1-1-2015

Terrae Incognitae: Integrative Genomic Analysis of Hnrnp L Splicing Regulation

Brian Sebastian Cole
University of Pennsylvania, colebr@mail.med.upenn.edu

Terrae Incognitae: Integrative Genomic Analysis of Hnrnp L Splicing Regulation

Abstract

Alternative splicing is a critical component of human gene control that generates functional diversity from a limited genome. Defects in alternative splicing are associated with disease in humans. Alternative splicing is regulated developmentally and physiologically by the combinatorial actions of cis- and trans-acting factors, including RNA binding proteins that regulate splicing through sequence-specific interactions with pre-mRNAs. In T cells, the splicing regulator hnRNP L is an essential factor that regulates alternative splicing of physiologically important mRNAs, however the broader physical and functional impact of hnRNP L remains unknown. In this dissertation, I present analysis of hnRNP L-RNA interactions with CLIP-seq, which identifies transcriptome-wide binding sites and uncovers novel functional targets. I then use functional genomics studies to define pre-mRNA processing alterations induced by hnRNP L depletion, chief among which is cassette-type alternative splicing. Finally, I use integrative genomic analysis to identify a direct role for hnRNP L in repression of exon inclusion and an indirect role for enhancing exon inclusion that supports a novel regulatory interplay between hnRNP L and chromatin. In two appendices, I present CLIP-seq studies of two additional RNA binding proteins: the splicing factor CELF2 and the RNA helicase DDX17. In conclusion, I provide comparisons of these three CLIP-seq studies, providing high-level insights into the capabilities and limitations of CLIP-seq. In sum, this dissertation expands our knowledge of hnRNP L splicing control in the context of broader studies of RNA binding proteins in multiple cell types and conditions.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Cell & Molecular Biology

First Advisor

Kristen W. Lynch

Keywords

Alternative splicing, CLIP-seq, hnRNP L, integrative genomics, RNA-seq, T cells

Subject Categories

Allergy and Immunology | Biochemistry | Bioinformatics | Immunology and Infectious Disease | Medical Immunology

TERRAE INCOGNITAE: INTEGRATIVE GENOMIC ANALYSIS OF hnRNP L SPLICING REGULATION

Brian Sebastian Cole

A DISSERTATION

in

Cell and Molecular Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation

____

Kristen W. Lynch, Ph.D.

Professor of Biochemistry and Biophysics

Graduate Group Chairperson

____

Daniel S. Kessler, Ph.D., Associate Professor of Cell and Developmental Biology

Dissertation Committee

Russ P. Carstens, M.D., Associate Professor of Medicine

Brian D. Gregory, Ph.D., Assistant Professor of Biology

Yoseph Barash, Ph.D., Assistant Professor Genetics

Stephen A. Liebhaber, M.D., Professor of Genetics and Medicine

TERRAE INCOGNITAE: INTEGRATIVE GENOMIC ANALYSIS OF hnRNP L SPLICING REGULATION

DEDICATION

I dedicate this dissertation to my thesis advisor, Kristen Lynch. Her wisdom and strength of character will be with me always.

ACKNOWLEDGMENT

I acknowledge the many who have supported me in my research, both technically and personally. I'd like to thank the legions of open source programmers who have worked for Freedom. Without open source programs, I'd never have been able to take the journey I've taken over the past five years. I'd like to thank my family and friends for all their support over the past five years: my mom, my dad, my brother, Leksa, Mike, Ian, Roland, Sheila and Dan, Matt, the Venez Voir crew, Kyle and Andy and everybody at PBG, Rohit and Christina and everybody at WHIT, Craig and Charles over at Penn Health, Jordan and Bren and Luigi and the Data Incubator crew from New York City, Matt and the Beer Club, Mr. Nap the mathiest lawyer who ever drummed sevens and fives, and everybody else who has kept me afloat.

ABSTRACT

TERRAE INCOGNITAE: INTEGRATIVE GENOMIC ANALYSIS OF hnRNP L SPLICING REGULATION

Brian Sebastian Cole

Kristen W. Lynch, Ph.D.

Alternative splicing is a critical component of human gene control that generates functional diversity from a limited genome. Defects in alternative splicing are associated with disease in humans. Alternative splicing is regulated developmentally and physiologically by the combinatorial actions of cis- and trans-acting factors, including RNA binding proteins that regulate splicing through sequence-specific interactions with pre-mRNAs. In T cells, the splicing regulator hnRNP L is an essential factor that regulates alternative splicing of physiologically important mRNAs, however the broader physical and functional impact of hnRNP L remains unknown. In this dissertation, I present analysis of hnRNP L-RNA interactions with CLIP-seq, which identifies transcriptome-wide binding sites and uncovers novel functional targets. I then use functional genomics studies to define pre-mRNA processing alterations induced by hnRNP L depletion, chief among which is cassette-type alternative splicing. Finally, I use integrative genomic analysis to identify a direct role for hnRNP L in repression of exon inclusion and an indirect role for enhancing exon inclusion that supports a novel regulatory interplay between hnRNP L and chromatin. In two appendices, I present CLIP-seq studies of two additional RNA binding proteins: the splicing factor CELF2 and the RNA helicase DDX17. In conclusion, I provide comparisons of these three CLIP-seq studies, providing high-level insights into the capabilities and limitations of CLIP-seq. In sum, this dissertation expands our knowledge of hnRNP L splicing control in the context of broader studies of RNA binding proteins in multiple cell types and conditions.

TABLE OF CONTENTS

LIST OF TABLES

MATERIALS AND METHODS

CHAPTER 1 – MAPPING TRANSCRIPTOME-WIDE HNRNP L-RNA INTERACTIONS BY COMPUTATIONAL ANALYSIS OF CLIP-SEQ DATA

CHAPTER 2 – DISCOVERY OF HNRNP L-REGULATED ALTERNATIVE SPLICING WITH RASL-SEQ AND MRNA-SEQ

CHAPTER 3 – INTEGRATIVE GENOMIC ANALYSIS OF HNRNP L SPLICING REGULATORY FUNCTION

Appendix 2 – CLIP-SEQ ANALYSIS OF CELF2-RNA INTERACTIONS IN T CELLS

CONCLUSION

LIST OF FIGURES

INTRODUCTION

Alternative pre-mRNA splicing is a nearly ubiquitous mechanism by which eukaryotic cells generate multiple protein-coding mRNAs from a single genetic locus[1]. As many as 95% of human genes that give rise to multiexon mRNAs generate more than one processed product by alternative pre-mRNA splicing[2,3,3]. Importantly, alternative splicing allows an abundance of distinct proteins to be encoded by a limited genome. Recent technological advances in high-throughput sequencing have made possible genome-wide studies of alternative splicing and the factors that regulate splicing. These high-throughput studies have necessitated commensurate software development efforts to process the large volumes of data generated. In this thesis, I present the design, analysis, and interpretation of several high-throughput sequencing studies that together aim to uncover the scope of alternative splicing in human lymphocytes, as well as key proteins that regulate alternative splicing.

The spliceosome catalyzes splicing through serial assembly and rearrangements

Pre-mRNA splicing is orchestrated by the stepwise assembly and two-step transesterification enzymatic catalysis of the macromolecular ribonucleoprotein (RNP) complex known as the spliceosome[4]. The spliceosome contains no preformed active site and instead relies upon a vectorial, multistep assembly and catalysis mechanism[5]. The core of the spliceosome is composed of five distinct subunits called the U1, U2, U5, and U4/U6 snRNPs (small nuclear ribonucleoprotein particles) that contain RNA and protein components. In addition to the snRNPs, the spliceosome contains more than

100 accessory proteins, together forming a heterogeneous and complex macromolecular machine[4].

During pre-mRNA splicing, the snRNPs assemble at their cognate splice sites on pre-mRNA in a stepwise manner, beginning with the recognition of the 5' splice site (5'ss) by the U1 snRNP, whose RNA component, the U1 snRNA, engages the 5'ss through RNA-RNA interactions. Recognition of the 3'ss at the other end of the intron is performed first through protein-RNA interactions: SF1 binds to the intronic branchpoint sequence (BPS) and the U2AF binds to the polypyrimidine tract and 3'ss[6]. The result of the recognition of the 5' and 3' splice sites is the E complex. Subsequent replacement of the proteins engaged at the 3'ss by the U2 snRNP forms the A complex.

After the A complex has formed, the trimer of the U4/5/6 "tri-snRNP" is recruited, forming the pre-catalytic B complex. Through a series of structured rearrangements, the U1 snRNP is displaced from the 5'ss, with U6 replacing it. These rearrangements generate a catalytically competent spliceosome, or C complex. The U4 snRNP is then released and the first transesterification reaction is catalyzed, in which a lariat structure is formed between the 5' end of the intron is cleaved and a lariat formation is created through attachment of the 5' end of the intron to the branchpoint adenosine. The second transesterification reaction cleaves the 3' end of the intron and ligates the exons spanning the intron together, resulting in release of the lariat and the splice pre-mRNA[7].

Regulation of splicing is achieved by RNA binding proteins

Importantly, the spliceosomal assembly pathway is characterized by several resolvable complexes that require multiple RNA-RNA, RNA-protein, and protein-protein interactions to allow progression to later complexes[8,9,9]. This mechanism provides many opportunities for regulation, enabling alternative splicing to exert fine-grained control over the structure of processed mRNAs. Regulation of alternative splicing is achieved by the interaction of trans-acting RNA binding proteins (RBPs) with cis-regulatory RNA sequences[5]. These interactions can serve to both positively and negatively regulate assembly and progression of the spliceosome at multiple steps of the spliceosomal assembly pathway.

Several well-studied examples of regulated alternative splicing highlight the importance of the location of protein-RNA interaction. One of the first systems used to understand splicing regulation was sex determination in *Drosophila melanogaster*. The RNA binding protein Sex lethal (Sxl) is expressed in females and represses splicing of male-specific mRNA isoforms. Biochemical studies of the Sxl target mRNA Transformer (Tra) indicated that Sxl binds to a sequence in the 3'ss of Tra exon 2, resulting in competition with the U2AF protein[10,11,11,12,12]. Additionally, Sxl can block binding of the U1 snRNP to the 5'ss of an exon in another target transcript, Msl2[13]. In both cases, Sxl binding prevents formation of the E complex, demonstrating a simple mechanism of competitive binding that results in splicing repression.

Another splicing regulatory protein involved in the *Drosophila* sex determination pathway is the Sxl target Transformer (Tra). Tra cooperates with a binding partner, Tra2, bind to a splicing enhancer sequence in the fourth exon of the doublesex (Dsx) pre-mRNA. When Tra and Tra2 bind to this enhancer, U2AF is recruited to the relatively weak 3'ss of the Dsx exon[14]. These two cases highlight the regulatory roles that RNA binding proteins can play in enhancing or repressing splicing, and additionally underscore the importance of the location of binding in determining splicing regulation.

Importantly, while these early studies implicated splicing regulation at the early steps of spliceosome assembly, subsequent research has identified instances of splicing regulation by RNA binding proteins at later steps (see below). While alternative splicing in humans is often complex and multifactorial, the insights gleaned from studies of the *Drosophila* sex determination pathway and the factors that regulate it provide a basis for investigating alternative splicing in humans.

Two major, conserved families of RNA-binding proteins (RBPs) with splicing regulatory functions are the SR and hnRNP protein families. SR proteins contain one or more RNA recognition motifs (RRMs) and a serine/arginine rich RS domain[15,16,16]. Both the RS and RRM domains may be involved in protein-RNA and protein-protein interactions. SR proteins were initially identified as splicing enhancers capable of activating splicing reactions in vitro[17,18,18,19,19,20,20]. The hnRNP proteins were originally identified by their association with pre-mRNA[21,22,22], and splicing regulatory roles for hnRNP proteins were later identified[23]. The development of high-throughput sequencing has enabled transcriptome-wide functional studies of SR and hnRNP proteins, and recent work has demonstrated that members of these protein families control virtually all aspects of RNA processing.

Several previously studied examples of alternative splicing events regulated by the interaction of SR and hnRNP proteins with *cis*-regulatory pre-mRNA motifs demonstrate overarching paradigms for the mechanisms by which these RNA-binding proteins are believed to exert control over pre-mRNA splicing. The human hnRNP A1 protein is a well-studied splicing factor with generally repressive activity[24]. hnRNP A1 can repress cassette exon inclusion by binding to exonic or intronic splicing silencer sequences, termed ESS and ISS respectively, and directly block formation of the spliceosome through steric hindrance[25,26,26]. In addition this bind-and-block mechanism, hnRNP A1 can bind to a high-affinity site and propagate interactions to neighboring pre-mRNA sites in a spreading mechanism. Importantly, this spreading allows hnRNP A1 to interfere with binding of splicing activators to nearby exonic splicing enhancer (ESE) sequences, as was observed in HIV tat exon 3[27]. This latter case highlights one instance in which multiple *trans*-acting RNA binding proteins interact with each other in an exonic context to regulate splicing, highlighting the integrative and combinatorial modes of regulation that characterize alternative splicing in humans. Another dimension of complexity in the regulation of alternative splicing is underscored by the fact that hnRNP A1 can also act as a splicing activator[28,29,29,30,30], an observation common to many other splicing factors, which have been observed to exert both positive and negative regulation of splicing.

