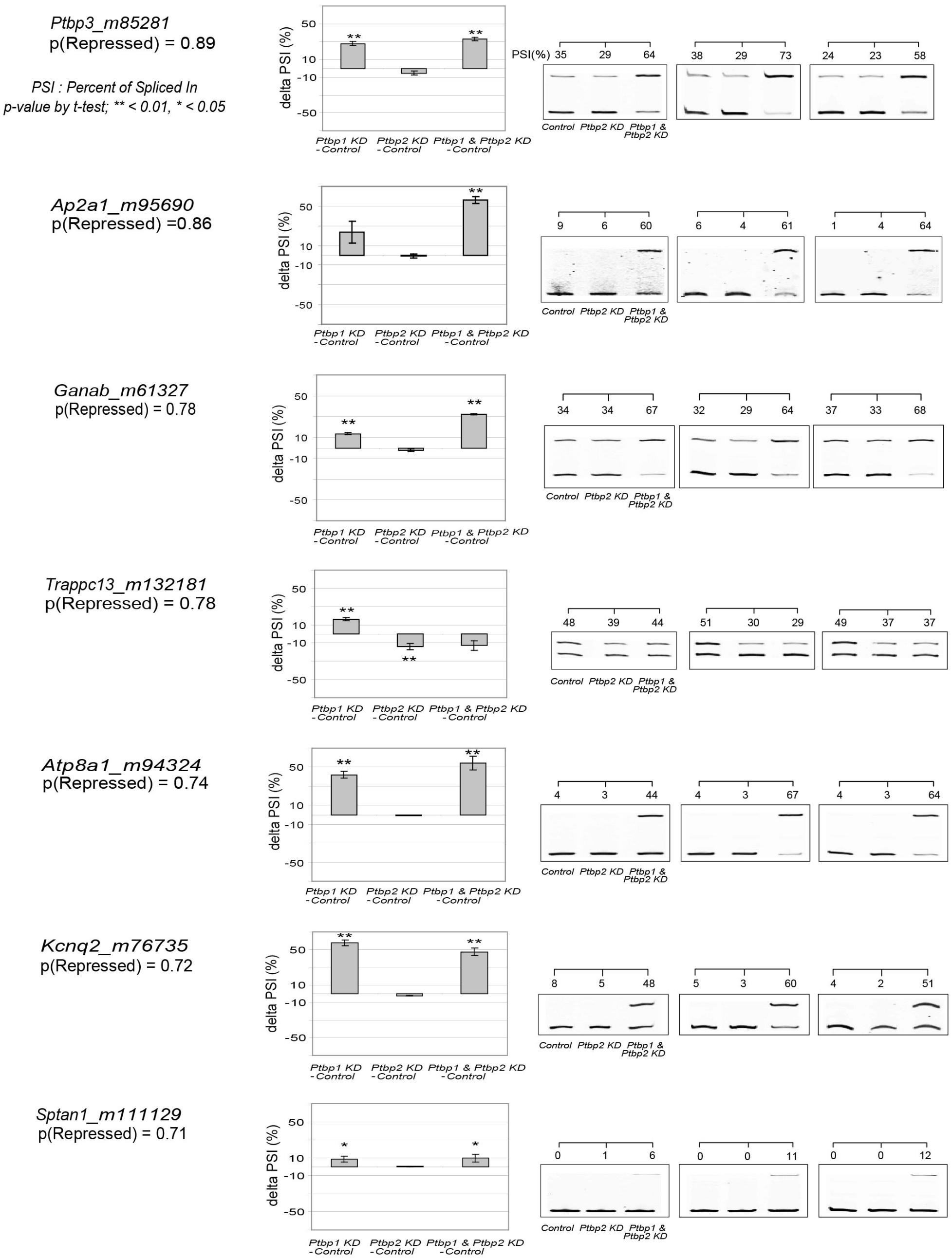
Supplementary Figure 3. A. Performance accessment of PTBP1-dependent splicing models. Receiver Operating Characteristic Curves in a leave-one-out cross validation for each logit, exon repression (left) and enhancement (right).



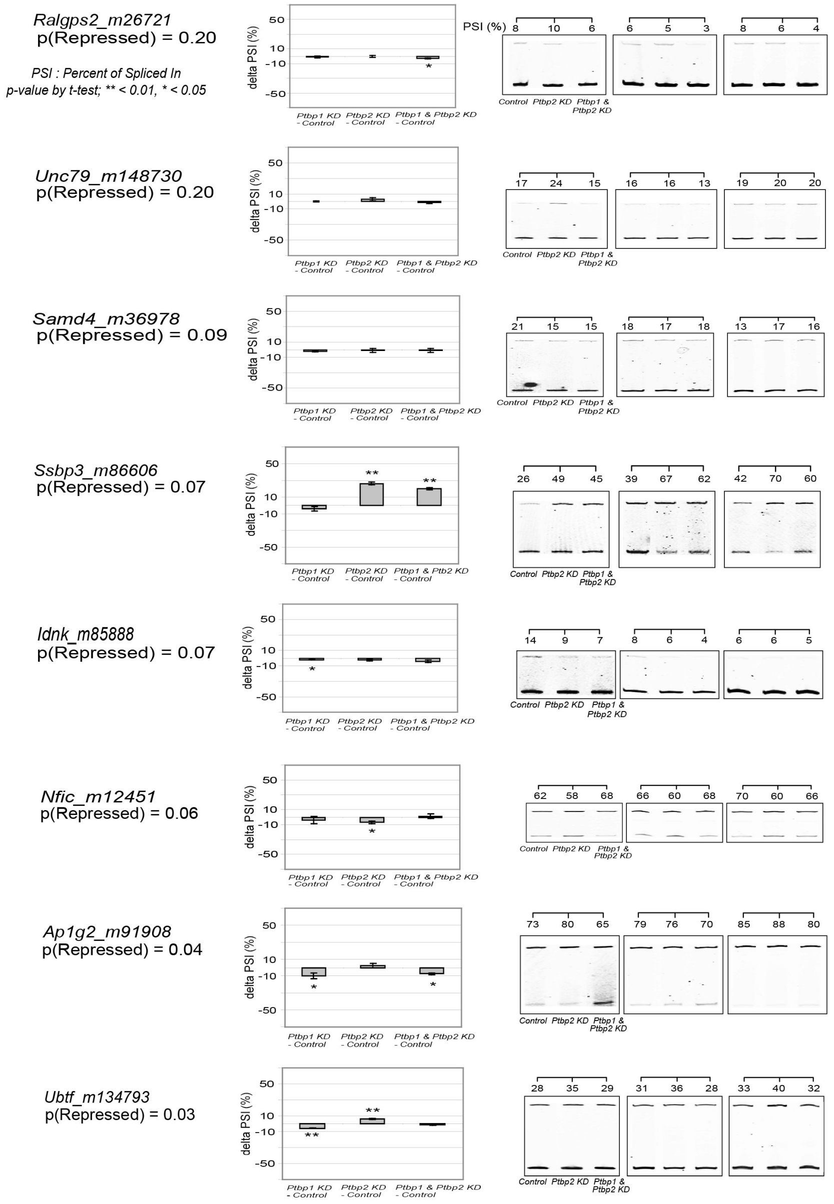
Supplementary Figure 3. B. Sensitivity and specificity were plotted across the whole threshold range. Sensitivity is defined as the percent of true repressed exons that are correctly predicted as repressed at the corresponding threshold. Specificity is defined as the percent of actual non-repressed exons that are correctly predicted as non-repressed exon at the corresponding threshold.



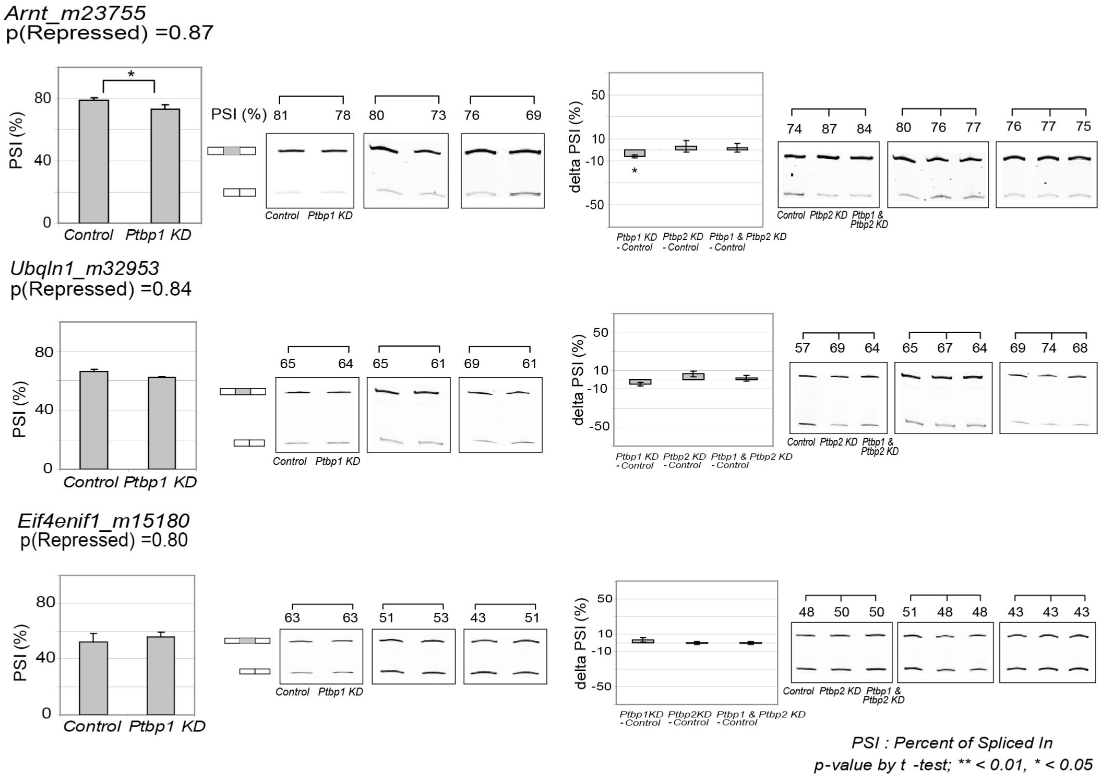
Supplementary Figure 4. Depletion of PTB proteins in Ptb gene knockdown experiments. Western blots show expression levels of PTBP1 and PTBP2 proteins in knockdown experiments. Numbers indicate quantified protein expression levels using fluorescence-labeled secondary antibody.



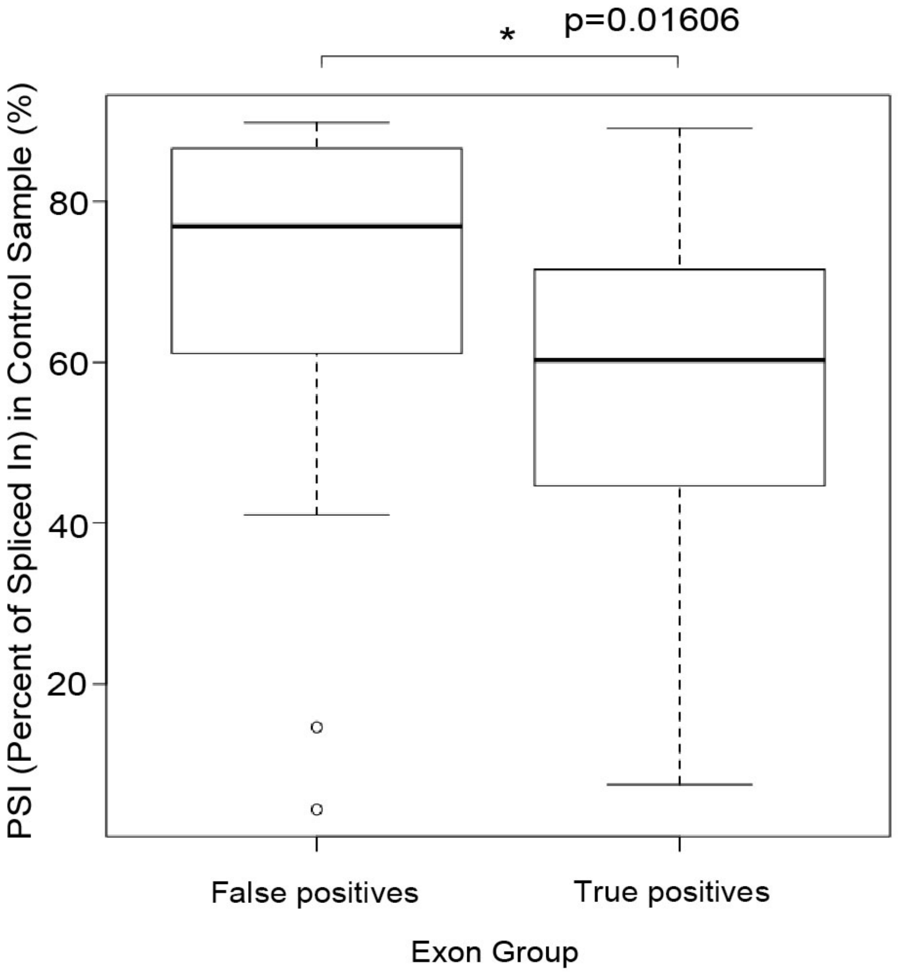
Supplementary Figure 5. A. PTBP1-repressed exon candidates with probability greater than 0.65 were tested further by RT-PCRs following Ptbp2 or Ptbp1 & Ptbp2 double knockdown. The graph shows averages ± SE of delta PSI (Percent of exon Spliced In) in biological triplicates. P-values were calculated from paired one-tailed t-tests with PSI values in control samples.



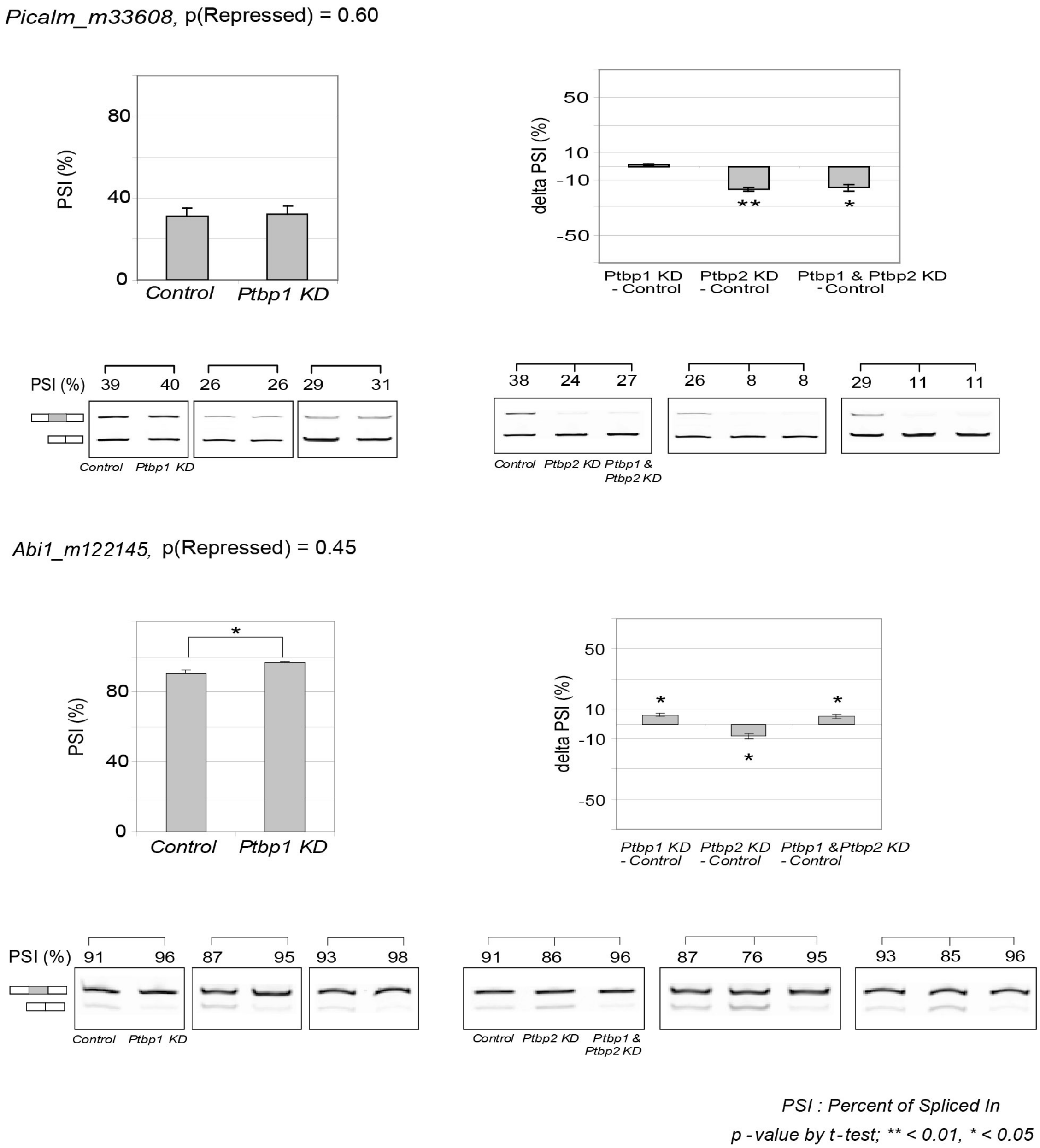
Supplementary Figure 5. B. Exons with low probabilities for PTBP1 repression (≤0.2) were tested further by RT-PCRs following Ptbp2 knockdown and Ptbp1 & Ptbp2 double knock down. The data shown are averages ± standard error of delta PSI (Percent of Spliced In) in biological triplicates. P-values were calculated from paired one-tailed t-tests with PSI values in control samples.



Supplementary Figure 6. A. Characteristics of false positive exons and intermediate scored exons. False positive exons did not show derepression in Ptb gene knockdown experiments. Exon inclusion level changes of three false positives exons in Ptbp1 knockdown, Ptbp2 knockdown, and Ptbp1 & Ptbp2 double knockdown experiments. P-values were calculated from biological triplicates using paired one-tailed t-tests.



Supplementary Figure 6. B. From RNA-seq data, we took twenty-nine false positive exons with high probability to be repressed (>0.55), but which were not de-repressed (delta PSI < 5%) upon Ptbp1 knockdown. Their initial exon inclusion levels were compared with those from thirty-six true positive exons. Statistical significance was assessed by one-tailed t-test.



Supplementary Figure 6. C. Two exons with intermediate scores showed complex responses to Ptb gene knockdown experiments. Exon inclusion level changes of two exons in Ptbp1, Ptbp2, and Ptbp1 & Ptbp2 knockdown experiments were shown. P-values were calculated from biological triplicates using paired one-tailed t-tests.

Supplementary Table 1. Trained PTBP1 splicing regulation model. The table presents a summary of the multinomial logistic regression model for PTBP1 splicing regulation. Estimated coefficients and their statistics are presented.

Logit	Variable	Coeff. (**β**)	Std. Err.	Odd Ratio	t	P>|t|	Signif. Code
$g_{repressed}(x) = \ln\left[\frac{p(repressed \mid x)}{p(Non-regulated \mid x)}\right]$	PTBP1 Binding Score for upstream intron(250nt)	0.7818	0.2363	2.1853	3.3079	0.0009	***
	PTBP1 Binding Score for exon	0.2840	0.1690	1.3285	1.6811	0.0928	.
	PTBP1 Binding Score for downstream intron(100nt)	0.4709	0.2008	1.6015	2.3454	0.0190	*
	3' Splice site strength	-0.5104	0.1874	0.6003	-2.7238	0.0065	**
	Constant	-0.9713	0.2818		-3.4464	0.0006	***
$g_{enhanced}(x) = \ln\left[\frac{p(enhanced \mid x)}{p(Non-regulated \mid x)}\right]$	PTBP1 Binding Score for upstream intron(250nt)	0.0429	0.2606	1.0439	0.1647	0.8692	
	PTBP1 Binding Score for exon	-0.1954	0.2046	0.8225	-0.9552	0.3395	
	PTBP1 Binding Score for downstream intron(100nt)	0.4864	0.2095	1.6264	2.3213	0.0203	*
	3' Splice site strength	-0.4840	0.2048	0.6163	-2.3632	0.0181	*
	Constant	-0.9816	0.2720		-3.6089	0.0003	***

‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Supplementary Table 2. Summary of PTBP1-dependent splicing model validation using RT-PCRs. The table shows validation results, sequence features, and predicted PTBP1 repression probabilities of eighteen exons. True positive exons are novel PTBP1-dependent exons identified by the

model.