Another well-studied splicing regulator in humans is SRSF1 (formerly denoted SF2 or ASF), an archetypical member of the SR protein family. In many described cases, SRSF1 positively regulates cassette exon splicing by binding to exonic splicing enhancer (ESE) sequences[31]. When SRSF1 is bound to the pre-mRNA, direct interactions with the U1 snRNP help to recruit U1 and initiate assembly of the spliceosomal E complex[32]. While this simple model of splicing enhancement is a mirror of the bind-and-block mechanism of splicing repression by hnRNP A1, genome-wide analysis of SRSF1-regulated splicing events has uncovered more complex scenarios[33]. In some instances, loss of SRSF1 results in compensatory splicing enhancement by other SR proteins, such as SRSF2. In other cases, there is negative coordination, in which loss of SRSF1 results in loss of binding by other SR proteins, leading to exon skipping. The mechanisms underlying the coordinate versus compensatory regulation remain unclear, but SRSF1 emphasizes another key concept in regulated alternative splicing: regulation of splicing likely involves a complex interplay of RNA binding proteins with complex interactions, motivating a systems-level understanding of splicing regulation.

hnRNP L is an essential regulator of alternative splicing in T cells

Human T lymphocytes are a critical cell line of the adaptive immune system that utilize alternative splicing to regulate the proteome during development and physiological activity[34]. One example of dynamic alternative splicing in T cells is the CD45 transmembrane protein tyrosine phosphatase, encoded by the gene PTPRC. The CD45 protein contains a heavily glycosylated extracellular domain that maintains the phosphatase activity by inhibiting homodimerization which would result in loss of activity[35,36,36]. This extracellular domain is encoded by three tandem alternative exons in the CD45 pre-mRNA which are increasingly skipped in response to T cell activation[37]. This splicing switch results in a shift toward CD45 isoforms that do not contain the glycosylated extracellular domain and therefore can homodimerize, resulting in loss of phosphatase activity in activated T cells.

Studies of sequences responsible for this activation-responsive splicing switch in CD45 exon 4 identified an exonic splicing silencer motif, ESS1[38]. The ESS1 motif establishes a reduced basal level of inclusion. Importantly, this basal repression poises the exon for further activation-induced repression enabled by the activation-dependent repressive activity of other factors such as PSF[39] and hnRNP LL[40].

Subsequent in vivo studies of hnRNP L have identified a physiological role for hnRNP L in T cell development and function. A lymphoid-specific hnRNP L gene ablation model system in Mus musculus displays defects in thymic development of T cells, with a specific block of the double negative to double positive stage observed among thymic pre-T cells[41] and defects in migration of hnRNP L -/- T cells in the periphery. The T cells of this mouse model display a noticeable shift towards higher molecular weight CD45 isoforms, indicating decreased repression of the exons that encode the activation-responsive extracellular domain. However, additional physiological defects in chemokine-induced migration among hnRNP L -/- T cells were observed, indicating that hnRNP L ablation results in additional physiological defects not attributable to CD45 misregulation alone.

hnRNP L is an abundant nuclear splicing factor with four conserved RRM domains, a glycine-rich N-terminal domain, and a proline-rich linker domain[42]. Biochemical evidence reveals that sequences within the latter two domains are required for exon repression, but maximal repressive activity requires at least one RRM sequence. The hnRNP L protein is ubiquitously expressed and is required for viability, and while the paralog protein hnRNP L-like exhibits cell type-specific expression[40,43,43], hnRNP L expression remains constant between unstimulated and stimulated T cells. Accordingly, hnRNP L repression of CD45 exon 4 splicing is observed in both T cell conditions.

A second exon in the CD45 pre-mRNA, exon 5, is also repressed by hnRNP L through exonic interactions, but the mechanism of repression is distinct for these two exons[44]. In the case of CD45 exon 5, hnRNP L binds to an exonic splicing silencer to block the activity of a neighboring splicing enhancer[45] that is bound by SF2/ASF to enhance exon inclusion. When this splicing enhancer is removed, hnRNP L regulation becomes enhancing, demonstrating that location of hnRNP L-RNA interaction alone is not sufficient to predict regulatory outcome. In addition, when the 3' and 5' splice sites of an hnRNP L-bound exon are weakened by mutagenesis, repression by hnRNP L is abolished and then reversed at extremely weak splice site strengths. In contrast, hnRNP L interaction with CD45 exon 4 recruits another splicing factor, hnRNP A1, to the exon, resulting in an extended interaction between the U1 snRNA and the 5' splice site of the exon, resulting in exon repression by blocking the transition of spliceosomal E complex to higher order complexes[46]. Taken together, these results support a model for context-dependent regulation of splicing by hnRNP L wherein splice site strengths and co-associated proteins form a combinatorial code that leads to positive or negative regulation of splicing by hnRNP L.

In addition to CD45, hnRNP L is directly involved in regulation of several other mRNAs. In a mechanism common to many splicing regulatory proteins across diverse species of life[47], hnRNP L protein binds to an evolutionarily conserved, CA-rich region in its own pre-mRNA upstream of an exon that contains a premature termination codon, a type of exon called a poison exon, as its inclusion leads to unproductive mRNAs that are degraded by the nonsense-mediated decay (NMD) pathway. This binding event is part of an autoregulatory mechanism wherein increased hnRNP L protein levels result in increased binding to the conserved, CA-rich region, which in turn results in increased inclusion of the poison exon, causing a shift toward unproductive mRNA isoforms that result in decreased protein expression[48]. This example also highlights a positive regulatory role for hnRNP L on alternative splicing, indicative of a potential direct splicing enhancer function for hnRNP L.

Another well-studied protein that is subject to alternative splicing is CD44. CD44 is a cell adhesion protein[49] encoded by a pre-mRNA that contains a variable exon whose inclusion is directly regulated by hnRNP L[50]. In this case, hnRNP L binds to CA repeats upstream of the variable exon to repress exon inclusion, providing evidence that hnRNP L can act as a splicing repressor from an upstream intronic binding site. Interestingly, the CD44 variable exon whose inclusion is repressed by hnRNP L, exon V10, has been implicated in leukocyte migration[51] as well as tumor progression[52], providing an additional instance of hnRNP L-mediated alternative splicing that is critical to T cell function and cancer biology.

In addition to the regulation of alternative splicing, hnRNP L-3'UTR interactions have been implicated in the hypoxia-induced stabilization of the VEGFA mRNA[53]. In this case, hypoxia induces cytoplasmic localization of hnRNP L protein, allowing it to interact with a CA-rich hypoxia-stability region in the 3'UTR of the VEFGA mRNA[54]. This interaction blocks miRNA-mRNA interactions that would otherwise result in translational repression, thereby increasing VEGFA translation. Importantly, this mechanism highlights an extranuclear function for hnRNP L that is induced by abnormal cellular conditions which underscores the importance of nuclear localization for hnRNP L.

While several cases of hnRNP L-regulated alternative splicing have been identified, including CD45 and HNRNPL, the impact of hnRNP L on the T cell transcriptome has not been determined, and the physical and functional targets of hnRNP L are unknown. Recent experimental advances with high-throughput sequencing technology have enabled genome-wide characterization of RNA targets of RNA binding proteins, and functional studies utilizing RBP knockdown followed by transcriptome sequencing and bioinformatic analysis have enabled detailed characterization of the direct and indirect functional targets of splicing factors.

Technologies for transcriptome-wide investigation of RNA binding proteins

The CLIP protocol (Crosslinking and immunoprecipitation) was developed to capture associations between proteins and the RNAs with which they are in direct contact[55,56,56]. This procedure utilizes UV irradiation to induce covalent crosslinks between proteins and RNA across short distances, and subsequent immunoprecipitation of an RBP of interest under stringent conditions followed by proteinase digestion purifies RNA targets of the protein under study. In a landmark study, Jernej Ule and Robert Darnell used CLIP to identify physical targets of the Nova splicing factor[57]. This study identified Nova-dependent splicing targets among physical targets of Nova, demonstrating that CLIP can be used to identify physical as well as functional targets of critical splicing regulators. Importantly, CLIP can utilize biological tissues as well as cells in culture, allowing physical targets to be identified in vivo, as was demonstrated for Nova in brain tissue.

The advent of affordable high-throughput sequencing technology allowed deep characterization of UV-crosslinked RNA-protein complexes, an experimental approach known as CLIP-seq or HITS-CLIP (Crosslinking and immunoprecipitation followed by high-throughput sequencing). This method was applied to Nova-RNA complexes isolated from brain tissue[58], greatly expanding the scope of discovered interaction sites. By digesting the RNA liberated from Nova-RNA complexes to short oligonucleotide fragments, the high-throughput sequencing reads generated from these fragments can be aligned to the reference genome of the organism under study, allowing identification of specific regions of protein-RNA interaction within physical target RNAs. This approach allowed identification of novel Nova-RNA interactions in 3'UTR regions, revealing a novel role for Nova in alternative polyadenylation.

The CLIP-seq method has been successfully applied to many RNA binding proteins, including splicing factors[59,60,60,61,61,62,62,63,63,64,64,65,65], the miRNA effector protein Argonaute[66,67,67,68,68], the RNA editing factor ADAR[69], oncogenic fusion proteins involving RNA-binding domains[70], RNA helicases[71], and others. Through these studies, CLIP-seq has emerged as a powerful tool for identifying sites of RBP-RNA interactions, a critical step in understanding the role a particular RBP plays in regulating the transcriptome in vivo. Recent efforts to centralize the findings of these experiments have led to the creation of a curated database containing CLIP-seq studies and their data[72].

Functional studies of splicing factors with next-generation sequencing reveal splicing regulatory activity

Technological advances have enabled wide-scale characterization of alternative splicing in eukaryotic transcriptomes. Early high-throughput studies of alternative splicing utilized microarray technology. Evidence of the applicability of microarrays to the discovery of alternatively spliced exons came from early genome and exome tiling array studies aimed at elucidating the structure of the human transcriptome[73]. This technology developed into exon-junction arrays[74,75,75] that provided probes focused on splicing junctions, enabling large-scale profiling of alternative splicing within and between cell types[76,77,77,78,78]. These pioneering studies provided evidence of the massive complexity of transcriptome processing and brought an appreciation of alternative splicing as a means of increasing proteomic complexity and diversity from a limited genome.

Comparative microarray studies were soon applied to the identification of functional targets of splicing factors. A study of brain tissue from Nova2-/- murine brain tissue identified hundreds of Nova-dependent splicing targets[55], paving the way for studies of other splicing factors in a plethora of different tissues and cell types[79]. One important overarching theme that arose from these studies was a coupling of pre-mRNA processing events, including splicing, to Pol II elongation, providing evidence for a co-transcriptional model of pre-mRNA processing[80,81,81].

While microarray technology facilitated wide-scale characterization of RNAs, the reliance upon pre-designed probes prevented de novo characterization of splice variants. Transcriptome characterization by high-throughput sequencing has recently enabled transcriptome-wide detection of novel splice variants in addition to those probed by splicing-sensitive microarrays[82,83,83]. High-throughput sequencing of cDNAs derived from cellular RNAs, or RNA-seq, spurned the development of new bioinformatics tools to meet the challenges of sequencing reads that align to the genome in multiple segments due to their origin in spliced mRNA molecules, such as the Tophat aligner that is aimed specifically at RNA-seq reads[84]. Alignment tools such as Tophat have enabled the development of downstream statistical algorithms for quantification of alternative splice junction utilization between two sample groups[85,86,86], allowing studies of alternative splicing controlled by differentiation, signaling, and specific splicing factors.

In this dissertation, I describe the application of CLIP-seq to identify the transcriptome-wide binding profile of hnRNP L in primary and cultured human CD4+ T lymphocytes. Analysis of hnRNP L CLIP-seq sites identifies novel cases of hnRNP L-mediated alternative splicing. I further describe depletion of hnRNP L followed by complementary high-throughput sequencing approaches, which allows transcriptome-wide characterization of hnRNP L-dependent alternative splicing events. Finally, I analyze hnRNP L occupancy in the context of hnRNP L-responsive alternative splicing events, providing mechanistic insights into direct and indirect regulation of alternative splicing by hnRNP L.