Category	Prediction	RT-PCR Result	Gene Name	Gene Description	mm9 Coodinates	PTB Binding Scores			3'	p(Repressed)
						Upstream Intron (250nt)	Exon	Downstream Intron (100nt)	Splice Site Strength	
True positive	Repressed	Repressed	*Ptbp3*	Polypyrimidine tract binding protein 3	chr4:59559021-59559054 (-)	3.80	0.57	1.66	0.73	0.89
True positive	Repressed	Repressed	*Ap2a1*	adaptor-related protein complex 2, alpha 1 subunit	chr7:52158832-52158897 (-)	3.35	-0.12	-0.89	-2.69	0.86
True positive	Repressed	Repressed	*Ganab*	alpha glucosidase 2 alpha neutral subunit	chr19:8982341-8982406 (+)	2.38	-0.02	0.75	-1.74	0.78
True positive	Repressed	Repressed	*Trappc13*	trafficking protein particle complex 13	chr13:104942245-104942262 (-)	2.23	1.02	1.56	0.01	0.78
True positive	Repressed	Repressed	*Atp8a1*	ATPase, aminophospholipid transporter (APLT)	chr5:68147889-68147933 (-)	2.02	-0.30	1.41	-2.13	0.74
True positive	Repressed	Repressed	*Kcnq2*	potassium voltage-gated channel subfamily KQT member 2	chr2:180838347-180838376 (-)	1.74	-0.17	-0.76	-4.15	0.72
True positive	Repressed	Repressed	*Sptan1*	spectrin alpha, non-erythrocytic 1	chr2:29884164-29884226 (+)	1.42	0.99	1.78	-0.73	0.71
True negative	Non-repressed	Non-repressed	*Ralgps2*	Ral GEF with PH domain and SH3 binding motif 2	chr1:158751518-158751595 (-)	-0.49	-1.70	2.08	-2.09	0.2
True negative	Non-repressed	Non-repressed	*Unc79*	unc-79 homolog (C. elegans)	chr12:104363819-104363935 (+)	-0.91	-0.58	2.86	-0.47	0.2
True negative	Non-repressed	Non-repressed	*Samd4*	sterile alpha motif domain containing 4	chr14:47672525-47672788 (+)	-0.17	0.43	-1.18	1.34	0.09
True negative	Non-repressed	Non-repressed	*Ssbp3*	single-stranded DNA binding protein 3	chr4:106699822-106699902 (+)	0.08	-1.13	-2.72	-0.21	0.07
True negative	Non-repressed	Non-repressed	*Idnk*	idnK gluconokinase homolog (E. coli)	chr13:58264199-58264242 (+)	-0.38	0.36	-1.65	0.99	0.07
True negative	Non-repressed	Non-repressed	*Nfic*	nuclear factor I/C	chr10:80866097-80866173 (-)	-1.00	-0.08	-2.64	-0.73	0.06
True negative	Non-repressed	Non-repressed	*Ap1g2*	AP-1 complex subunit gamma-like 2	chr14:55723419-55723495 (-)	-1.59	-0.59	-3.18	-1.41	0.04
True negative	Non-repressed	Non-repressed	*Ubtf*	upstream binding transcription factor, RNA polymerase I	chr11:102172240-102172350 (-)	-0.79	-1.01	-0.96	1.85	0.03
False positive	Repressed	Non-repressed	*Arnt*	aryl hydrocarbon receptor nuclear translocator	chr3:95270715-95270759 (+)	3.48	-0.36	2.53	0.09	0.87
False positive	Repressed	Non-repressed	*Ubqln1*	ubiquilin 1	chr13:58282183-58282266 (-)	2.88	0.98	-0.06	-1.17	0.84
False positive	Repressed	Non-repressed	*Eif4enif1*	eukaryotic translation initiation factor 4E transporter	chr11:3133991-3134062 (+)	3.45	0.28	0.42	0.60	0.8

Supplementary Table 3. Enriched gene ontology categories for novel PTBP1-repressed exons. The table lists ontology entries enriched in genes with predicted PTBP1-repressed exons (probability score of exon repression > 0.65). Whole mouse internal exons were used as the control set, and p-value cut off was 0.05. Gene ontology analysis was performed using the GOTM web server.

Gene ontology category	Description	Gene ontology ID	number of reference genes in the category	number of genes in the gene set and also in the category	expected number in the category	Ratio of enrichment	p value adjusted by the multiple test adjustment
biological process	calcium ion transport	0006816	116	39	18.88	2.07	0.0008
	chromosome organization	0051276	282	73	45.89	1.59	0.0025
	establishment of organelle localization	0051656	25	13	4.07	3.20	0.0042
	regulation of microtubule cytoskeleton organization	0070507	19	11	3.09	3.56	0.0043
	synaptic transmission, glutamatergic	0035249	15	9	2.44	3.69	0.0140
	ubiquitin dependent protein catabolic process	0006511	105	32	17.09	1.87	0.0140
	mitotic metaphase/anaphase transition	0007091	16	9	2.60	3.46	0.0185
	vesicle organization	0016050	26	12	4.23	2.84	0.0185
	establishment of protein localization	0045184	591	127	96.18	1.32	0.0221
	intracellular transport	0046907	357	82	58.10	1.41	0.0262
cellular component	presynaptic membrane	0042734	15	11	2.42	4.55	0.0002
	terminal button	0043195	6	5	0.97	5.17	0.0116
	nuclear pore	0005643	37	15	5.97	2.51	0.0116
	apical plasma membrane	0016324	67	22	10.81	2.04	0.0116
	condensed chromosome	0000793	52	19	8.39	2.26	0.0116
	postsynaptic membrane	0045211	109	32	17.58	1.82	0.0116
	axon part	0033267	17	9	2.74	3.28	0.0116
	T tubule	0030315	11	7	1.77	3.94	0.0116
	microtubule cytoskeleton	0015630	291	67	46.95	1.43	0.0200
	extracellular matrix part	0044420	80	24	12.91	1.86	0.0223
	sarcoplasmic reticulum	0016529	30	12	4.84	2.48	0.0229
	myosin filament	0032982	13	7	2.10	3.34	0.0244
	COPI vesicle coat	0030126	10	6	1.61	3.72	0.0244
	neuron projection membrane	0032589	5	4	0.81	4.96	0.0272
	kinetochore	0000776	17	8	2.74	2.92	0.0272
	leading edge membrane	0031256	5	4	0.81	4.96	0.0272
	postsynaptic density	0014069	14	7	2.26	3.10	0.0311
molecular function	calcium channel activity	0005262	61	29	10.01	2.90	0.0000
	ATP binding	0005524	1237	291	203.05	1.43	0.0000
	active transmembrane transporter activity	0022804	262	68	43.01	1.58	0.0012
	ubiquitin thiolesterase activity	0004221	53	21	8.70	2.41	0.0012
	motor activity	0003774	109	34	17.89	1.90	0.0021
	helicase activity	0004386	105	32	17.24	1.86	0.0042
	cytoskeletal protein binding	0008092	327	79	53.68	1.47	0.0042
	transcription coactivator activity	0003713	59	21	9.68	2.17	0.0059
	peptidase activity, acting on L amino acid peptides	0070011	481	107	78.96	1.36	0.0072
	ligase activity	0016874	310	74	50.89	1.45	0.0072
	ATPase activity, coupled	0042623	167	45	27.41	1.64	0.0072
	calmodulin binding	0005516	105	31	17.24	1.80	0.0088
	GTPase binding	0051020	54	19	8.86	2.14	0.0102

CHAPTER 3

Redundancy and Divergence of Targets for Splicing Factor Paralogs, PTBP1 and PTBP2.

INTRODUCTION

Gene duplication is a major evolutionary force that proliferate and diverse gene pools in population genetics (Hughes, 1994; Zhang, 2003). Sequence duplication is generated by transposition of mobile elements, which can be fixed in genome by selection (Hughes, 1994). Duplicated genes often survive when confer a benefit to organism. They provide specialized or novel function not traced back to their ancestors. Alternatively, they may serve extra copies of genes when the gene products are in high demand (Zhang, 2003).

Paralogous genes are evolutionarily related genes within a species by gene duplication events (Hughes, 1994). They can play either redundant or divergent roles in a cell. Paralogous genes of splicing factors and transcription factors have both binding and regulation targets (Singh and Hannenhalli, 2008). They can follow one of several evolutionary scenarios (Singh and Hannenhalli, 2008; Teichmann and Babu, 2004). The paralogous genes may be fully redundant and recognize the same ribonucleotide targets and gene targets. On the other hand, they could exhibit an entirely different set of binding and regulation targets. In some cases, paralogous genes show a mixed behavior, where they are partially redundant and only a few functional parts (i.e. domains) are conserved. An example of this would be paralogous genes that have the identical binding targets, but they regulate different genes due to different tissue-specific activity or partner proteins.

Splicing is the primary mechanism for regulation of gene expression (Chen and Manley, 2009). A precursor messenger RNA (pre-mRNA) transcribed from a protein-coding gene contains non-coding segments called introns. Splicing removes introns and ligates coding segments, exons, during pre-mRNA maturation. In mammalian cells, splicing is regulated extensively by several hundreds of splicing factors, allowing generation of frequent alternative splicing events (Pan et al., 2008; Wang et al., 2008). Splicing factors often bind to their binding sites near the target exon using their RNA binding domain (RBD) or RNA recognition motif (RRM) and regulate whether the alternative exon will be included or excluded in the final mRNA (Black, 2003).

There are many splicing factor paralogs and it raises an important question – which evolutionary scenario would describe the case of splicing factor paralogs? The studies that address this question will enlighten evolutionary roles of gene duplication events in splicing networks. Furthermore, it will allow us to understand how splicing factor paralogs help cells to achieve sophisticated splicing regulations.

Polypyrimidine tract binding (PTB) proteins are strong splicing factors. Four PTB paralogous genes have been reported in mammals including PTBP1 (hnRNPI), PTBP2, ROD1 (Regulator of Differentiation 1 or PTBP3), and smPTB (Keppetipola et al., 2012). PTB paralogous have different expression patterns (Boutz et al., 2007b). The PTBP1 is broadly and abundantly expressed in most cells except differentiating neurons and muscle cells. On the contrary, three paralogous genes (PTBP2, ROD2, and smPTB) exhibit more restricted expressions (Keppetipola et al., 2012). PTBP2 is expressed in neuronal cell types, whereas ROD1 and smPTB in hematopoietic cells and smooth muscle, respectively (Keppetipola et al., 2012).

PTBP1 and its neuronal paralogs, PTBP2, share over 70% of the amino acid sequence identity (Keppetipola et al., 2012). Additionally, only two amino acids are different within their RNA interface region (Keppetipola et al., 2012). It has been reported that PTBP1 and PTBP2 have similar affinity to a pyrimidine rich sequence within the upstream intron of the c-src-N1 exon (Keppetipola et al., unpublished). It seems PTBP1 and PTBP2 have similar RNA affinity at least for pyrimidine rich sequences (Keppetipola et al., unpublished).

Despite the similarity of these splicing factors at the protein level, PTBP1 and PTBP2 are expressed mutually exclusively in most cell lines and tissues (Boutz et al., 2007b). Expression of PTBP2 is tightly down regulated under the presence of PTBP1 at both posttranscriptional and translational levels (Boutz et al., 2007b). Previous studies revealed that PTBP1 enhances nonsense-mediated decay (NMD) of PTBP2 transcripts by repressing inclusion of an alternative exon (Boutz et al., 2007b). Also, PTBP2 transcripts that are not NMD subjects show inefficiently translation upon the presence of PTBP1. This tight regulation of PTBP1 to PTBP2 suggests that two paralogous genes might be evolved to ensure that each protein regulates different splicing networks.

Recently, CLIP-seq (crosslinking and immunoprecipitation followed by sequencing) technology was developed and enabled researchers to identify millions of in vivo targets of RNA binding proteins (Licatalosi et al., 2012; Ule et al., 2005). Using PTBP1 CLIP-seq data, previously we were able to derive a predictive binding model for PTBP1. In this study, we expanded the model to PTBP2 using recently available PTBP2 CLIP-seq data (Licatalosi et al., 2012). Two PTBP1 and PTBP2 binding models enable us to compare their RNA binding properties and targets systemically.

With an advance of sequencing technology, RNA-sequencing enables researchers to assay the transcriptome in cells (Mortazavi et al., 2008; Wang et al., 2009). Splicing factor dependent alternative exons have been identified by measuring exon inclusion levels while manipulating expression levels of splicing factors. In this study, we down-regulated individual PTB protein (PTBP1 or PTBP2) or both PTBP proteins in mouse N2A (neuroblastoma) cells (Boutz et al., 2007b) and performed RNA-seq. As a result, we were able to identify and characterize exon targets of PTB protein paralogs. We expect our results will reveal an evolutionary role of gene duplication events in splicing network.