MATERIALS AND METHODS

CLIP-seq read processing and alignment

Raw CLIP-seq reads are single, continuous oligonucleotide sequences in FASTQ formatted-files, which include sequencing basecall quality scores for each nucleotide encoded in PHRED quality scores. The first step in the processing raw CLIP-seq reads is to remove 3' sequencing adaptors and any sequence that might remain to the 3' of the end of the adaptor, which might result from an extremely short CLIP-seq fragment and a long sequencing read length. I utilized the cutadapt version 0.9.4, invoking the cutadapt program with default options and providing the RL3 linker sequence (5'-GTGTCAGTCACTTCCAGCGG-3').

Next, homopolymeric stretches of 6 or more nucleotides of the same basecall were removed from the 3' end of the reads. The FASTQ quality string was also trimmed to match the length of the trimmed sequence read. The resulting trimmed reads were discarded if they were fewer than 8 nucleotides in length, and the remaining reads were utilized for alignment. Homopolymeric runs were removed with the program *trim_homopolymeric_ends* of the clipseq_analysis distribution, a collection of Perl programs and libraries I wrote to process the CLIP-seq datasets described in this dissertation.

To obtain alignments for trimmed CLIP-seq reads, bowtie version 0.12.7 was used to map reads against the hg19 build of the human genome, allowing for a maximum of 2 mismtaches and disallowing more than 1 alignment position per read. Because the CLIP-seq library preparation involves 39 cycles of PCR, multiple reads that align to the same genomic coordinates are not guaranteed to represent separate RNA fragments that were crosslinked to the protein under study. For this reason, any set of two or more aligned reads that share the same 5’ coordinate was reduced to only 1 representative, and all other alignments from that 5’ coordinate were discarded. To collapse these potential PCR duplicates, I used the program *collapse_duplicates* in the clipseq_analysis distribution. For hnRNP L and DDX17 CLIP-seq experiments, one barcode was applied to each sequencing library, therefore libraries from replicates within sample groups were combined after collapsing PCR duplicates. For CELF2 CLIP-seq, however, a new strategy was employed to aid discrimination of PCR duplicates: before the PCR reactions, the post-doctoral researcher performing the CLIP-seq library preparation, Dr. Ganesh Shankarling, split each replicate into three aliquots and applied a different barcode to each of aliquot. Subsequent processing, alignment, and removal of potential PCR duplicates from each barcoded library separately allowed me to pool the three barcoded aliquots from each individual replicate such that there can now be a maximum of three alignments with a 5’ end at any given genomic coordinate. This approach thus allowed me to discriminate *bona fide* multi-copy RNA-protein complexes from PCR duplicates.

After PCR duplicates were removed and replicates were combined, the aligned reads were ready to be searched for peaks: regions of CLIP-seq coverage that exceed that which is expected by random chance. Initially, I had attempted to call peaks on each replicate separately and then combine them, however discovery power suffered under this approach: overlap between replicates was less than 50%, and the hnRNP L binding site within CD45 exon 4, a site of known hnRNP L-RNA interaction, was lost (see results sections below). In accordance with methods employed by prior CLIP-seq analyses, I combined replicates within sample groups before peak discovery rather than calling peaks on individual replicates and attempting to define shared sites based on overlap between individual replicates (figure M1). In order to preserve replicate support, I arrived upon a strategy in which peaks were called on the pooled replicates from each sample group, and then individual peaks were discarded if they did have coverage in the aligned reads of at least 2 replicates from that sample group. The resulting strategy provides a balance between the greater discovery power provided by pooled replicates within sample groups and the requirement that all CLIP-seq peaks had replicate support among constituent high-throughput sequencing libraries.

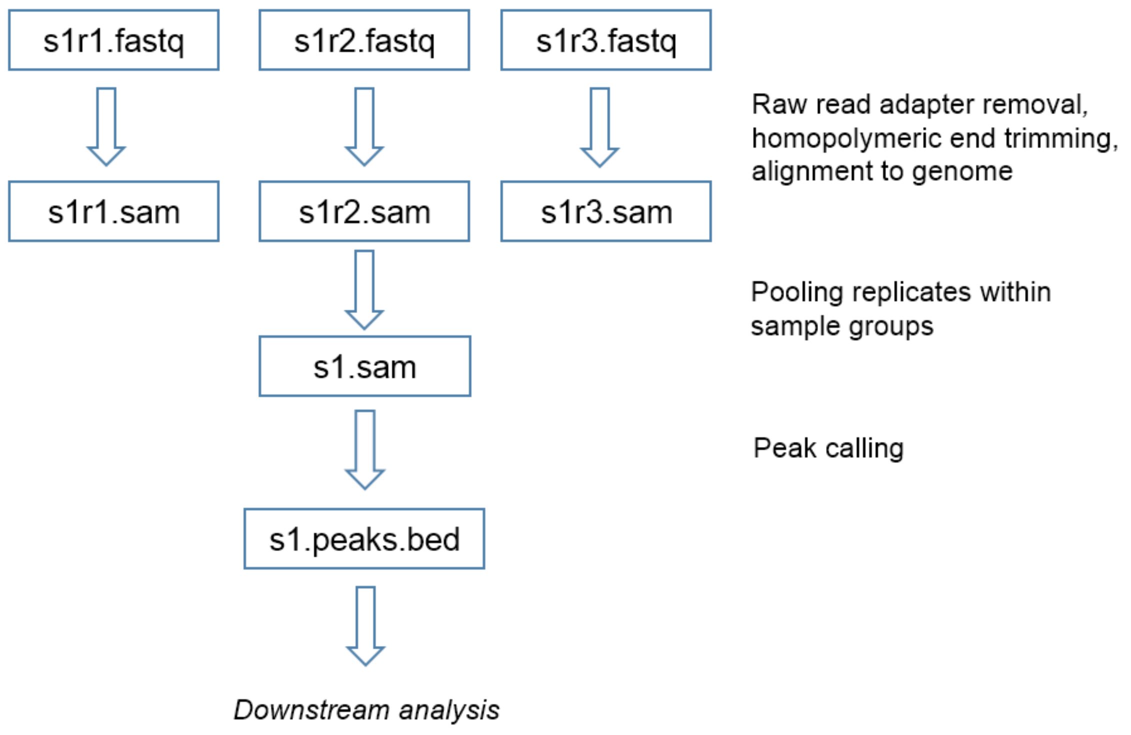

s1r1.fastq
s1r2.fastq
s1r3.fastq
Raw read adapter removal, homopolymeric end trimming, alignment to genome
s1r1.sam
s1r2.sam
s1r3.sam
Pooling replicates within sample groups
s1.sam
Peak calling
s1.peaks.bed
Downstream analysis

Figure M1. CLIP-seq processing pipeline flowchart. One sample group (“s1”) of a representative CLIP-seq experiment with three replicates (“r1-3”) is displayed. File formats at each step of the pipeline denote the type of data contained within: fastq = raw reads, sam = aligned reads, bed = CLIP-seq peaks.

CLIP-seq binding site definition: peak calling

CLIP-seq aligned reads form broad coverage traces across the genome, with most areas experiencing low or no coverage, and some areas exhibiting tall peaks of coverage. In order to isolate these signal peaks and remove the background coverage, I developed an implementation of an empirical peak calling algorithm that utilizes iterative within-transcript permutation of aligned reads to define an empirical false discovery rate (FDR) associated with each coverage height observed within the transcript, thus allowing a minimum peak height for each transcript to be separately computed. I based this implementation on a method initially developed in the Yeo and Ule groups, but no source code was available. For this reason, the algorithm had to be reimplemented to be of use in my CLIP-seq analyses.

This algorithm first iterates over each transcript in a provided set of transcripts, for instance the refSeq transcriptome annotation. Within each transcript, the CLIP-seq reads that align to that transcript are isolated and grouped into “islands” of nonzero coverage bounded by zero coverage. The maximum coverage of each island is computed through an associative array, and the number of islands at each height in the transcript are tabulated. This table is transformed into a table of empirical p-values defined as the sum of all islands with maximum coverage as or more extreme than the given maximum coverage. Next, iterative permutations of the coordinates of the aligned reads within that transcript are generated. At each iteration, each read is assigned a random start coordinate within the transcript such that the read is still contained within the transcript and does not hang off of the end of the transcript. This is achieved by generating a pseudorandom number between zero and the last valid start position within the transcript, which is defined as the length of the transcript minus the length of the current aligned read. After all reads aligned to the transcript are thus permuted, islands of overlapping aligned reads bounded by regions of zero coverage are identified and a table of empirical p-values of the number of islands at or greater than each maximum coverage value is generated. This process is repeated iteratively, and after the last iteration, a new table of the average and standard deviation of the empirical p-value at each maximum coverage value is computed from all of the iterations of permutations. This table is then converted to an FDR table, where the FDR value for each maximum coverage is defined as the mean p-value for that height plus one standard deviation, then this quantity is divided by the observed cumulative distribution value for that height.

From this FDR table, a minimum peak height for that specific transcript is generated. Given a provided FDR cutoff of 0.001, which is also parameterized and therefore can be customized by the end-user, the minimum peak height is defined as the height in the FDR table that has an FDR value of at most 0.001. The islands of overlapping coverage in the aligned CLIP-seq reads are then called as a peak if their maximum coverage is at least the minimum peak height for that transcript. Continuing this iteration across all transcripts in the provided transcriptome annotation allows comprehensive, transcriptome-wide discovery of CLIP-seq peaks.

This implementation is available as both a standalone program called *discover_peaks* and a library of refactored subroutines in the clipseq_analysis distribution. Importantly, several computational optimizations are utilized in this algorithm.

First, to provide a balance between memory consumption and execution speed, the algorithm first discovers the chromosomes contained within the aligned CLIP-seq reads files to be searched for peaks and iterates separately over each strand of each chromosome encountered, building a data structure of aligned reads and transcripts on that strand of that chromosome. This strand-specific algorithm prevents the entirety of the transcriptome and the entirety of the aligned CLIP-seq reads from needing to be contained within memory at any given point in time, greatly decreasing the required memory and allowing execution on a commodity personal computer instead of a dedicated compute cluster node with increased memory.

Second, any transcript that does not have at least one CLIP-seq alignment is immediately discarded and never subjected to randomization. Additionally, any transcript that has CLIP-seq alignments that do not overlap each other is also discarded because a maximum coverage of 1 will always be generated by iterative permutation.

Third, an associative array of CLIP-seq coverage is constructed during the parsing of CLIP-seq reads that align to the current strand of the current chromosome as the alignments are parsed. This data structure allows rapid extraction of maximum coverage from each observed island of overlapping CLIP-seq reads. Additionally, while associative array lookup is significantly slower than array lookup given identical numbers of elements, the sparsity of CLIP-seq coverage across a given strand of a given chromosome of the human genome is such that an array of CLIP-seq coverage would contain a vast majority of zero values across its indices: a sparse array. In contrast, an associative array does not instantiate key-value pairs at all for nucleotides that had no CLIP-seq coverage, and the resulting data structure consumes much less memory than an analogous array implementation.

Finally, significant refactoring of subroutines and variables between this and other programs in the clipseq_analysis distribution was achieved by migrating code into modules (Table M1). In addition to the benefits of code reuse, this allows unit and integration testing to cover multiple specific programs by testing modular subroutines in addition to high-level subroutines defined in specific programs that import those modules. Documentation for low-level subroutines is confined to modules, while documentation in programs focuses on higher-level logic and algorithm descriptions. This modularization of refactored subroutines also aids future enhancements to the code via a version-controlled repository.

Program name	Program type	Description
Clipseq.pm	Perl module	Library of subroutines for I/O, permutations, and data structures for genomic coverage, empirical p-values, and FDR
Z_score.pm	Perl module	Library of subroutines for sequence extraction, *k*mer counting, and data structures for motif enrichment analysis
discover_peaks	Perl program	Implementation of the empirical randomization peak discovery algorithm of Yeo and Ule
compute_z_scores	Perl program	Motif enrichment algorithm focused on *k*mers within CLIP-seq peaks
binding_mapper	Perl program	Prints plottable datafiles containing total, average, or normalized complexity of CLIP-seq binding within and around provided intervals (e.g. exons)
centric_binding_mapper	Perl program	Similar to binding_mapper, but computes binding relative to center of provided intervals, useful for miRNA hairpins e.g.
compute_randomized_overlap	Perl program	Computes Z-scores for overlap between two provided sets of CLIP-seq peaks using within-transcript randomization

Table M1. The clipseq_analysis distribution. Programs and modules (libraries) developed for computational analysis of CLIP-seq datasets. Documentation, unit and integration testing, and version control are provided in the distribution.