METHODS AND RESULTS

Binding codes for PTBP1 and PTBP2 are redundant.

To assess binding properties of PTB proteins, we took a set of RNA sequences bound to PTB proteins from previously published CLIP data (Licatalosi et al., 2012; Xue et al., 2009). Then, we trained hidden markov models for the binding targets of PTBP1 and PTBP2. Overall, derived models were very similar suggesting that PTBP1 and PTBP2 have similar RNA binding properties (Figure 1). Indeed, average dissimilarity between two models based on 10,000 sequences was only 0.12 (Rabiner, 1989). When we plotted the probability distribution for sixty-four possible triplets, we could observe that PTBP2 does not prefer guanosine-interrupted pyrimidines to some extent (Figure 1). In the future, it will be interesting to validate this observation and dissect subtle binding differences between PTBP1 and PTBP2.

We also applied both binding models to random and pre-mRNA sequences of internal mouse exons to assess overall distribution of binding scores for PTBP1 and PTBP2 (Figure 2). Again we observed that most sequences have similar binding scores although few sequences from actual pre-mRNA sequences prefer one protein to the other. These results support that PTBP1 and PTBP2 have redundant binding characteristics.

PTB profile centered clustering identifies PTB protein dependent exons.

Next, we sought to collect exon targets of PTB proteins to investigate splicing codes. We down-regulated single PTB protein or both PTB proteins in mouse neuroblastoma cells (N2A) and assayed RNA and PTB proteins from the same cells. We confirmed that PTBP2 is up regulated when PTBP1 is knocked down (Figure 3). Using RNA extracted from these cells, we generated 100bp paired-end RNA-seq libraries. From the sequencing results, we were able to estimate exon inclusion levels of 3,503 cassette exons (Figure 3) (Wu et al., 2011). We could observe that inclusion ratios of some exons are dependent on PTBP1 or PTBP2 protein levels. Some exons show complex responses to PTB proteins.

In order to identify splicing factor dependent exons, the most straightforward approach would be comparing exon inclusion levels between the control sample and splicing factor knock down samples. However, because of the regulation between PTBP1 and PTBP2, we could not merely treat those differential exons as one PTB protein dependent exons. In this study, we took protein expression profiles and calculated correlations with inclusion level profiles of alternative exons (Zhu et al., 2002). This splicing factor-centered way enables us to identify a PTB dependent exon in more reasonable way. When the correlation is close to 1, the exon inclusion level is concord with a PTB protein expression. In this case, a PTB protein might enhance splicing of this exon. On the contrary, an exon with the correlation close to -1 might be a repressed exon by a PTB protein. As a result, we observed that PTBP1 has quite different distributions of correlations with PTBP2 (Figure 4). PTBP1 have more exons that have correlation values close to either 1 or -1, which indicate that there are more exons regulated by PTBP1 than PTBP2. We observed that PTBP1 also has more repressed exons than enhanced ones. This confirms our hypothesis that PTBP1 acts more as a splicing repressor than an enhancer. Splicing targets of PTBP1 and PTBP2 are quite divergent as shown in an x-y plot of correlations for PTBP1 versus PTBP2 (Figure 4). It suggests that PTBP1 and PTBP2 have their own target exons although they bind similar RNA sequences. In the future, it will be extremely interesting to investigate whether this divergence of splicing targets are originated from switching in their expression or having different splicing codes. It will clarify whether PTB paralogs are survived because they are specialized or served novel function.

TABLES AND FIGURES

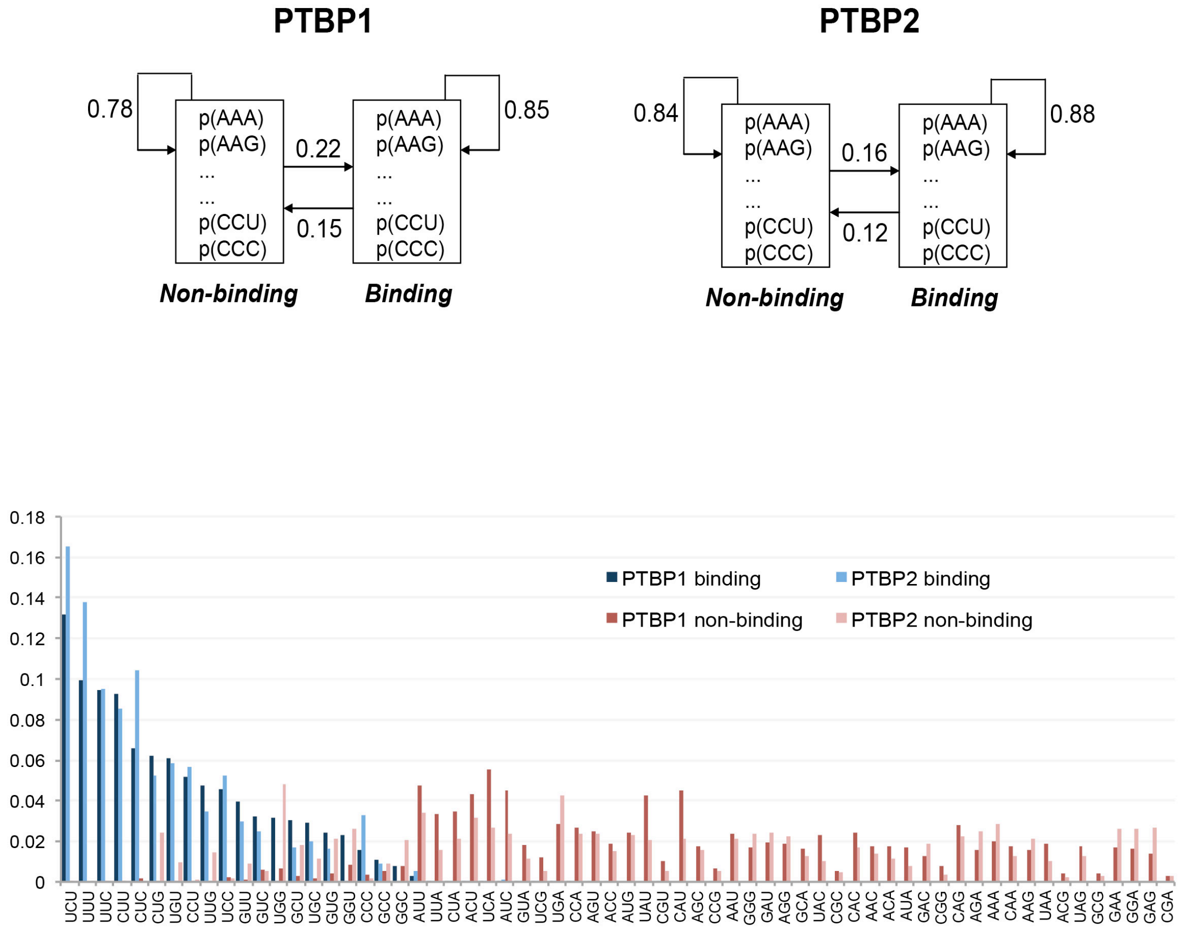
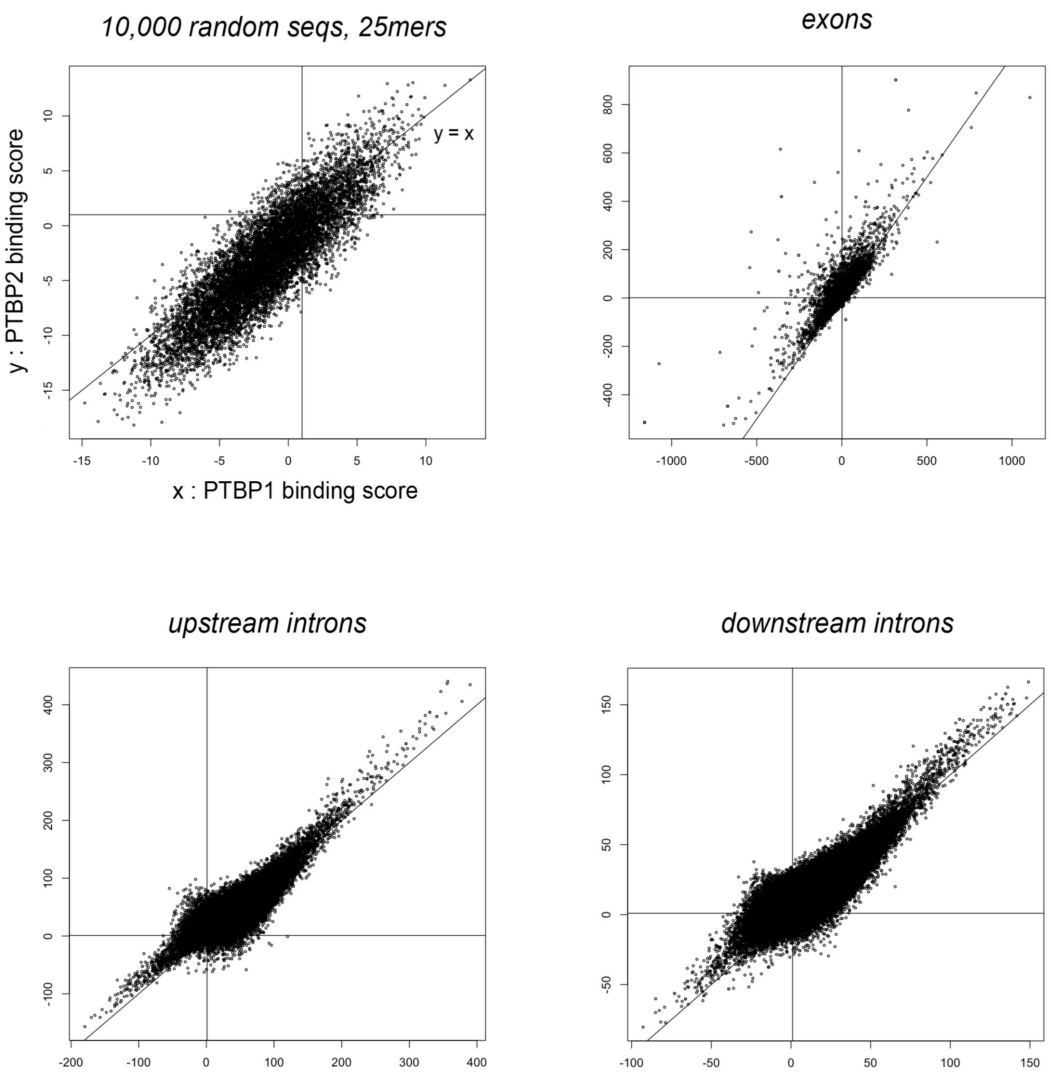


Figure 1. Binding models for PTBP1 and PTBP2 are similar suggesting redundant binding codes for PTB proteins. PTBP1 and PTBP2 binding models were trained by PTBP bound RNA sequences from published PTBP1-CLIP and PTBP2-CLIP data. The diagram presents the structure of the PTBP hidden markov model (HMM) and trained transition probabilities. Emission probabilities to observe a certain triplet in PTBP binding state and non-PTBP binding state are plotted for the two models.



10,000 random seqs, 25mers

exons

upstream introns

downstream introns

Figure 2. Most random and pre-mRNA sequences were scored similarly by both PTBP1 and PTBP2 model. Each dot represents an RNA sequence.