To establish version control and provide source code for future users, I initialized a git repository of this and other CLIP-seq programs, described above and below, and hosted this repository as a public repository on GitHub, the world's largest webserver for git repositories. This repository allows users to download the most recent version of the distribution, which contains specific programs tailored to individual tasks like CLIP-seq peak calling as well as modules of refactored subroutines, embedded documentation, unit and integration tests, and an automated installation mechanism specific to Perl distributions. The programs contained within this distribution have no dependencies outside of the Perl core, although bedtools, the genome arithmetic distribution, can be utilized to increase execution speed. Additionally, adapter removal and alignment functionalities are not provided by the clipseq_analysis distribution as well-established tools such as cutadapt and bowtie are already widely distributed. The clipseq_analysis distribution is available at https://github.com/bryketos/clipseq_analysis for download and collaborative development.

CLIP-seq motif enrichment analysis

To identify enriched motifs within CLIP-seq peaks, I implemented another empirical algorithm, *compute_z_scores*, that relies upon iterative within-transcript permutation and has found use in analysis of CLIP-seq motifs in prior studies. In this motif enrichment strategy, CLIP-seq peaks are first converted to FASTA files containing genomic sequences. Sequences are extracted from these FASTA files and a table of each *k*mer is constructed, relating the frequency of each *k*mer within the dataset to the total number of *k*mers in the entire sequence space within that dataset, in this case the number of sliding substrings of length *k* within all CLIP-seq peaks. Next, iterative permutation is performed exactly as for CLIP-seq aligned reads in the case of the CLIP-seq peak caller, except CLIP-seq peaks are permuted within transcripts, and an additional requirement is enforced that a permuted CLIP-seq peak cannot overlap a previously permuted CLIP-seq peak within that transcript at that iteration. This requirement is necessary to preserve the legitimacy of the permuted peaks, because the original CLIP-seq peaks cannot, by definition, overlap one another.

Within each iteration, at the end of permutation of CLIP-seq peaks within each transcript, coordinates for the permuted CLIP-seq peaks are written to a BED file. At the end of iterations, BED files are converted to FASTA. A sequence table of *k*mer frequencies for each permuted set of binding sites is then extracted, and this table is converted to a mean and standard deviation for each *k*mer across all sets of permuted CLIP-seq peaks. Finally, for each *k*mer encountered within the original CLIP-seq peaks, the frequency of the *k*mer is utilized to report the z-score: the number of standard deviations above or below the mean frequency among permuted CLIP-seq peaks. A high z-score indicates strong enrichment and a negative z-score indicates depletion.

This implementation is also provided within the clipseq_analysis distribution as *compute_z_scores*, and several subroutines are shared between the CLIP-seq peak caller and the CLIP-seq motif enrichment programs.

CLIP-seq peak overlap comparison

To compare the fraction of CLIP-seq peaks that are shared between two sets of peaks, I applied the refactored permutation subroutines developed for motif enrichment analysis to a new program: *compute_randomized_overlap*. Two sets of CLIP-seq peaks are first compared by computing the fraction of peaks in the first set that overlap peaks in the second set. This is most rapidly achieved using bedtools intersect, however an alternative pure Perl implementation is also provided that uses a hierarchical genomic coverage data structure. This fraction, expressed as the union divided by the intersection, is then compared to a similar fraction achieved by comparing the first set of peaks to a permuted copy of the second set of peaks. Several optimizations are additionally implemented to speed computation. First, if a transcript contains peaks only in the first set of peaks, permutation is entirely unnecessary as it is impossible to achieve any outcome that creates an overlap. Instead of permuting the second set of peaks, the total unshared peaks are simply incremented by the number of peaks in the second set of peaks and randomization is skipped. Second, memory footprint is minimized by discarding any transcript that does not have peaks in either set of CLIP-seq peaks, therefore runtime is also reduced. Third, a strand-specific algorithm is implemented that only operates on a given strand of a given chromosome at a time, thereby reducing memory footprint at the expense of runtime, however this strategy will facilitate parallelization in the future, as no state is shared between computations on separate strands. Fourth, the hierarchical data structure of genomic coverage utilizes associative arrays instead of conventional arrays to account for the sparseness of coverage, an analogy to sparse matrix algorithms which do not instantiate matrix entries at all for zerovalued or undefined elements. Together, these optimizations increase runtime and minimize resource consumption, allowing efficient operation on a single processor. This allows a compute cluster user to conduct several comparisons concurrently.

Gene ontology analysis

To analyze enriched gene ontology (GO) terms in sets of genes of interest, the Database for Annotation, Visualization, and Integrated Discovery (DAVID) version 6.7 was utilized. Where applicable, expressed transcripts within the given cell type were provided as a background, and Biological Process (GO_BP_FAT), Molecular Function (GO_MF_FAT), and KEGG pathways were searched. P-values are reported with the multiple hypothesis testing correction of Benjamini and Hochberg.

Splicing analysis from RASL-seq data

The Fu lab at UCSC provides pre-processing of raw RASL-seq data that generates a spreadsheet of junction pairs with numbers of reads mapping to each member of a junction pair. This necessary first step eliminates the need for downstream analysts to understand the specific probe pool and barcoding strategy utilized by RASL-seq and allows researchers to provide simple analyses of junction counts to generate predictions.

To obtain significant alternative splicing predictions from RASL-seq data, junction pairs were first discarded if the average number of reads across replicates was less than 10. The remaining junction pairs were analyzed by a dependent, two-sample T-test in Microsoft Excel, generating p-values for intergroup comparisons, e.g. hnRNP L knockdown versus mock depletion sample groups in unstimulated JSL1 T cells. Significant predictions were then filtered based on two criteria: first, the p-value for the T test must be less than 0.05, demonstrating significant difference of inclusion levels between the two sample groups, and second that the inclusion level difference must be at least 10%. Inclusion levels are calculated for each alternative splicing event as the total inclusion reads divided by the sum of the inclusion and exclusion reads, generating a Percent Spliced In (PSI) value, a portable metric that allows direct comparison between experiments. Inclusion level changes between sample groups are thus reported as deltaPSI, or the PSI value of the second sample group (e.g. hnRNP L depletion in unstimulated cells) minus the PSI value of the first sample group (e.g. mock depletion in unstimulated cells). Positive deltaPSI values thus indicate an increase in inclusion upon hnRNP L depletion, which indicates potential repression of splicing by hnRNP L, and negative deltaPSI values indicate the potential for enhancement by hnRNP L.

Splicing analysis from mRNA-seq data

To analyze alternative splicing using mRNA-seq data, I utilized the Multivariate Analysis of Transcript Splicing (MATS) software. MATS is a Python pipeline that utilizes algorithms implemented in the Scipy and Numpy packages to perform Bayesian inference and estimate the significance of inclusion level changes observed in RNA-seq aligned reads. MATS internally utilizes Tophat to perform alignment, therefore raw sequence reads in FASTQ format can be provided as input. MATS has many parameters that must be fine-tuned to the specific RNA-seq dataset under analysis, and the developers of the MATS software recommended that I explore a range of different parameter values when I observed that the default parameter settings generated no statistically significant alternative splicing predictions.

To find the set of optimal parameter values for this RNA-seq experiment, the unstimulated and stimulated sample groups that were not subjected to hnRNP L depletion were utilized, and performance metrics were extracted from confusion matrices utilizing a set of 169 existing RT-PCR results performed in at least triplicates in unstimulated and stimulated cells (see Results). The optimal parameter set was defined as the set that maximized Positive Predictive Value (PPV), a measure of the fraction of significant alternative splicing predictions that are validated by RT-PCR. This set of parameters also resulted in the highest overall accuracy (ACC). The parameters thus discovered were –c 0.001, -analysis P, and the ReadsOnTargetAndJunctionCounts scoring method output files. This set of optimized parameters was then utilized to discover hnRNP L-responsive alternative splicing by comparing the hnRNP L-depleted to mock-depleted sample groups in unstimulated and stimulated conditions (separately). A third sequencing replicate was performed but the sequencing and/or library quality was too low to be used.

The output from MATS was then filtered for the significance and magnitude. The p-value for the significance of differential exon inclusion between the two sample groups under comparison must be below 0.05 and the magnitude of the differential inclusion (deltaPSI) must be greater than 10% in either direction.

Stringent union of RASL-seq and RNA-seq alternative splicing predictions

After mRNA-seq libraries were prepared, sequenced, and analyzed, the Fu lab at UCSD subsequently performed a series of RASL-seq analyses of splicing in JSL1 and CD4+ T cells. Multiple sample groups were analyzed by RASL-seq, including hnRNP L depletion samples that were generated by lentiviral transduction. Data from RASL-seq were subsequently merged with data from RNA-seq to increase discovery power.

To incorporate the significant alternative splicing predictions from both RASL-seq and MATS analyses, a variety of methods were explored. First, the union and intersection were calculated. For the union, any exon that had a prediction of significant alternative splicing in either the RASL-seq or MATS analysis were retained. This was found to be too liberal as some of the predictions were in opposite directions in the two experiments and thus should be avoided. At the other extreme, the intersection was found to be too conservative because the total number of junctions queried by RASL-seq is more than an order of magnitude lower than the total number of junctions queried by RNA-seq. To find a compromise between these two extremes, I developed a method called the “stringent union” in which the union of RASL-seq and MATS analyses is first compiled, then any prediction from one experiment that had a deltaPSI value of less than 5% in the same direction in the other experiment was discarded from further analysis. This stringent union approach provides a balance between liberal and conservative analyses: as for the optimization of MATS parameters, the gold standard RT-PCR results from stimulation-responsive alternative splicing were utilized as a guide, and indeed a high false positive rate was found to result from the union of RASL-seq and MATS, and a high false negative rate was found to result from the intersection. I therefore conclude that the stringent union successfully incorporates splicing data from both experiments where an exon is queried by both, and allows the much greater breadth-of-coverage from RNA-seq to expand the scope of predicted alternative splicing events.

Differential gene expression analysis

To analyze gene expression changes from RNA-seq alignments, counts of aligned reads per transcript in the refSeq transcriptome annotation were generated. A linear model was used to test the significance of differential gene expression between two sample groups (limma). First, read counts were normalized by library size and variance of the observed mean (voom). Transcripts with fewer than 1 read per million in at least half of sequencing libraries were discarded. Empirical Bayes fitting was used to fit the model and extract p-values. Significant gene expression changes were defined as genes with at least 1.5 log2 change in either direction with an accompanying p-value less than 0.05.

Integrating CLIP-seq with splicing predictions

An informative step in integrating CLIP-seq and splicing predictions is to generate graphs of the CLIP-seq binding signal within and around regulated exons. Regions of overrepresented or underrepresented binding can subsequently identified. In this analysis, it is important to control for peaks that are of variable height by investigating the fraction of regulated exons that have a CLIP-seq peak at each nucleotide.

To achieve this, I first extracted coordinates for 350nt intervals containing 50nt of exonic sequence and 300nt of exon-proximal intronic sequence were generated for the C1 exon 5'ss (the exon upstream of the alternative exon), both splice sites of the alternative exon, and the 3'ss of the C2 exon. Next, for each nucleotide in each of these intervals, the fraction of cassettes containing a CLIP-seq peak at that position was computed. Output files relating the fraction of intervals with at least one CLIP-seq peak at each position were then generated. Plots of the fraction-bound at each position were then generated.

Combined with the CLIP-seq peak caller described above, this analysis provides a means of describing RNA binding protein occupancy patterns within and around regulated exons. Source code is available in the clipseq_analysis distribution.

Modular splice site scoring

The maxEntScan algorithm was developed to provide a quantitative splice site score separately for 5’ and 3’ splice sites based on the maximum entropy of splice site-snRNA interaction. I refactored the MaxEntScan source code available on the developer’s website into a self-testing, self-documenting, and self-installing Perl module available on the Comprehensive Perl Archive Network (CPAN).

Binomial motif enrichment analysis of exonic and periexonic intervals

In order to investigate potentially enriched sequence features within and around exons of interest, for example exons that are enhanced by hnRNP L, I developed an analysis that compares the fraction of intervals of interest containing each *k*mer to the fraction of intervals containing that *k*mer of the same type (e.g. 300nt upstream of an exon) from all internal refSeq exons. First, the intervals upstream of the exon, within the exon, and downstream of the exon are extracted for the exons of interest. For the upstream interval, I did not extract sequences that include the 3’ss as defined by the maxEntScan algorithm, namely the 20nt of intronic sequence to the immediate 5’ of the exon of interest. For the exonic region, the first and last 3nt of the exon were not extracted for the same reason. For the downstream intronic interval, the first 6nt of the intron were not extracted. This was performed so as not to conflate splice site signals with cis-regulatory motifs that are targets of potential coregulators of splicing as my primary interest was on sequences that are not part of the core splice sites themselves.