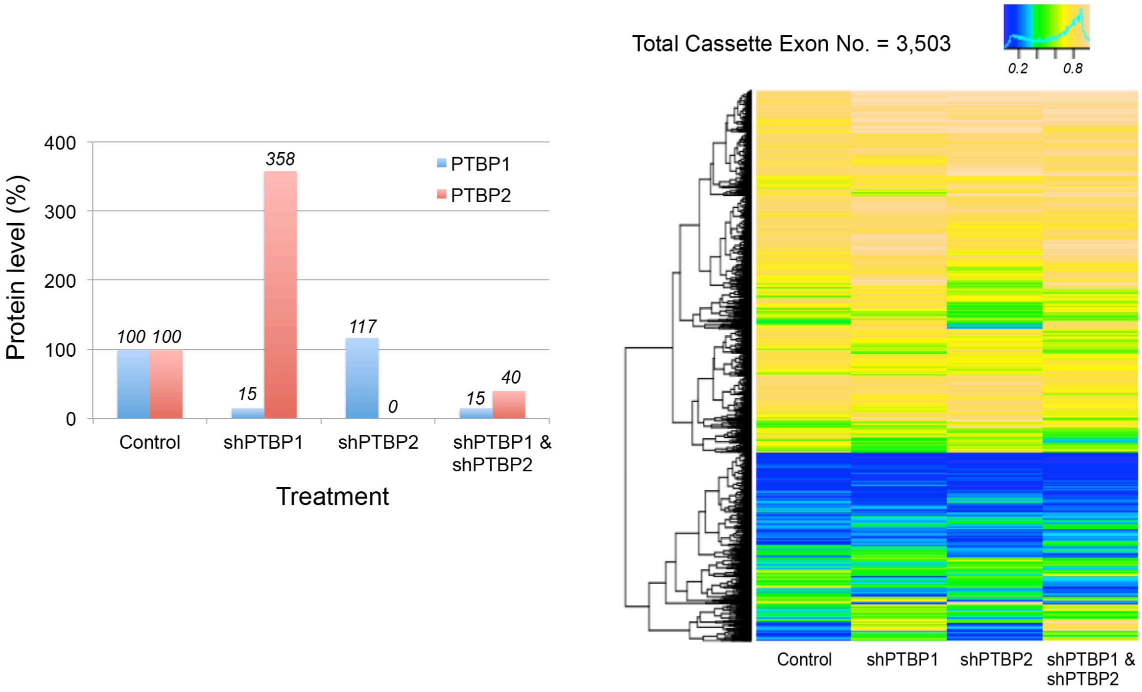


Figure 3. Protein expression and exon inclusion profiles from cells after PTB protein knockdown with short hairpin RNAs.

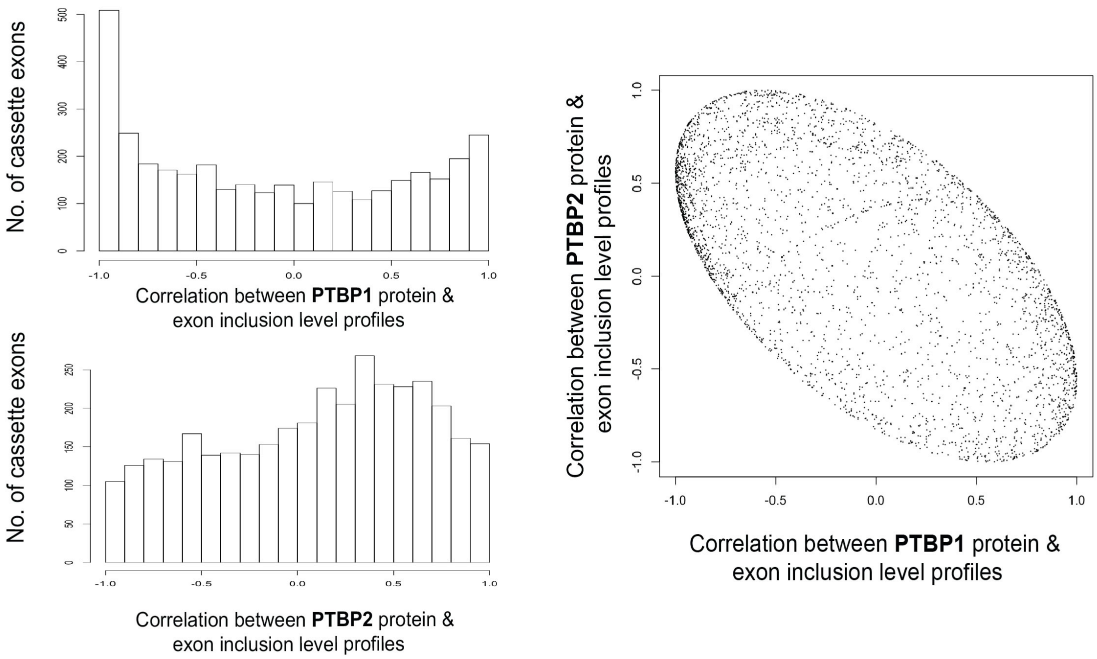


Figure 4. PTBP1 and PTBP2 are correlated differently with exon inclusion levels. Left panel displays histograms of correlations and right panel shows an x-y plot of correlations for PTBP1 and PTBP2. In this figure, each dot represents an alternative exon.

CHAPTER 4

Integrative analysis of in vivo binding and splicing targets of PTBP2 from mouse mutant studies

INTRODUCTION

Mouse mutants are useful to study function of genes in mammalian genetics and development (Hedrich Hans, 2004). A target gene in mouse can be inactivated by sequence driven mutagenesis. The sequence driven mutagenesis is achieved by designed gene targeting vectors containing a mutated gene sequence and a resistance marker gene, and optionally short motifs for recombination (Hedrich Hans, 2004). When the gene targeting vectors are introduced to cells, a mutated gene sequence and a resistance marker gene are inserted into the genome by homologous recombination (Hedrich Hans, 2004). The resistance marker gene enables selection of cells in which the homologous recombination event occurred. Later, the resistance marker gene can be removed by expression of recombinase (Hedrich Hans, 2004).

Mouse mutants are generated using embryonic stem (ES) cells and chimeric mice (Hedrich Hans, 2004). The gene targeting vectors are transfected into embryonic stem (ES) cells. Then, ES cells with the desired mutations are confirmed by testing their genomic DNA. Those ES cells with the correct mutation are used to generate chimeric mice (Hedrich Hans, 2004). By intercrossing chimeric mouse siblings, researchers are able to generate mouse mutants with a mutated target gene in both alleles (Hedrich Hans, 2004).

Complete knock out or inactivation of the target gene sometimes results complex phenotype or embryonic lethality (Hedrich Hans, 2004). In those cases, researchers can adopt a conditional gene targeting approach in order to more easily and specifically investigate the phenotype. The conditional gene targeting approach enables researchers to mutate the target gene in a particular cell type and/or time point. Therefore, researchers can investigate function of the gene in a desired cellular context (Hedrich Hans, 2004).

Conditional gene-targeting vectors have recombination sites that respond to a specific recombinase (Hedrich Hans, 2004). Thus, introducing a specific recombinase expression vector can inactivate the target gene in ES cells. Similarly, mutant mice can be bred with transgenic mice expressing the specific recombinase (Hedrich Hans, 2004). As a result, the target gene will be inactivated in desired cell types and/or time points. Recently, the Cre-loxP technique has been used to achieve conditional gene targeting (Hedrich Hans, 2004). The Cre-loxP technique uses Cre recombinase and its target sequence, loxP. Mice with the target gene flanked by loxP sites are crossed with transgenic mice expressing Cre under control of a specific promoter (Hedrich Hans, 2004). The promoter is inducible in specific cell types or at specific developmental time points; therefore the target gene can be designed to be conditionally inactivated.

Previously, two groups generated PTBP1 mouse mutants and studied the roles of PTBP1 in mammalian development (Shibayama et al., 2009; Suckale et al., 2011). Both studies reported embryonic lethality of homozygous PTBP1 knock out mice (PTBP1-/-), suggesting that PTBP1 is required for the embryonic development. In two studies PTBP1 null mice died at slightly different time points. One study observed that PTBP1 null mice died shortly after implantation (between after E3.5 and before E6.5) (Shibayama et al., 2009). In the other study, PTBP1-/- embryos were able to implant but died between after E7.5 and before E12 with growth defects (Suckale et al., 2011). The former study performed additional cell cycle analysis experiments in PTBP1-/- ES cells and observed a prolonged G2/M phase of the cell cycle (Shibayama et al., 2009). Although the two mutant mice lines showed differences in life span of PTBP1 null mice, both studies suggest that PTBP1 affects cell proliferation and may be critical for the embryonic development.

Neurons undergo dynamic changes in their splicing network during differentiation. Expression levels of the splicing factors PTBP1 and PTBP2 change dramatically during the neuronal differentiation (Qin et al., unpublished)(Boutz et al., 2007b). PTBP1 is highly expressed in non-neuronal cells, glial cells and neuronal progenitor cells (Keppetipola et al., 2012). Conversely, PTBP2 is first lowly expressed in the neuronal progenitor cells and becomes up regulated during differentiation into post-mitotic neurons. Later, in mature neurons, expression of PTBP2 is down regulated (Zheng et al., 2012). The expression pattern of these two proteins seems to distinguish the splicing network of three phases of neuronal differentiation: that of undifferentiated neuronal progenitor cells, differentiating neurons, and mature neurons (Zheng et al., 2012)(Qin et al., unpublished). PTBP1 is known to repress neuronal-specific exons in non-neuronal cells and neuronal progenitor cells. However, the role of PTBP2 in neurons has been not studied extensively.

To investigate role of the PTBP2 in neuronal development, PTBP2 null mice were generated and studied. Two independent studies reported that PTBP2 null mice died soon after birth, though they were not embryonic lethal as PTBP1 null mice were (Licatalosi et al., 2012) (Qin et al., unpublished). To investigate the neuronal function of the PTBP2 gene, two additional conditional PTBP2 null mice were generated (Qin et al., unpublished). One is the PTBP2-NesKO, which nullifies PTBP2 in central nervous system (CNS) neurons using a Nestin promoter driven Cre recombinase. The other is PTBP2-EmxKO, which nullifies PTBP2 in developing cortices using an Emx promoter driven Cre recombinase. The PTBP2-NesKO mice initiated breathing but died within one hour after birth, suggesting that PTBP2 in CNS neurons is required for postnatal survival. Interestingly, PTBP2-EmxKO mice survived longer, up to three weeks after birth. However, the mice eventually died at P21 with degenerated cortices, suggesting that developing neurons requires PTBP2 (Qin et al., unpublished).

PTBP2 is a splicing factor with multiple RNA binding domains (Keppetipola et al., 2012). Coupled with high-throuput assays, the PTBP2 null mice studies enabled us to investigate binding and splicing targets of PTBP2 in the mouse brain. Previously, PTBP2-CLIP was performed from E18.5 cortices of wild type littermates in the PTBP2 null mice study (Licatalosi et al., 2012). The CLIP experiment identified 19,767 CLIP significant clusters transcriptome-wide. RNA was isolated from neocortices of PTBP2 -/- mice and their wild littermates at E18.5, and differential splicing changes were profiled using microarrays.

RNA-SEQ technology has been developed to profile the transcriptome using a deep sequencing technology (Wang et al., 2009). Two PTBP2 conditional knock out mice, PTBP2-Nestin KO and PTBP2-Emx KO were subject to RNA-SEQ. In this study, I investigated splicing event changes in those PTBP2 conditional knock out mice (PTBP2-Nestin KO and PTBP2-Emx KO) from RNA-SEQ results. It revealed global characteristics of splicing regulation for PTBP2 in two transcriptome. I also performed an integrative analysis of in vivo binding and splicing targets of PTBP2 from knock out mice studies. I expect this integrated study will address how multiple PTBP2 null mice studies agree with each other. In mammalian cells, several hundreds of splicing factors are existed and alternative exons are often regulated by multiple factors. Alternative exons identified in independent studies multiple times will be strong candidates for direct targets of PTBP2. Those candidate exons can be studied further in depth to reveal the molecular mechanism of PTBP2 null mice phenotypes.

METHODS AND RESULTS

RNA was isolated from P0 brain of PTBP2-Nestin KO and P5 cortex of PTBP2-Emx KO (Qin *et al.,* unpublished). The RNA-SEQ libraries were generated and subject to 100bp-paired end sequencing. Hundreds of millions (M) of reads were mapped back to a mouse exon duo or trio database using the SpliceTrap tool (Wu et al., 2011). From RNA-SEQ libraries of PTBP2-NesKO and matched wild type mice, 129M and 132M reads were mapped, respectively. And from RNA-SEQ libraries of PTBP2-EmxKO and the wild type mice, 233M and 236M read were mapped.

Among total alternative splicing events from this initial database, most exons (80%) passed an expression filter (Wu et al., 2011), and inclusion levels of those exons were estimated. The total number of assayable alternative splicing events, 16,891 and 16,832, respectively, was very similar between two studies (Table1 and Table2). Though the numbers of mapped reads from the PTBP2-EmxKO mice study were about twice more than read numbers for the PTBP2-NesKO mice study, the number of assayable exons was quite similar. It is possible that 20% of the remaining exons are not expressed or are very lowly expressed in brains.

In this study, we defined an alternative exon as a PTBP2 regulated exon if its inclusion level changed over 10% upon loss of PTBP2. In total, 3,631(21%) and 2,421 (14%) PTBP2-dependent alternative events were identified from the PTBP2-NesKO mouse study and the PTBP2-EmxKO mouse study (Table1 and Table2). PTBP2-EmxKO mice show fewer alternative splicing changes than PTBP2-NesKO mice do. The RNA from PTBP2-EmxKO mice was isolated from cortices rather than from whole brains. Thus, it is possible that splicing events are less variable in the PTBP2-EmxKO mouse study because they are restricted to a specific brain region.

Interestingly, in both studies, numbers of repressed exons occupied about 60% of regulated exons (Table1 and Table2). It suggests PTBP2 acts as a splicing repressor more than as an enhancer. Splicing studies with the paralogous gene of PTBP2, PTBP1, also reported more PTBP1 repressed exons than PTBP1 enhanced exons. While both PTBP1 and PTBP2 can enhance exon inclusion, it seems there are more exons repressed by PTB proteins (Keppetipola et al., 2012).

Next, we investigated whether PTBP2-dependent alternative exons from these two studies overlap. We found that only 7% to 19% of alternative splicing events agreed with each other, suggesting that there are quite distinct alternative splicing events specified to each PTBP2 knock out mouse line (Figure 1). It is possible that different alternative splicing events are detected because RNA samples from the two mouse lines were isolated at different developmental time points (P0 vs. P5) and brain regions (whole brain vs. cortex).

We also compared the PSI (Percent Spliced In) changes upon knocking out PTBP2 in the two mouse lines. We subtracted the PSI of wild type mice from the PSI of PTBP2KO mice, creating in a delta PSI value. If an alternative exon was repressed by PTBP2, the delta PSI value was positive, while if the exon was enhanced by PTBP2 the delta PSI value was negative. From a plot of pairwise delta PSI values (Figure 2), moderate correlation was observed (Pearson correlation = 0.51). We found 465 alternative exons consistently regulated by PTBP2 in the same direction in both mouse lines. Some exons were regulated in opposite ways. These exons are likely indirect targets of PTBP2.

The PTBP2 dependent alternative exons identified in our study could be either direct or indirect targets of PTBP2. To address whether these exons had known binding sites for PTBP2 (Licatalosi et al., 2012), we searched PTBP2 CLIP clusters near the identified exons. For each exon, we searched regions between from 5’ neighbor exon to the 3’ neighbor exon of the PTBP2 dependent exon. We observed that more than half of simple cassette exons (53% from the Nestin-PTBP2KO and 56% from the Emx-PTBP2KO) had at least one known PTBP2 CLIP cluster within neighbor exon boundaries. The ratio was significantly higher than random chance (binomial test, p-values < 2.2e-16 in both lines). Thus, the exons identified in our study are enriched with direct targets of PTBP2. Interestingly, we could not observe statistically significant enrichment of CLIP clusters for other alternative splicing events such as intron retention, 5’ splice site, and 3’ splice site. It is possible that PTBP2 may bind near the target exon to regulate skipping or inclusion of the target exon entirely. Since the number of exons with other splicing events was smaller than that of simple cassette exons, we cannot rule out the possibility that they could not show statistically significant p-values.

We compared our RNA-SEQ results with previous splicing array results (Figure 3) from a previous study where RNA was isolated from cortices of PTBP2 KO mice at E18.5 (Licatalosi et al., 2012). This study also defined PTBP2 dependent exons, which showed more than 10% inclusion levels with loss of PTBP2. Among 393 regulated exons identified from this array, 173(44%) exons were assayed in our RNA-SEQ data, and 50% of PTBP2 repressed exons from the array were repressed in both PTBP2-NesKO and PTBP2-EmxKO mice. However, only 22% of PTBP2 enhanced exons from the array were also enhanced in both PTBP2-NesKO and PTBP2-EmxKO mice. Again, this suggests that enhanced exons are enriched with indirect targets and thus are more mouse-line specific than repressed exons are. When we visualized pairwise delta PSI values from the array and RNA-SEQ datasets (Figure 4 and Figure 5), we observed that some exons were not PTBP2-responsive in the RNA-SEQ experiments. Nevertheless, we rarely see that exons are regulated in different directions.

Finally, we were able to compile a list of exons with their delta PSI values in different PTBP2 knock out studies and presence of a PTBP2 CLIP cluster near the exon. We expect the resource is useful to identify the molecular mechanism of different phenotypes as well as to study direct PTBP2 target exons.

TABLES AND FIGURES

Table 1. Summary of PTBP2 regulated alternative splicing events identified from PTBP2-NesKO mouse study.

Alternative Splicing Event Type	No. Expressed exons	No. Regulated exons (%)	No. Repressed exons (%)	No. Enhanced exons (%)
Cassette exon	10,603	2967 (28%)	1,757 (17%)	1,210 (11%)
Alternative 3' splice site	3,276	359 (11%)	189 (6%)	170 (5%)
Alternative 5' splice site	2,136	205 (10%)	111 (5%)	94 (4%)
Intron retention	876	100 (11%)	61 (7%)	39 (4%)
Sum	16,891	3631 (21%)	2,118 (13%)	1,513 (9%)

Table 2. Summary of PTBP2 regulated alternative spicing events identified from PTBP2-EmxKO mouse study.

Alternative Splicing Event Type	No. Expressed exons	No. Regulated exons (%)	No. Repressed exons (%)	No. Enhanced exons (%)
Cassette exon	10,277	1,980 (19%)	1,220 (12%)	760 (7%)
Alternative 3' splice site	3,454	234 (7%)	124 (4%)	110 (3%)
Alternative 5' splice site	2,214	140 (6%)	84 (4%)	56 (3%)
Intron retention	887	67 (8%)	33 (4%)	34 (4%)
Sum	16,832	2,421 (14%)	1,461 (9%)	960 (6%)

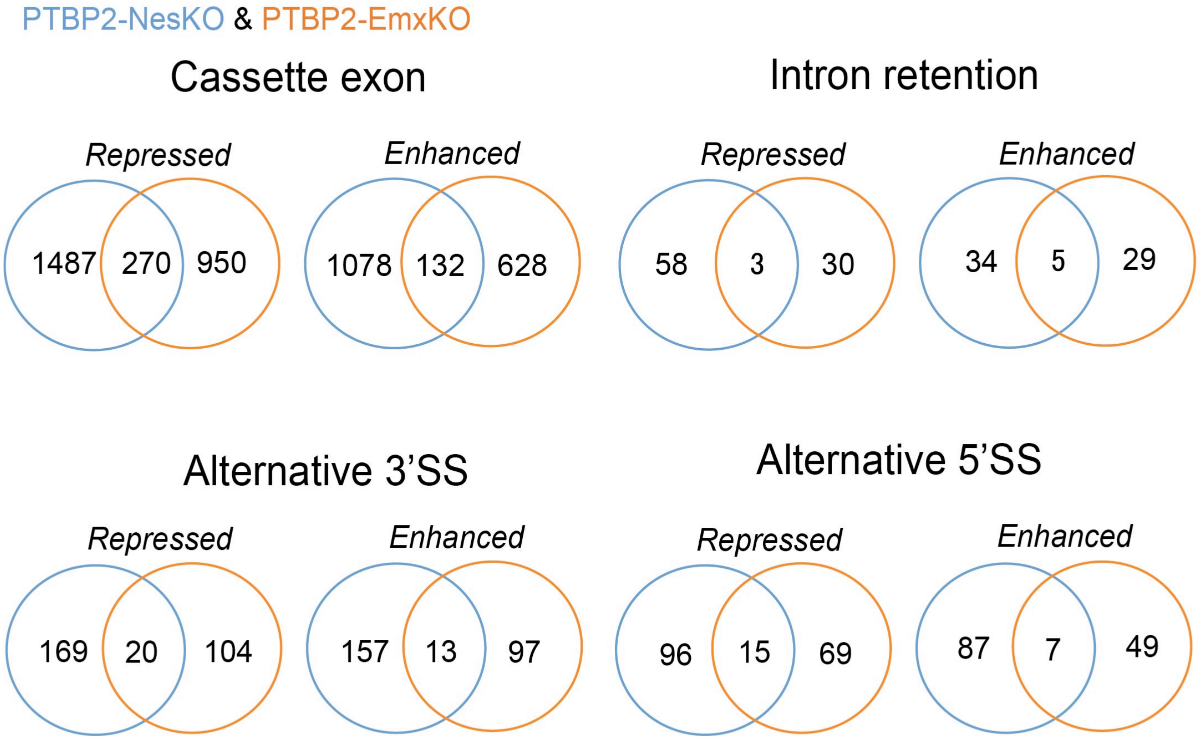


Figure 1. Overlaps of identified PTBP2 regulated exons from two conditional PTBP2 knock out mouse lines, PTBP2-NesKO and PTBP2-EmxKO.