For each of these intervals, I extracted sequences around all refSeq internal exons, namely exons that are not first or last in all transcripts in the refSeq transcriptome annotation that have at least 3 total exons. Then, the fraction of intervals from each dataset containing each *k*mer were extracted. This analysis allows comparison by the binomial test because each interval can either contain or not contain a given *k*mer. As a computational optimization, I used an associative array instead of a one-zero matrix to represent *k*mer occurrences within each interval of interest. This considerably improves execution speed as the intervals are typically short, 50-280 nucleotides, and the number of unique *k*mers contained within each interval is a small subset of the number of possible strings of length k. Additionally, some *k*mers are not encountered at all in sets of exons that contain only a few hundred intervals, and this sparsity is reflected in the uninstantiated nature of the associative array data structure, whose uninstantiated values are undefined and thus may be used in Boolean expressions, in which instance they return a False value.

An output file was then generated for each interval: upstream, exonic, and downstream, containing each *k*mer encountered in the input regions on a separate row. Columns included the fraction of input intervals that containing that *k*mer, the fraction of corresponding refSeq intervals that contained that *k*mer, and the total numbers of each. This was then utilized as input into the R statistical package, from which the exact binomial test was called and p-values and confidence intervals were extracted. Importantly, the binomial test in R was automated from a Perl program by invoking R as a subprocess via a named pipe. Instructions and input data were passed to R without the need to separately generate an R script or an R input datafile, thus removing the filesystem from the interprocess communication (IPC) entirely. In this way, the invoking program can dynamically handle errors and parse output from R by utilizing the standard error (STDERR) and standard input/output (STDIN/STDOUT) output streams. This provides an automated alternative to invoking R manually and utilizes the speed and extended numerical precision of the R statistical package for performing statistical tests such as the binomial test.

To control for multiple hypothesis testing, p-values from individual binomial tests were adjusted by the Bonferroni correction, where a new alpha level was computed as 0.05/4**k, which for hexamers (*k*mers of length 6nt) is approximately equal to 1.22e-5. All *k*mers with p-values below this corrected alpha level were then aligned together with ClustalW2 and the multiple sequence alignments were used to generate sequence logos with WebLogo version 2.8.2.

Software

Software development was performed in Emacs v23.1.1 and later. Programs, scripts, one-liners, and interactive computation utilized Perl v5.10.1 and later, Python v2.7.2 and later, R v3.0.1 and later, and platform-dependent versions of the GNU compiler collection, the Bourne-again shell, the GNU core utilities, and the standard library headers. In addition to software packages mentioned in the above methods sections and core/standard libraries, software distributions from the Comprehensive Perl Archive Network, the Python Package Index, the Enthought Python Distribution (numpy and scipy), the Comprehensive R Archive Network, the Synaptics Package Manager, and the Bioconductor Project were utilized. All software use was performed in accordance with provided licenses.

Computation was performed on the Penn Genome Frontiers Institute's compute cluster and the Penn Medicine Academic Computing Services' High Performance Compute Cluster (HPC) in addition to commodity computing using multiple distributions of the GNU/Linux system as well as mintty version 1.2.0.1 and associated Cygwin distributions.

Cell culture, cell stimulation, and hnRNP L depletion

JSL1 cells were cultured in RPMI medium with 15% heat-inactivated fetal bovine serum (FBS). Stimulation of JSL1 cells was achieved by supplementing culture medium with the phorbol ester PMA (Sigma-Aldrich) at a final concentration of 20 ng/mL.

CD4+ cells were purified from human peripheral blood mononuclear cells to a purity of at least 90% by the Human Immunology core at the University of Pennsylvania (IRB #811028). These cells were cultured in RPMI medium supplemented with 10% FBS. Stimulation of CD4+ cells was achieved by the addition of antibodies to human CD3 and CD28 (Clontech).

Protein depletion by antisense morpholino oligonucleotide (AMO) was achieved by electrotransfection of 20 million cells that are first pelleted and washed twice with serum-free RPMI medium. Cells were then resuspended in 400uL for electoporation. Control samples (mock transfection) were electroporated with no AMO and knockdown samples were electroporated with 10uL of 1nmol/uL AMO. Electroporated cells were allowed to recover in RPMI medium supplemented with fetal bovine serum for 24 hours before stimulation.

RNA extraction, RT-PCR splicing assay, and qRT-PCR gene expression assay

RNA was isolated with the RNA-bee (Tel-Test) reagent and protocol. Semiquantitative radiolabeled RT-PCR assay was carried out as described previously[37]. Briefly, reverse transcription of isolated RNA was achieved by annealing reverse primer to total RNA at 90 degrees, with the reaction subsequently cooled to 43 degrees before addition of MMLV reverse transcriptase and RT-PCR master mix containing dNTPs. Reverse transcription was incubated for 30 minutes at 43 degrees, then heated to 95 degrees for 5 minutes.

PCR was carried out with 2.5ng each of 32-P end-labeled forward primer and unlabeled forward primer and 5ng unlabeled reverse primer. A mixture of PCR cycle numbers was utilized to determine the linear range of detection for the given analyte. Primer sequences for all RT-PCR primers I designed are listed in Table M2.

Gene expression changes were assayed using qRT-PCR. Total RNA isolated as described above were reverse transcribed with a cDNA reverse transcription kit (Applied Biosystems) which uses random primers. In biological and technical triplicates, 5uL of cDNA were loaded with 20uL of SYBR green PCR master mix (Applied Biosystems) into optical plates (Applied Biosystems) and primers that span exon-exon junctions to minimize the possibility of genomic DNA amplification. For each primer pair, a standard curve of 4 serial dilutions, each of 1:5 dilution ratio, was analyzed on the same plate to enable quantitation. Additionally, a no-template control and a no-RT control were analyzed in every plate. After 40 cycles of amplification on the SDS7000 qRT-PCR thermal cycler (ABI), standard curves were inspected for linearity and PCR products were analyzed by 1.5% agarose gel electrophoresis to confirm expected amplicon size and no nonspecific amplification. Quantitation was by ABI Prism software and normalized to ACTB quantitations achieved by qRT-PCR from the same RT-PCR reactions, and gene expression changes were computed as the average of the log2 (knockdown / mock-depleted) for hnRNP L depletion and as log2 (stimulated / unstimulated) for stimulation-responsive differential gene expression analyses. qRT-PCR primer sequences are included in Table M2.

Gene	Forward primer sequence	Reverse primer sequence	Amplicon sizes	Exon number	Primer type
CDK5RAP2	CCA AAA GTT AAT TCT GGC TGA AGC AGT GAT GG	GCA AGC TGG CAA GGT CAT CAG GTG GGC	127/250	E22-E24	RT-PCR
DMD	CCC AGG CAG AGG CCA AAG TGA ATG GC	CTC CAT CGC TCT GCC CAA ATC ATC TGC C	246/278	E76-E79	RT-PCR
H2AFY	GTC CAC CAA GAC GTC CAG GTC TGC C	GCT TCT TCT GGG ATG GAG ACT TGG CC	352/459	E2-E4	RT-PCR
KRBOX4	GTT GCG AAG CCA GAT GTG ATC TTC AGG	GTT CTG GAT TCT TGA CCG CTT TCA TCC	224/268	E5-E7	RT-PCR
SIRT2	GGA CAG AGC GGT CGG TGA CAG CC	CGC TCT GCA TGT ACC GGG CCA CC	165/219	E1-E4	RT-PCR
TPD52L2	GTC ACT CTG CGC CAG GTC CTG GC	CAT GTC TCC AAG CTT CCT GCT GAT GGC	234/294	E3-E5	RT-PCR
ZNF232	GGG TGA GGG CTG TAA GTG GCG CG	CTG GTC TCA TAC TCA CAA GAC TGT TCC	238/384	E2/E3	RT-PCR
MTRR	TGT TAC ATG CCT TGA AGT GAT GAG GAG G	GCC GGG CTC CAA GCT CTT GAA GTC G	387/233	E2-E4	RT-PCR
DOCK7	GGA GGA TCA GTG CAT TAT GCC ACA ATG GC	TTG ACG TCT CTG TGT GCG AAG ACA TAC G	295/205	E22-E24	RT-PCR

MARS	CGA AAT GAG ACT GTT CGT GAG TGA TGG C	GAT CTT GCA GTA ATG GGT ATA GGG CTC C	515/210	E1-E5	RT-PCR
PPP2R5E	TCC GTC AGA AAA GCC AGA CAG AAG AGG	CCC TTG AGG CAT CGT GGC CAC ACT T	292/195	E2-E4	RT-PCR
ITGA6	GAC TGT AGC TCA GTA TTC GGG AGT ACC	AAA TCA GTC CTC AGG GAT TGA GCA GGC	475/345	E24-E26	RT-PCR
FYN	AGA GAG CTG CAG GTC TCT GCT GCC G	CTC GGT GAC GAT GTA GAT GGG CTC C	310/168	E9A/E9B-E10	RT-PCR
FYN-E9-B	GTT TCG CTG AAG TGT GGC TTG GTA CC	N/A	310/168	E9A/E9B-E10	RT-PCR
PABPC1	GGA ACC AAG AGA CCG AGG CCT TCC C	CCG GCT GCT GGA AGT TCA CAT ACG C	394/118	E1-E2	RT-PCR
RBMX	GGT CAT TCC AGT TCA CGT GAT GAC TAT CC	TCC ACC ATA TCC GTC ACG TGA GCT GC	228/154	E5-E7A/E7B	RT-PCR
RBMX-E7B-R	N/A	CTT TAT CTA CTG TGA ATC AAT CAG CAC TCC	228/154	E5-E7A/E7B	RT-PCR
MIF	TAC ATC GCG GTG CAC GTG GTC CCG	CTG TCC GGG CTG ATG CGC AGG CG	173	E2	RT-PCR
HSP90AB1-E6-F	GAC CAA GCC TAT TTG GAC CAG AAA CCC	GCC AAC ATG CAA AGG CTT CTC ACA CC	131	E6/I6	RT-PCR
HEXB	GTG AAG TCT TCA CTA CCA TCC AGC CC	TAC TGA ACA CTT GAC ATG TGG CTA ATG C	172		RT-PCR
APPL2	TCA CTT GAG GCC AGG AGT TCA AGA CC	CTC ACT ACA ACC TCC GCC TCC TGG G	169		RT-PCR
RC3H1	CCC ACA AAA CTC CAT GAA GAA TTA AGC C	CCA TAA ATG TGG ATT ATG ACT CTT GGG AT	140		RT-PCR
NCK2	GGA AGA ACA GCC TGA AGA AGG GCT CC	GTT TGT TCT TCC CTG ACG CTT TAA GGG	822/100	E3-5	RT-PCR
IKZF2	GGT GAA CGC CCC TTC CAC TGT AAC C	GAC AGC AGG TCT CTC AAA AGG CAC C	365/227	E5-7	RT-PCR
PAK1	GCC GAG AGG AGC TGA GCG AGC GC	GAT ATT TGA TGT CTG AAG CAA GCG GGC	385/174	E1-4	RT-PCR
ACTB	GCAAAGACCTGTACGCCAAC	AGTACTTGCGCTCAGGAGGA	144	E5-6	qPCR
EGR1	GCAGCAACAGCAGCAGCAGC	CGTTGTTCAGAGAGATGTCAGG	111	E1-2	qPCR

CD7	GATCTCCTTCCTCCTCGGGC	CCTCGTACACCACACATGCC	116	E3-4	qPCR
TNFAIP3	AACTGGTGTCGAGAAGTCCGG	AGAGACTCCAGTTGCCAGCG	189	E2-3	qPCR
B2M	GGCCTTAGCTGTGCTCGCG	CAATGTCGGATGGATGAAACCC	149	E1-2	qPCR
TAF1D	GCAGAGGATCTGGCTTCCC	GCTTCAATGATTCTTTCAGGTGG	155	E3-4	qPCR
TRIB2	CGTGCATCTGCACAGCGG	CATAGGCTTTGGTCTCACCC	150	E1-2	qPCR
CTSW	CCGCTAGAGCTGAAAGAGGC	AGGTCACTGAATGGAGTCACC	200	E2-3	qPCR
CD1C	AGCTCTTCTTCTCCCAGGTGG	GTCCAGCCATCCTGAGCCC	137	E1-2	qPCR
VGF	TGTCTCCGGCAGCCTCTTGG	AGGCTGCGCCTCAGGGCG	119	E1-2	qPCR
TRIB1	CGAGCGCGAGCATGTGTCC	TGGCAGCTGGATGTAAGGCC	115	E1-2	qPCR
SCG2	GAAGCTCGCCCGGAGAACG	TGAAATGAAGCTGCTTCAGCC	171	E1-2	qPCR
GPR84	CTTCCATTATAGAAAGAATTGAAGG	GTCACAGCCACCACCACCC	144	E1-2	qPCR
IER3	GCACCGAAAGCGCAGCCGC	CTTCAGCCATCAGGATCTGGC	137	E1-2	qPCR
LIME1	GGTGGCCGAGTATGCCCGC	CCCTGGAGTACAGGACGTCC	119	E5-6	qPCR
IL32	GCTCCTTGAACTTTTGGCCG	CGTCCTGATTCTGCATTTTGC	129	E2-4	qPCR
EGR2	AGATGAACGGAGTGGCCGG	GAAGGTCTGGTTTCTAGGTGC	122	E1-2	qPCR
IGLL1	CTCGGTCACTCTGTTCCCG	GGGTACCATCTGCCTTCCAGG	121	E3	qPCR
TMEM173	CTAGCTCCCTGCAGCTGG	CAGGCCCGCACAGTCCTCC	113	E3-4	qPCR
LYL1	GCTGCAAGAACAGTGCTGGG	GGGCAGGCGCTGGGCTGG	151	E1-2	qPCR
GZMA	CTCTCAGTTGTCGTTTCTCTCC	AGTGAGCTGCAGTCAACACCC	175	E1-2	qPCR