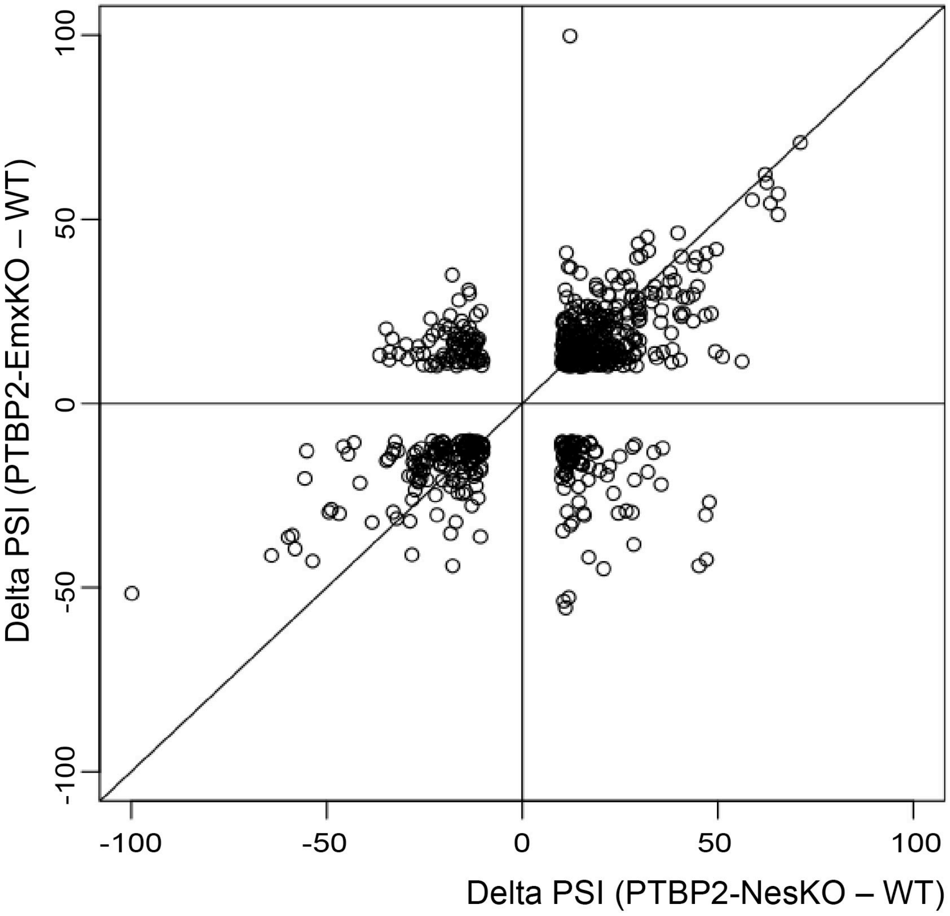


Figure 2. Pairwise comparison of the delta PSI (Percent Spliced In of an exon in wild type mice - Percent Spliced In of the exon in PTBP2 knock out mice) from PTBP2-NesKO and PTBP2-EmxKO mouse studies.

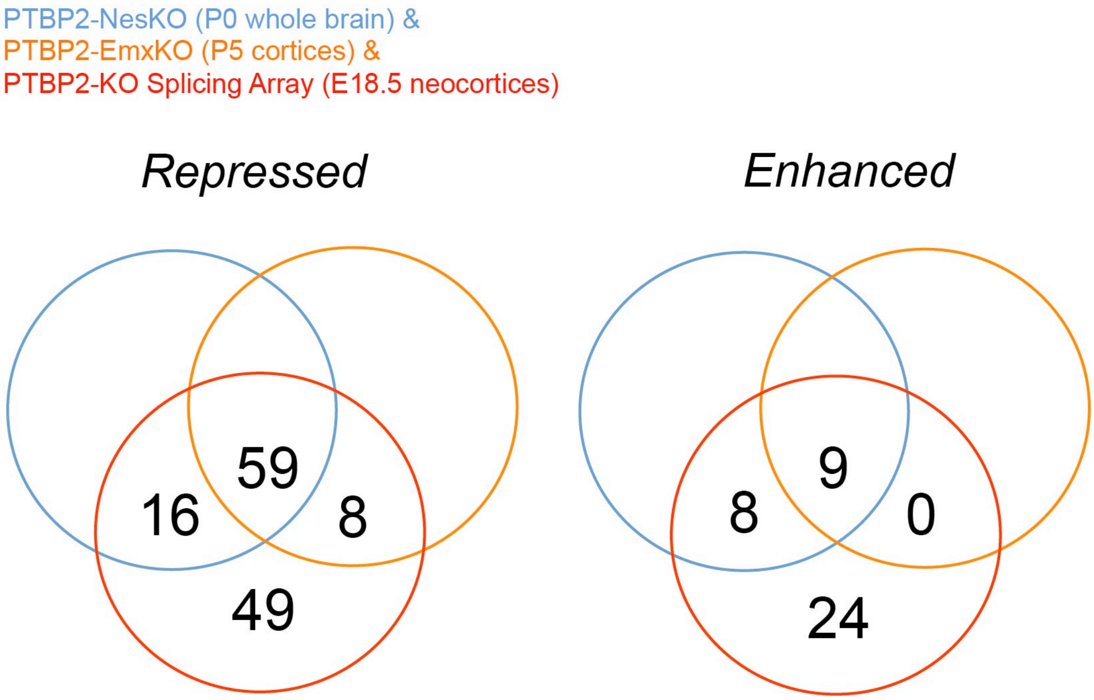


Figure 3. Overlaps of identified PTBP2 regulated exons from splicing microarray on PTBP2 null mouse study and RNA-SEQ experiments using two conditional PTBP2 knock out mouse lines, PTBP2-NesKO and PTBP2-EmxKO.

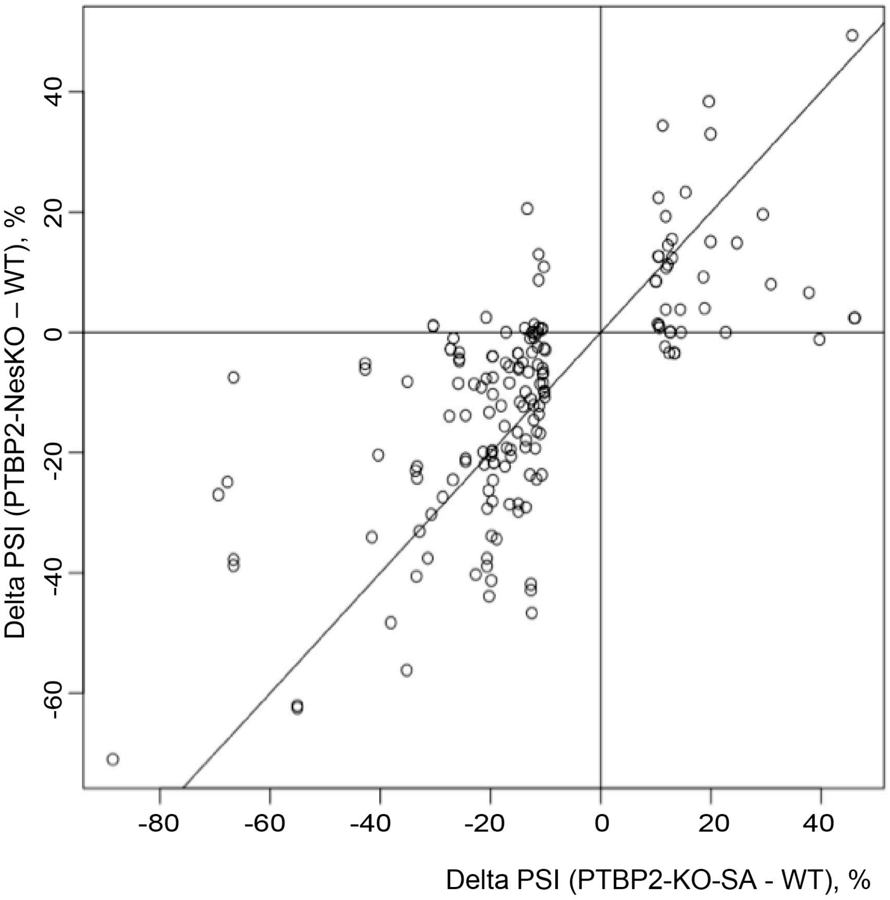


Figure 4. Comparison of the delta PSI (Percent Spliced In of an exon in wild type - Percent Spliced In of the exon in PTBP2 knock out mice) values for the PTBP2 regulated exons from PTBP2-KO array and PTBP2-NesKO RNA-SEQ.

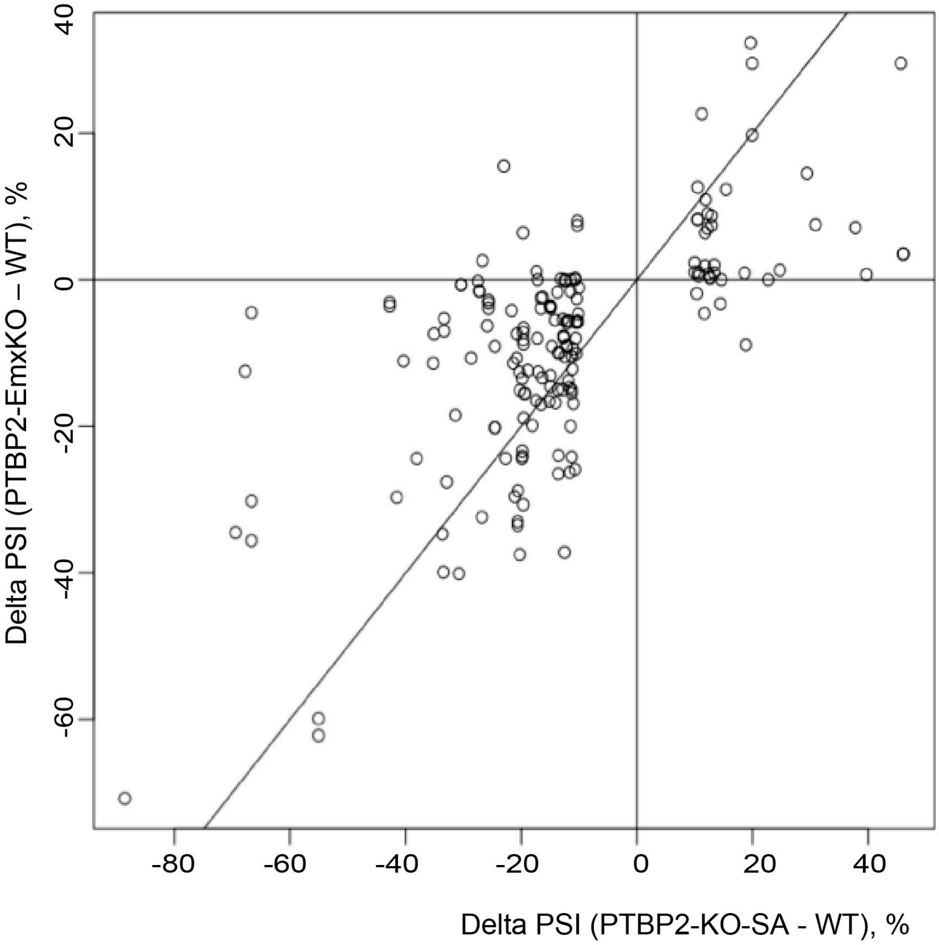


Figure 5. Comparison of the delta PSI (Percent Spliced In of an exon in wild type - Percent Spliced In of the exon in PTBP2 knock out mice) values for the PTBP2 regulated exons from PTBP2-KO array and PTBP2-EmxKO RNA-SEQ.

REFERENCES

Amir-Ahmady, B., Boutz, P.L., Markovtsov, V., Phillips, M.L., and Black, D.L. (2005). Exon repression by polypyrimidine tract binding protein. RNA *11*, 699-716.

Ashiya, M., and Grabowski, P.J. (1997). A neuron-specific splicing switch mediated by an array of pre-mRNA repressor sites: evidence of a regulatory role for the polypyrimidine tract binding protein and a brain-specific PTB counterpart. RNA *3*, 996-1015.

Ast, G. (2004). How did alternative splicing evolve? Nature reviews Genetics *5*, 773-782.

Auweter, S.D., Oberstrass, F.C., and Allain, F.H.T. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? Nucleic Acids Research *34*, 4943-4959.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53-59.

Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem *72*, 291-336.

Blencowe, B.J. (2006). Alternative splicing: New insights from global analyses. Cell *126*, 37-47.

Boutz, P.L., Chawla, G., Stoilov, P., and Black, D.L. (2007a). MicroRNAs regulate the expression of the alternative splicing factor nPTB during muscle development. Gene Dev *21*, 71-84.

Boutz, P.L., Stoilov, P., Li, Q., Lin, C.H., Chawla, G., Ostrow, K., Shiue, L., Ares, M., Jr., and Black, D.L. (2007b). A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. Genes Dev *21*, 1636-1652.

Carmel, I., Tal, S., Vig, I., and Ast, G. (2004). Comparative analysis detects dependencies among the 5' splice-site positions. RNA *10*, 828-840.

Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A., and Johnson, J.M. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nature genetics *40*, 1416-1425.

Cech, T.R. (1986). The generality of self-splicing RNA: relationship to nuclear mRNA splicing. Cell *44*, 207-210.

Chen, M., and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat Rev Mol Cell Biol *10*, 741-754.

Darnell, R.B. (2010). HITS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley Interdiscip Rev RNA *1*, 266-286.

Daubner, G.M., Clery, A., Jayne, S., Stevenin, J., and Allain, F.H. (2012). A syn-anti conformational difference allows SRSF2 to recognize guanines and cytosines equally well. The EMBO journal *31*, 162-174.

Durbin, R. (1998). Biological sequence analysis : probabalistic models of proteins and nucleic acids (Cambridge, UK New York, Cambridge University Press).

Fairbrother, W.G., Yeo, G.W., Yeh, R., Goldstein, P., Mawson, M., Sharp, P.A., and Burge, C.B. (2004). RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic acids research *32*, W187-190.

Gabut, M., Chaudhry, S., and Blencowe, B.J. (2008). SnapShot: The splicing regulatory machinery. Cell *133*, 192 e191.

Garcia-Blanco, M.A., Jamison, S.F., and Sharp, P.A. (1989). Identification and purification of a 62,000-dalton protein that binds specifically to the polypyrimidine tract of introns. Genes Dev *3*, 1874-1886.

Gooding, C., Roberts, G.C., and Smith, C.W. (1998). Role of an inhibitory pyrimidine element and polypyrimidine tract binding protein in repression of a regulated alpha-tropomyosin exon. RNA *4*, 85-100.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.C., Munschauer, M*., et al.* (2010). PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. Journal of visualized experiments : JoVE.

Hedrich Hans, G.R.B., Peter Petrusz, Gillian Bullock (2004). The laboratory mouse (Elsevier Academic Press).

Hosmer, D.W., and Lemeshow, S. (2000). Applied logistic regression, 2nd edn (New York, Wiley).

Hughes, A.L. (1994). The Evolution of Functionally Novel Proteins after Gene Duplication. P Roy Soc B-Biol Sci *256*, 119-124.

Izquierdo, J.M., Majos, N., Bonnal, S., Martinez, C., Castelo, R., Guigo, R., Bilbao, D., and Valcarcel, J. (2005). Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. Molecular cell *19*, 475-484.

Jin, Y., Suzuki, H., Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K., and Inoue, K. (2003). A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. EMBO J *22*, 905-912.

Keppetipola, N., Sharma, S., Li, Q., and Black, D.L. (2012). Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. Crit Rev Biochem Mol *47*, 360-378.

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2011). iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. J Vis Exp.

Lamichhane, R., Daubner, G.M., Thomas-Crusells, J., Auweter, S.D., Manatschal, C., Austin, K.S., Valniuk, O., Allain, F.H., and Rueda, D. (2010). RNA looping by PTB: Evidence using FRET and NMR spectroscopy for a role in splicing repression. Proc Natl Acad Sci U S A *107*, 4105-4110.

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature methods *7*, 709-715.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X.N., *et al.* (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464-U422.

Licatalosi, D.D., Yano, M., Fak, J.J., Mele, A., Grabinski, S.E., Zhang, C.L., and Darnell, R.B. (2012). Ptbp2 represses adult-specific splicing to regulate the generation of neuronal precursors in the embryonic brain. Gene Dev *26*, 1626-1642.

Llorian, M., Schwartz, S., Clark, T.A., Hollander, D., Tan, L.Y., Spellman, R., Gordon, A., Schweitzer, A.C., de la Grange, P., Ast, G*., et al.* (2010). Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. Nature structural & molecular biology *17*, 1114-1123.

Makeyev, E.V., Zhang, J., Carrasco, M.A., and Maniatis, T. (2007). The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. Mol Cell *27*, 435-448.

Matlin, A.J., Clark, F., and Smith, C.W.J. (2005). Understanding alternative splicing: Towards a cellular code. Nat Rev Mol Cell Bio *6*, 386-398.

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nature genetics *30*, 13-19.

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

Nagai, K., and Mattaj, I.W. (1994). RNA-protein interactions (Oxford ; New York, IRL Press at Oxford University Press).

Oberstrass, F.C., Auweter, S.D., Erat, M., Hargous, Y., Henning, A., Wenter, P., Reymond, L., Amir-Ahmady, B., Pitsch, S., Black, D.L*., et al.* (2005). Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science *309*, 2054-2057.

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

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res *37*, e123.

Perez, I., McAfee, J.G., and Patton, J.G. (1997). Multiple RRMs contribute to RNA binding specificity and affinity for polypyrimidine tract binding protein. Biochemistry *36*, 11881-11890.

Rabiner, L.R. (1989). A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition. P Ieee *77*, 257-286.

Ray, D., Kazan, H., Chan, E.T., Pena Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat Biotechnol *27*, 667-670.

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

Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol *132*, 365-386.

Schuelke, M. (2000). An economic method for the fluorescent labeling of PCR fragments. Nature biotechnology *18*, 233-234.

Shapiro, M.B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic acids research *15*, 7155-7174.

Sharma, S., Kohlstaedt, L.A., Damianov, A., Rio, D.C., and Black, D.L. (2008). Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. Nat Struct Mol Biol *15*, 183-191.

Shen, H.H., Kan, J.L.C., Ghigna, C., Biamonti, G., and Green, M.R. (2004). A single polypyrimidine tract binding protein (PTB) binding site mediates splicing inhibition at mouse IgM exons M1 and M2. Rna-a Publication of the Rna Society *10*, 787-794.

Shibayama, M., Ohno, S., Osaka, T., Sakamoto, R., Tokunaga, A., Nakatake, Y., Sato, M., and Yoshida, N. (2009). Polypyrimidine tract-binding protein is essential for early mouse development and embryonic stem cell proliferation. FEBS J *276*, 6658-6668.

Singh, L.N., and Hannenhalli, S. (2008). Functional diversification of paralogous transcription factors via divergence in DNA binding site motif and in expression. PLoS One *3*, e2345.

Singh, R., Valcarcel, J., and Green, M.R. (1995). Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science *268*, 1173-1176.

Smith, C.W.J. (1998). RNA-protein interactions : a practical approach (Oxford ; New York, IRL Press at Oxford University Press).

Spellman, R., Llorian, M., and Smith, C.W.J. (2007). Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. Molecular cell *27*, 420-434.

Spellman, R., and Smith, C.W. (2006). Novel modes of splicing repression by PTB. Trends Biochem Sci *31*, 73-76.

Suckale, J., Wendling, O., Masjkur, J., Jager, M., Munster, C., Anastassiadis, K., Stewart, F., and Solimena, M. (2011). PTBP1 Is Required for Embryonic Development before Gastrulation. Plos One *6*.

Tang, Z.Z., Sharma, S., Zheng, S.K., Chawla, G., Nikolic, J., and Black, D.L. (2011). Regulation of the Mutually Exclusive Exons 8a and 8 in the CaV1.2 Calcium Channel Transcript by Polypyrimidine Tract-binding Protein. J Biol Chem *286*, 10007-10016.

Teichmann, S.A., and Babu, M.M. (2004). Gene regulatory network growth by duplication. Nature genetics *36*, 492-496.

Ule, J., Jensen, K., Mele, A., and Darnell, R.B. (2005). CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods *37*, 376-386.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212-1215.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B.J., and Darnell, R.B. (2006). An RNA map predicting Nova-dependent splicing regulation. Nature *444*, 580-586.

Valcarcel, J., and Gebauer, F. (1997). Post-transcriptional regulation: the dawn of PTB. Curr Biol *7*, R705-708.

Wahl, M.C., Will, C.L., and Luhrmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell *136*, 701-718.

Wang, E.T., Cody, N.A., Jog, S., Biancolella, M., Wang, T.T., Treacy, D.J., Luo, S., Schroth, G.P., Housman, D.E., Reddy, S*., et al.* (2012). Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. Cell *150*, 710-724.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

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

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

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119*, 831-845.

Wu, J., Akerman, M., Sun, S., McCombie, W.R., Krainer, A.R., and Zhang, M.Q. (2011). SpliceTrap: a method to quantify alternative splicing under single cellular conditions. Bioinformatics *27*, 3010-3016.

Xing, Y., Stoilov, P., Kapur, K., Han, A., Jiang, H., Shen, S., Black, D.L., and Wong, W.H. (2008). MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. RNA *14*, 1470-1479.

Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.S., Zhang, C., Yeo, G., Black, D.L., Sun, H., *et al.* (2009). Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. Mol Cell *36*, 996-1006.

Yeo, G.W., Coufal, N.G., Liang, T.Y., Peng, G.E., Fu, X.D., and Gage, F.H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nature structural & molecular biology *16*, 130-137.

Zhang, C., Frias, M.A., Mele, A., Ruggiu, M., Eom, T., Marney, C.B., Wang, H., Licatalosi, D.D., Fak, J.J., and Darnell, R.B. (2010). Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. Science *329*, 439-443.

Zhang, J.Z. (2003). Evolution by gene duplication: an update. Trends Ecol Evol *18*, 292-298.

Zheng, S., Gray, E.E., Chawla, G., Porse, B.T., O'Dell, T.J., and Black, D.L. (2012). PSD-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2. Nat Neurosci *15*, 381-U202.

Zhu, Z., Pilpel, Y., and Church, G.M. (2002). Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. J Mol Biol *318*, 71-81.