CLEC11A	GAGAGGGAGGCCCTGATGC	CCTGGTCCTCCTCCATCTCC	144	E1-2	qPCR
C1orf233	GATGCGCGCCCCGCCGC	GGGCCCTCGGGCAGCACC	137	E1	qPCR
TIMP1	CTTCTGGCATCCTGTTGTTGC	GTGTCCCCACGAACTTGGCC	146	E2-3	qPCR

Table M2. RT- and qRT-PCR primer sequences.

mRNA-seq Library Preparation

Illumina TRU-seq v2 paired-end high-throughput polyA mRNA sequencing libraries were prepared according to the manufacturer's protocol. Briefly, 1ug of total RNA extract as described above was diluted to 50uL with ultrapure water and mixed with 50uL of RNA purification beads (poly-dT beads provided with kit), mixed, and incubated for 5 minutes at 65 degrees. Beads were magnetically held as supernatant was aspirated, then beads were washed with 200uL of bead washing buffer. After removing the bead washing buffer, mRNA was eluted from the beads with 50uL of elution buffer in a 2 minute 80 degree incubation followed by bead extraction with provided bead-binding buffer.

Purified mRNAs from the above polyA purification were mixed with 19.5uL of elute-prime-fragment buffer for fragmentation at 94 degrees for 4 minutes to generate an expected median fragment size of 160nt as per the manufacturer's protocol.

Fragmented mRNAs were then subjected to first-strand synthesis using the supplied first strand master mix with an incubation at 25 degrees for 10 minutes, 42 degrees for 50 minutes, and then 70 degrees for 15 minutes. Second strand synthesis was achieved by incubation at 16 degrees for 1 hour in the presence of second strand master mix. Products were then purified with Ampure beads at room temperature for 15 minutes before magnetic stationing of beads, discarding of supernatant, washing with 200uL of fresh 80% EtOH two times, and subsequent resuspension of cDNA products with 52.5uL of resuspension buffer per sample.

cDNA ends were repaired with the 40uL of end repair mix at 30 degrees for 10 minutes before Ampure bead purification as described above. Repaired cDNAs were then eluted from the beads with 17.5uL of resuspension buffer. Adenylation was performed with 12.5uL of A-tailing mix at 37 degrees for 30 minutes before proceeding immediately to adapter ligation.

Adapters were individually added to each sample according to a unique barcoding strategy in which each sample received its own barcode. This strategy allowed flexible multiplexing wherein each sample could be pooled with any other sample in the set of prepared libraries within the same lane of the flow cell. To ligate adapters onto fragmented cDNAs, separate samples of 2.5uL of adapters and 2.5uL of ligation mix were added to each cDNA sample before incubation at 30 degrees for 10 minutes. Ligated products were then purified with Ampure beads as above, this time repeating the purification twice. 20uL of purified products were aspirated and subjected to PCR fragment enrichment with 5uL of the provided PCR primer cocktail and 25uL of the provided PCR master mix for 13 cycles of amplification with the manufacturer's provided PCR cycling program.

PCR products were then purified by a single round of Ampure bead purification before 1uL of the resulting libraries were used for Bioanalyzer analysis to verify concentration of the libraries and the distribution of fragment sizes. The resulting libraries were submitted to the Next Generation Sequencing core at the University of Pennsylvania for normalization, pooling, and high-throughput sequencing on the Illumina HiSeq 2000 apparatus.

CHAPTER 1 - MAPPING TRANSCRIPTOME-WIDE hnRNP L-RNA INTERACTIONS BY COMPUTATIONAL ANALYSIS OF CLIP-seq DATA

Introduction

RNA-based gene regulation encompasses many universal processes that are essential to shaping the composition and function of the proteome in eukaryotic cells[1]. In particular, mechanisms such as alternative splicing, alternative 3′-end processing, and microRNA (miRNA)-directed processes control not only the level of expression of a transcript but also the distinct protein isoforms encoded by a given gene. Therefore, such regulatory mechanisms allow for both the expansion and the control of genetic information.

Virtually all processes of RNA-based gene regulation are controlled by the activity of a family of RNA binding proteins known as hnRNPs (heterogeneous nuclear ribonucleoproteins)[87,88,88,89,89,90,90]. Most members of the hnRNP family are ubiquitously expressed and bind to RNA substrates through RRM (RNA recognition motif) or KH (hnRNP K homology) domains[89]. Depending on the location of binding and associated proteins, hnRNPs have been shown to either enhance or repress the inclusion of particular exons, promote or inhibit splicing efficiency, alter the use of competing 3′ cleavage and polyadenylation sites, control mRNA stability, and regulate miRNA access to target genes[87,88,88,89,89,90,90]. All hnRNPs that have been well studied appear to be capable of carrying out all of these activities; therefore, the location of binding appears to be a primary determinant of whether and how a specific hnRNP controls the expression of a particular gene[87,88,88,89,89,91,91].

Given the intricacy of T cell development and function, it is not surprising that RNA-based gene regulation is increasingly recognized as a critical determinant of the growth and activity of T cells[92,93,93]. In particular, one hnRNP for which there is much evidence of a functional role in T cell biology is hnRNP L[94,95,95,96,96,97,97]. hnRNP L is a 65-kDa hnRNP family member that contains 4 RRM domains spaced throughout the length of the protein. These RRMs bind preferentially to CA repeat sequences[98], although at least one biologically relevant target sequence of hnRNP L does not conform to a strict CA repeat motif[94].

hnRNP L was first implicated in T cell biology through its role in regulating the splicing of the CD45 gene, which encodes a transmembrane phosphatase essential for T cell activation[94,95,95,96,96,97,97,99,99]. The CD45 gene contains three cassette exons (exons 4 to 6) that are independently regulated at the level of alternative splicing to control phosphatase activity. We and others have shown previously that hnRNP L is a key determinant of CD45 splicing and expression[94,95,95,97,97,100,100]. Each of the three CD45 variable exons contains an exonic splicing silencer (ESS) that is constitutively bound by hnRNP L[96,101,101]. The binding of hnRNP L to these ESSs directly induces skipping of these exons both in vivo and in vitro[44,44,94,95,95,96,96].

Recent investigation of the in vivo consequences of hnRNP L ablation in mouse thymocytes revealed a broad impact on thymic cellularity, T cell development, and the egress of mature T cells to the periphery[97]. The effect of hnRNP L on CD45 splicing may account for some of the T cell development phenotypes observed; however, dysregulation of CD45 splicing is not sufficient to explain all of the functional defects[102,103,103]. Therefore, the phenotypes of hnRNP L-deficient mice suggest that hnRNP L mediates a broad range of yet unidentified RNA-regulatory events critical to T cell development and function.

Here we have used in vivo cross-linking and immunoprecipitation (CLIP)[55,104,104] to comprehensively identify the spectrum of hnRNP L targets within the transcriptome of human peripheral CD4+ T cells. In agreement with the idea that the primary role of hnRNP L in T cells is the regulation of alternative splicing, we observe extensive hnRNP L RNA interactions in the introns of protein-coding genes. While a subset of hnRNP L binding profiles may differ in different cell states, we find significant overlap between the hnRNP L binding profiles in the two primary functional states of CD4+ cells (resting and activated), as well as between those in primary CD4+ cells and JSL1 Jurkat cells, a common T cell model cell line. Such an overlap suggests a broadly conserved role for hnRNP L in T cell physiology. Importantly, we use the conserved binding sites for hnRNP L to identify several hnRNP L-regulated alternative splicing events in genes known to impact T cell development and function, and we demonstrate that 5′ splice site (5′ss) strength is a strong predictor of hnRNP L-regulated exons. Together, our data greatly expand the understanding of the cellular activity of hnRNP L, provide a transcriptome-wide profile of hnRNP L RNA interactions in human T cells, and identify hnRNP L-dependent splicing regulation of cellular pathways as critical for T cell development and immune function.

Results

hnRNP L has been well documented to control the splicing of the *CD45* gene in both mouse and human T cells[94,95,95,96,96,97,97]. However, the dramatic developmental defect observed in hnRNP L-deficient thymocytes, together with the high abundance of this protein in T cells, suggests that hnRNP L controls the expression of a large set of functionally important genes. Therefore, to begin to understand the physiological impact of hnRNP L on T cell function, I worked with a postdoctoral fellow in the lab, Dr. Ganesh Shankarling, to map the *in vivo* association of hnRNP L with mRNAs and pre-mRNAs in primary human T cells using crosslinking and immunoprecipitation followed by high-throughput sequencing[55,104,104]. Additionally, CLIP was performed in parallel in JSL1 cells, a monoclonal Jurkat T cell line that is a model for primary T cells[94,95,95,97,97,105,105].

All previous studies of in T cells have shown hnRNP L to function similarly in resting and activated cell states, with no data suggesting a widespread change in the binding specificity of this protein in response to T cell stimulation[54,54,95]. Nevertheless, since our goal is to understand the role of hnRNP L in promoting T cell function, Ganesh Shankarling, performed CLIP in parallel in resting (unstimulated) cells and cells activated through the T cell receptor, since these two cell conditions represent critical states of T cell physiology. Briefly, purified CD4$^+$ T cells were obtained from three healthy donors. For each donor, half the cells were stimulated in culture with antibodies against CD3 and CD28 (T cell receptor and coreceptor), while the other half were maintained in medium alone. Direct protein-RNA interactions were fixed in living cells by treatment with UV light, which induces covalent cross-links between proteins and the RNAs to which they are directly bound[55]. Cells were then lysed; RNA was fragmented to a size range of 30 to 110nt; and hnRNP L RNA complexes were stringently purified using a well-described antibody to endogenous human hnRNP L (figure 1.1). The efficiency of the immunoprecipitation and the consistency of hnRNP L expression in resting and stimulated CD4$^+$ T cells are shown in Fig. 2a. Following isolation of the hnRNP L RNA complexes from cells, RNAs were released from the protein, tagged with RNA linkers, and subjected to high-throughput sequencing.

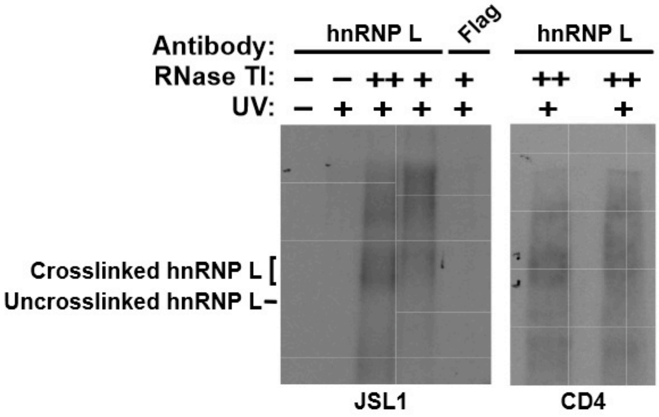


Figure 1.1. Autoradiograms of hnRNP L-RNA complexes isolated for sequencing. Representative autoradiogram from the CLIP procedure conducted from JSL1 cells (left panel) and CD4+ cells (right panel). Cells were subjected to UV crosslinking, digested with varying amounts of RNase T1, and subjected to immunoprecipitation using anti-hnRNP L or control (FLAG) antibodies. The immunoprecipitated RNA-protein complexes were resolved on 10% bis-tris NuPAGE gels. The brackets denote hnRNP L RNA-protein complexes containing RNAs of ~30-110 nucleotides. The line denotes the point of migration of the uncrosslinked hnRNP L protein. RNA-protein complexes were excised from the gel for further processing. (Figure courtesy of Ganesh Shankarling.)

hnRNP L RNA interaction profiles in T cells

Dr. Shankarling obtained a total of ~200 million reads from the 3 pools of unstimulated CD4$^+$ cells and ~100 million reads from the stimulated samples (Fig. 2b), which I proceeded to analyze with computational genomics approaches. In each case, more than 80% of reads mapped unambiguously to the genome, corresponding to a final total of 13 to 15 million unique alignments (Figure 1.2b, Table 1.1). Of these unique aligned reads (i.e., “CLIP tags”), ~23% mapped within protein-coding transcripts (Figure 1.2b, refSeq alignments), 6% to established noncoding RNAs, 19% to antisense RNAs, and the remaining 51% to mitochondrial RNAs or RNAs deriving from intergenic regions of the genome (Table 1.2). Notably, the numbers of unique alignments, as well as the genomic distributions of reads, are virtually identical for the resting and stimulated samples despite the 2-fold differential in raw reads. Thus, the sequencing depth of the stimulated samples is essentially a saturating sampling of hnRNP L binding and that the increased sequencing depth from the resting samples provides little extra discovery. Of further note, the majority of intergenic alignments were typically represented isolated reads (singletons: not overlapping any other aligned read), suggesting that these are due to spurious binding events and/or background noise in the sequencing (Table 1.2).

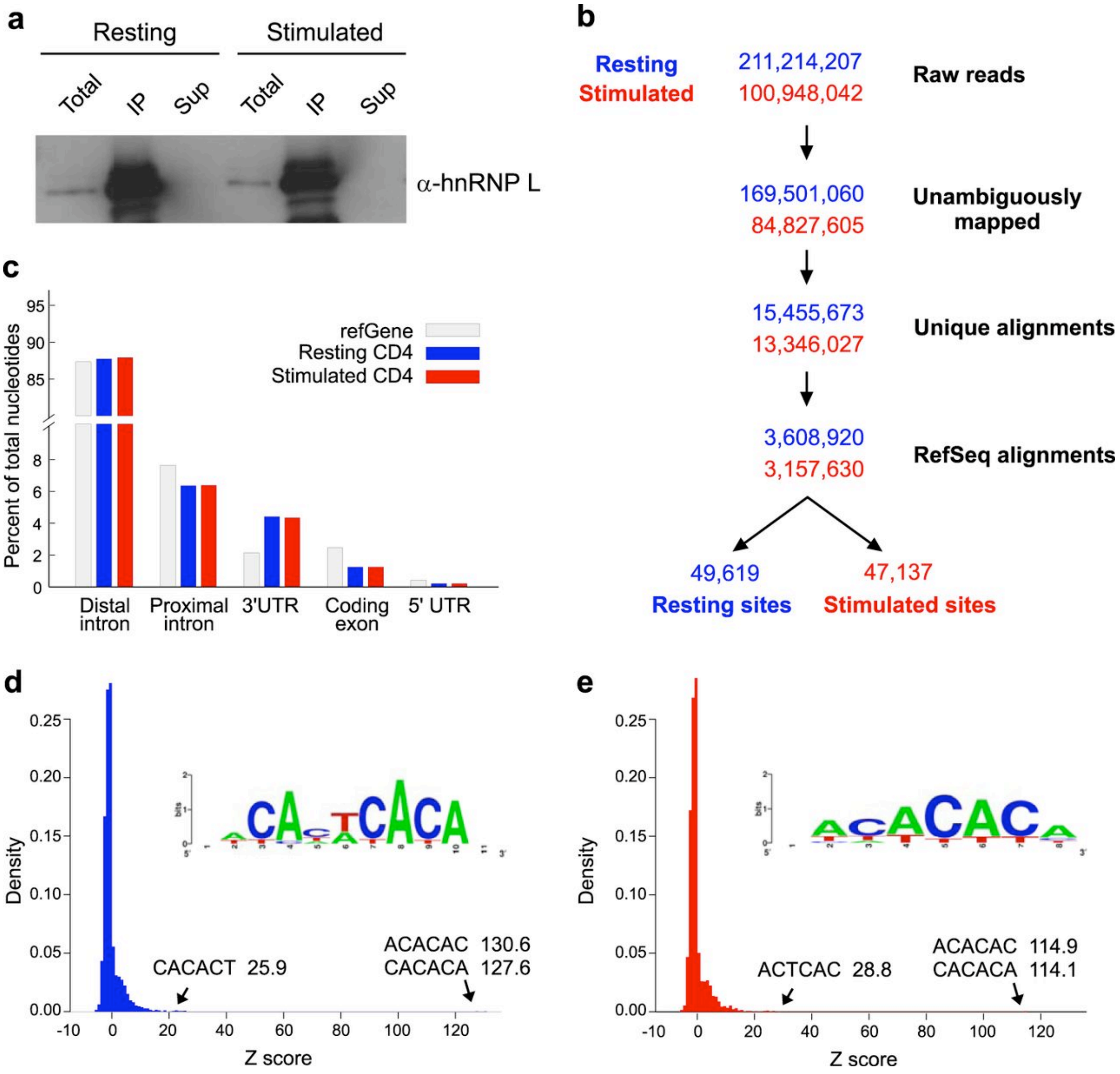
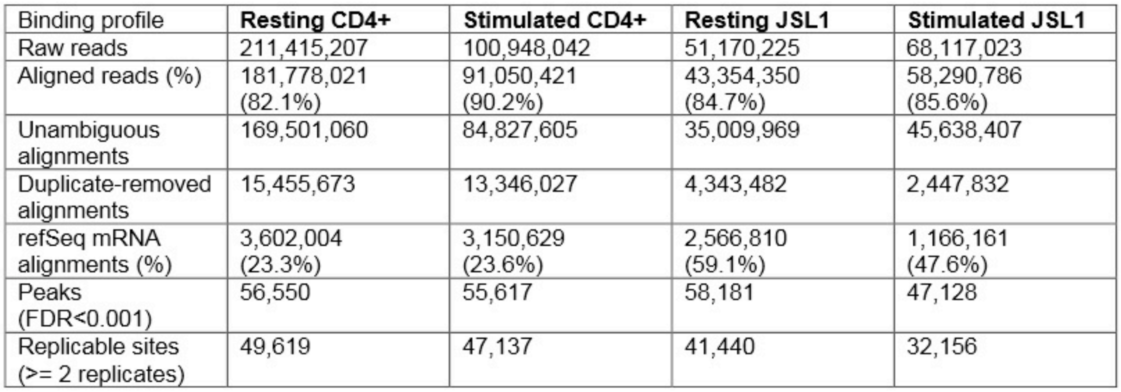


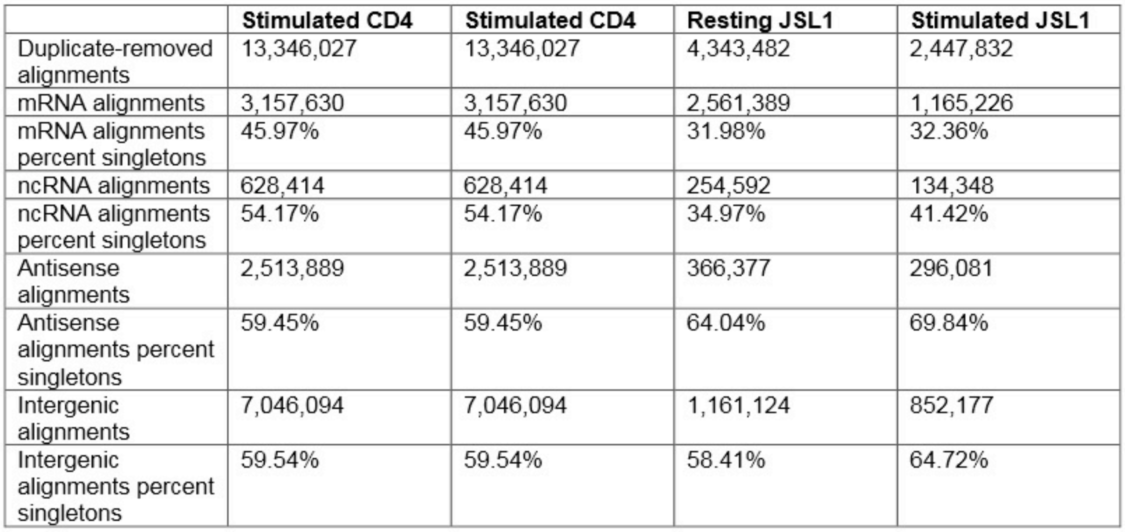
Figure 1.2. Transcriptome-wide hnRNP L-RNA interactions in primary human CD4$^+$ T cells revealed by CLIP-seq. (a) Western blot of hnRNP L expression in resting and anti-CD3- and anti-CD28-stimulated human CD4$^+$ T cells. Shown are both total expression (Total) and the efficiency of immunoprecipitation (IP) versus the protein remaining uncollected (Sup). Note that "Total" and "Sup" levels are 5% of IP levels. (b) Flow chart of analysis of CLIP-seq reads obtained from CD4$^+$ cells from three independent donors. Each sample was analyzed before and after stimulation by anti-CD3 and anti-CD28. Data from resting CD4$^+$ cells are shown in blue, while data from stimulated CD4$^+$ cells are shown in red. Numbers of reads passing key filters in the analysis are shown, including the final number of binding sites defined within refSeq transcripts in resting and stimulated human CD4$^+$ cells (see Materials and Methods and Table S1 in the supplemental material for details). (c) Distribution of hnRNP L binding sites that map to each indicated feature of RefSeq mRNAs compared to the distribution of each feature in the total refSeq transcriptome. (d and e) Z-scores for the enrichment of hexamers within binding sites in resting (d) and stimulated (e) cells were calculated by comparing observed hexamer frequencies within CLIP-defined hnRNP L binding sites to randomized binding profiles within bound

transcripts. (Insets) The top 20 hexamers were aligned to generate sequence logos. (Panel a. courtesy of Ganesh Shankarling.)



Binding profile	Resting CD4+	Stimulated CD4+	Resting JSL1	Stimulated JSL1
Raw reads	211,415,207	100,948,042	51,170,225	68,117,023
Aligned reads (%)	181,778,021 (82.1%)	91,050,421 (90.2%)	43,354,350 (84.7%)	58,290,786 (85.6%)
Unambiguous alignments	169,501,060	84,827,605	35,009,969	45,638,407
Duplicate-removed alignments	15,455,673	13,346,027	4,343,482	2,447,832
refSeq mRNA alignments (%)	3,602,004 (23.3%)	3,150,629 (23.6%)	2,566,810 (59.1%)	1,166,161 (47.6%)
Peaks (FDR<0.001)	56,550	55,617	58,181	47,128
Replicable sites (>= 2 replicates)	49,619	47,137	41,440	32,156

Table 1.1. Alignment and processing statistics for hnRNP L CLIP-seq. Total data points are listed for each major step of the CILP-seq alignment and processing pipeline, as described in Materials and Methods. "Aligned reads" refer to the total initial alignment of CLIP reads from all three samples of a particular cell type/condition to the human genome index hg19. "Unambiguous" and "duplicate-removed" are as described in Materials and Methods. Those alignments that fell within portions of the genome overlapping a refSeq mRNA were then identified. These "refSeq mRNA alignments" were then used to define binding sites ("preliminary peaks") as described in Materials and Methods. Final reported refSeq binding sites ("replicable sites") were generated by merging preliminary peaks that fell within a 50nt window of each other, and removing sites that were not supported by reads from at least 2 biologic replicates. (See table 5 for numbers of sites at different replicate stringencies.)

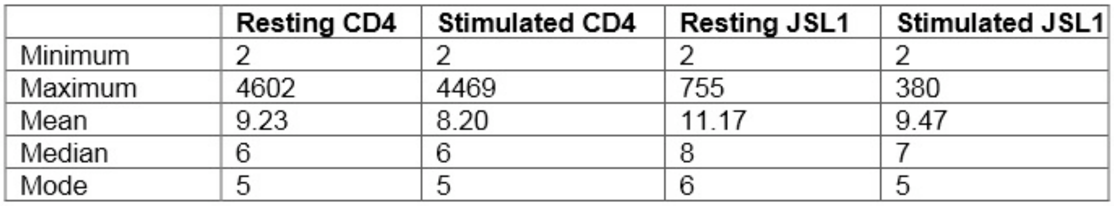


	Stimulated CD4	Stimulated CD4	Resting JSL1	Stimulated JSL1
Duplicate-removed alignments	13,346,027	13,346,027	4,343,482	2,447,832
mRNA alignments	3,157,630	3,157,630	2,561,389	1,165,226
mRNA alignments percent singletons	45.97%	45.97%	31.98%	32.36%
ncRNA alignments	628,414	628,414	254,592	134,348
ncRNA alignments percent singletons	54.17%	54.17%	34.97%	41.42%
Antisense alignments	2,513,889	2,513,889	366,377	296,081
Antisense alignments percent singletons	59.45%	59.45%	64.04%	69.84%
Intergenic alignments	7,046,094	7,046,094	1,161,124	852,177
Intergenic alignments percent singletons	59.54%	59.54%	58.41%	64.72%

Table 1.2. CLIP-seq alignments by genomic feature and percent singletons. Total CLIP-seq alignments, with duplicates removed, were assigned to one of four types of genomic feature, in decreasing order of precedence: refSeq mRNAs, refSeq ncRNA or UCSC lincRNA, antisense to refSeq mRNA, or intergenic (all remaining alignments). For each resulting pool of

unique alignments, the percentage of alignments that did not overlap any other alignment was calculated (singletons).

Because our primary interest is to understand the role of hnRNP L in shaping protein expression in T cells, I focused on those reads within protein-coding transcripts (Figure 1.2b, refSeq alignments). In order to identify a reliable binding profile of hnRNP L within transcripts, I defined binding sites empirically, using an iterative permutation algorithm similar to published methods that accounts for transcript length and sequencing depth-of-coverage by comparing observed CLIP-seq alignment distributions to those expected by random chance[106] (see Materials and Methods). To identify sites of reproducible hnRNP L RNA interaction, I required that a binding site be represented in at least two of three biological replicates. By this criterion I identified, in total, 49,619 sites of hnRNP L binding in resting CD4$^+$ cells and 47,137 in anti-CD3- and anti-CD28-stimulated cells (Figure 1.2b). Importantly, the overlap between biological samples was high: ~85% of total peaks met the requirement of being present in at least two of the replicates. Moreover, on average, each site was supported by 8 to 12 reads, although a subset of sites were supported by many more (Table 1.3).



	Resting CD4	Stimulated CD4	Resting JSL1	Stimulated JSL1
Minimum	2	2	2	2
Maximum	4602	4469	755	380
Mean	9.23	8.20	11.17	9.47
Median	6	6	8	7
Mode	5	5	6	5

Table 1.3. Read statistics for binding sites defined from hnRNP L CLIP-seq. Minimum, maximum, mean, median, and mode number of reads comprising the binding sites defined in table S1 for each experimental condition.

As expected from general predictions of hnRNP function in pre-mRNA splicing, the majority of the binding sites I identified occur within proximal (within 300nt of an exon) and distal intronic regions (Figure 1.2c). Furthermore, hnRNP L binding sites are depleted within coding exons but are enriched in 3′ UTR exons (Figure 1.2c), in agreement with previously identified roles for hnRNP L in the regulation of 3′-end processing and the modulation of miRNA regulation[54,54,90]. Finally, hexamer enrichment analysis revealed a strong preference for CA repeat elements, as evidenced both in the 2 most enriched hexamers and by multiple sequence alignment of the top 20 enriched hexamers (Figure 1.2d and e). Such a bias toward CA repeats is anticipated from previous biochemical studies of the binding specificity of hnRNP L[98]. In sum, the concurrence of the locations and sequence bias of the CLIP-identified hnRNP L binding sites with those from previous studies, together with the presence of sites of known hnRNP L RNA regulatory interactions within CLIP-derived binding profiles (see below), provides confidence that I have reliably identified major binding sites of hnRNP L across the transcriptome of CD4$^+$ T cells.

In order to correlate our findings in primary CD4$^+$ cells to Jurkat cells and to determine the utility of Jurkat cells for future mechanistic studies of hnRNP L function, Ganesh prepared CLIP-seq libraries in parallel with the CD4+ libraries described above using JSL1 Jurkat cells (Figure 1.1). As with the CD4$^+$ cells, Ganesh used triplicate biological samples of JSL1 cells grown in medium alone (resting) or stimulated with the phorbol ester PMA, which mimics T cell signaling in these cells[37]. In these experiments, Ganesh collected a total of 51 million and 68 million reads from the resting and stimulated cells, respectively, from which I defined 41,440 binding sites in resting cells and 32,156 binding sites in stimulated cells by using the criteria described for CD4$^+$ cells

(Figure 1.3a). Notably, the distribution of transcript features bound by hnRNP L in JSL1 cells is similar to that in CD4$^+$ cells (Figure 1.33b). Additionally, the sequence motifs enriched within hnRNP L binding profiles are consistent both with previous experiments[98] and with the results for CD4$^+$ primary T cells (Figure 1.3c and d). Interestingly, using expression data for resting and stimulated JSL1 cells from previous studies[107], I found that there is no general correlation between the density of CLIP tags aligning to a gene and its overall expression level (Figure 1.4). This lack of correlation of CLIP detection and gene expression confirms that the abundance of CLIP tags is a true reflection of the binding preference of hnRNP L.

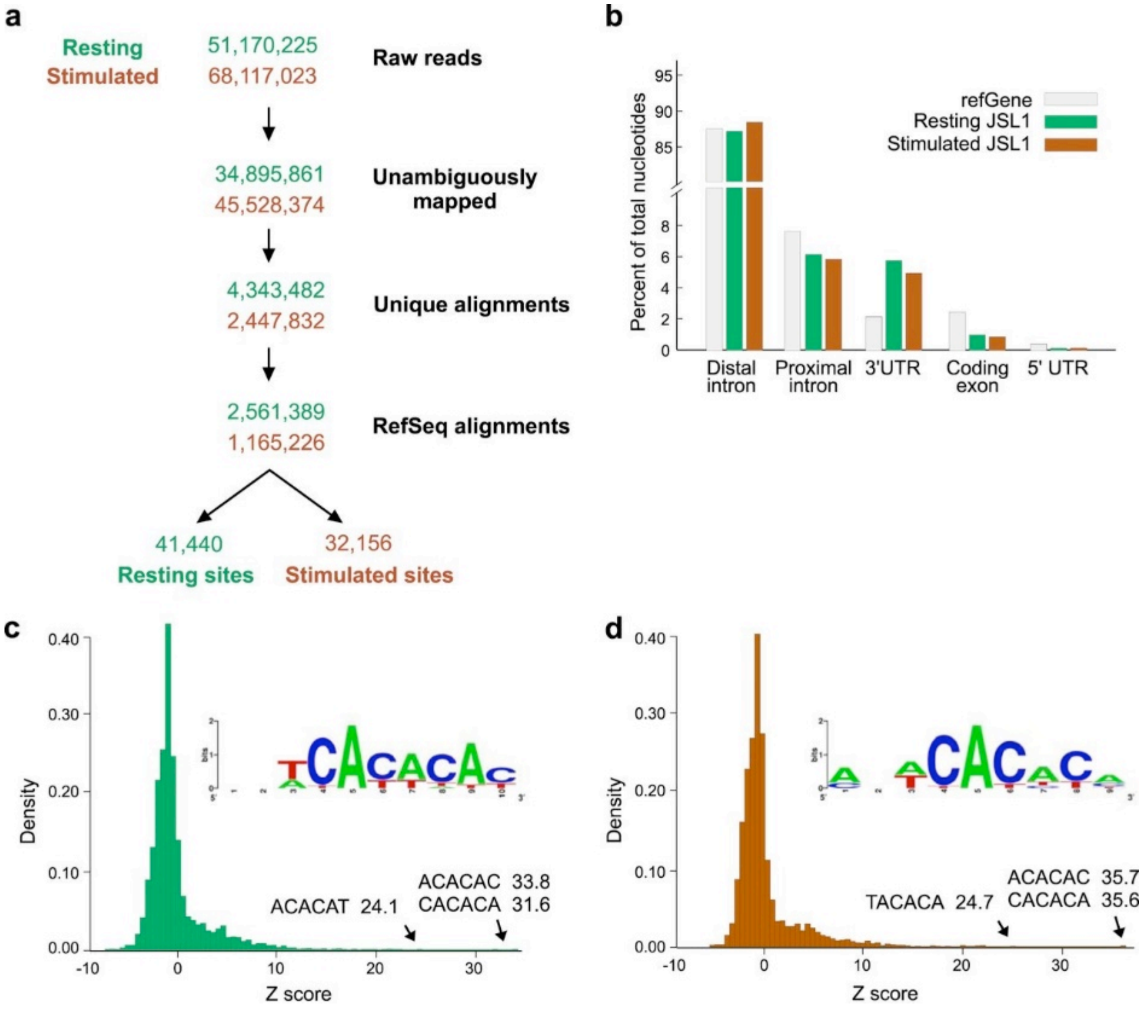


Figure 1.3. Transcriptome-wide hnRNP L-RNA interaction profiles obtained in JSL1 T cells. (a) Six biological replicates of JSL1 T cells, representing triplicate samples of resting and PMA-stimulated cells, were subjected to CLIP-seq analysis. Data were processed by a pipeline identical to that used to analyze hnRNP L binding sites in CD4+ cells. (b) Nucleotides of each type of transcript feature were enumerated within hnRNP L binding sites for both resting and stimulated conditions. Proximal introns are defined as intronic regions within 300nt of an exon. (c and d) Z-scores for the enrichment of hexamers within binding sites in resting (c) and stimulated (d) cells were calculated by comparing observed hexamer frequencies within CLIP-defined hnRNP L binding sites to randomized binding profiles within bound transcripts. (Insets) The top 20 hexamers were aligned to generate sequence logos.

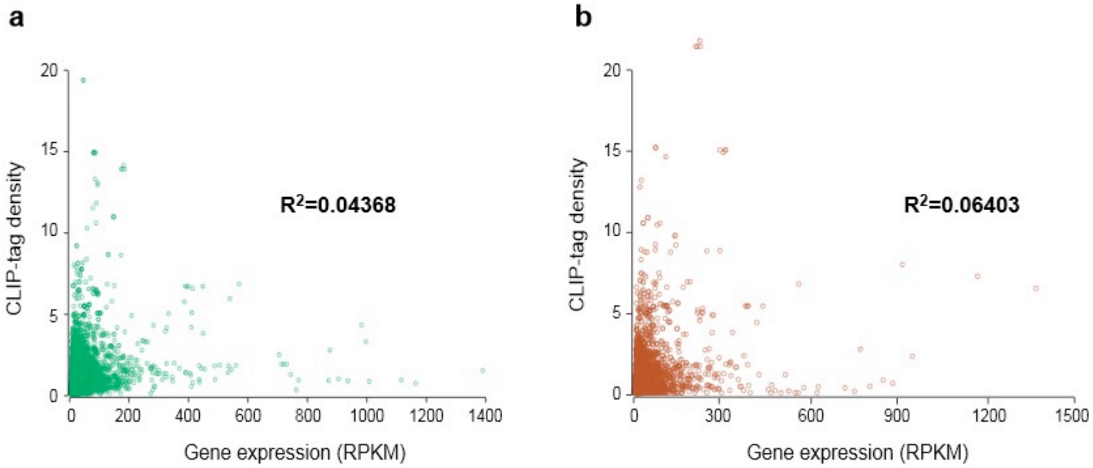


Figure 1.4. Gene expression levels do not globally correlate with CLIP tag density. CLIP tag density for resting (a) or stimulated (b) JSL1 cells was computed as RPKM by enumerating total uniquely aligned CLIP tags for each gene, then dividing by the gene length in kilobases, then dividing by the total number of uniquely aligned CLIP tags in the dataset. This normalized binding signal was compared to gene expression RPKM values obtained previously by RNA-seq (Martinez et al., 2012) and adjusted R^2 values were obtained by simple linear regression.

CLIP-seq identifies consistent binding profiles in JSL1 and CD4$^+$ T cells

Given the similarity between the sequence features and genomic annotations of the hnRNP L binding profiles obtained in CD4$^+$ and JSL1 T cells, I asked how consistent the binding of hnRNP L was between cell types and growth conditions. By calculating the percentage of total overlapping nucleotides for the two cell types, or for the two conditions, I found significantly greater overlap between the hnRNP L CLIP samples from the four cell populations than between binding profiles subjected to permutation (Figure 1.5a). For each cell type, I also investigated the number of peaks in resting cells that fell within 50nt of a peak in the corresponding stimulated cells (Figure 1.5b and c). For both CD4$^+$ and JSL1 cells, at least one-third of the peaks are shared between the resting and stimulated conditions by this logic. I defined a further ~50% of binding sites as “biased,” based on the observation of reads in both cell states, although these reads reach significance thresholds under only one of the two conditions. Indeed, at most ~20% of hnRNP L binding sites in any cell appear to be truly condition specific, in that reads are identified in only one of the growth states investigated. While this minority population of condition-specific binding events may be of interest (see below), our data clearly demonstrate that the bulk of hnRNP L binding is conserved between primary and cultured T cells as well as between resting and stimulated states. Specifically, I identified a set of 4,585 common hnRNP L binding regions that are present in all four cell types analyzed. These common regions occupy 2,460 genes in the T cell transcriptome. Importantly, among these common hnRNP L binding sites, I observed the two best characterized hnRNP L functional sites of interaction, namely, the ESS1 regulatory element in *CD45* exon 4[94] (Figure 1.5d) and an autoregulatory intronic site in *HNRNPL*[108] (Figure 1.5e).

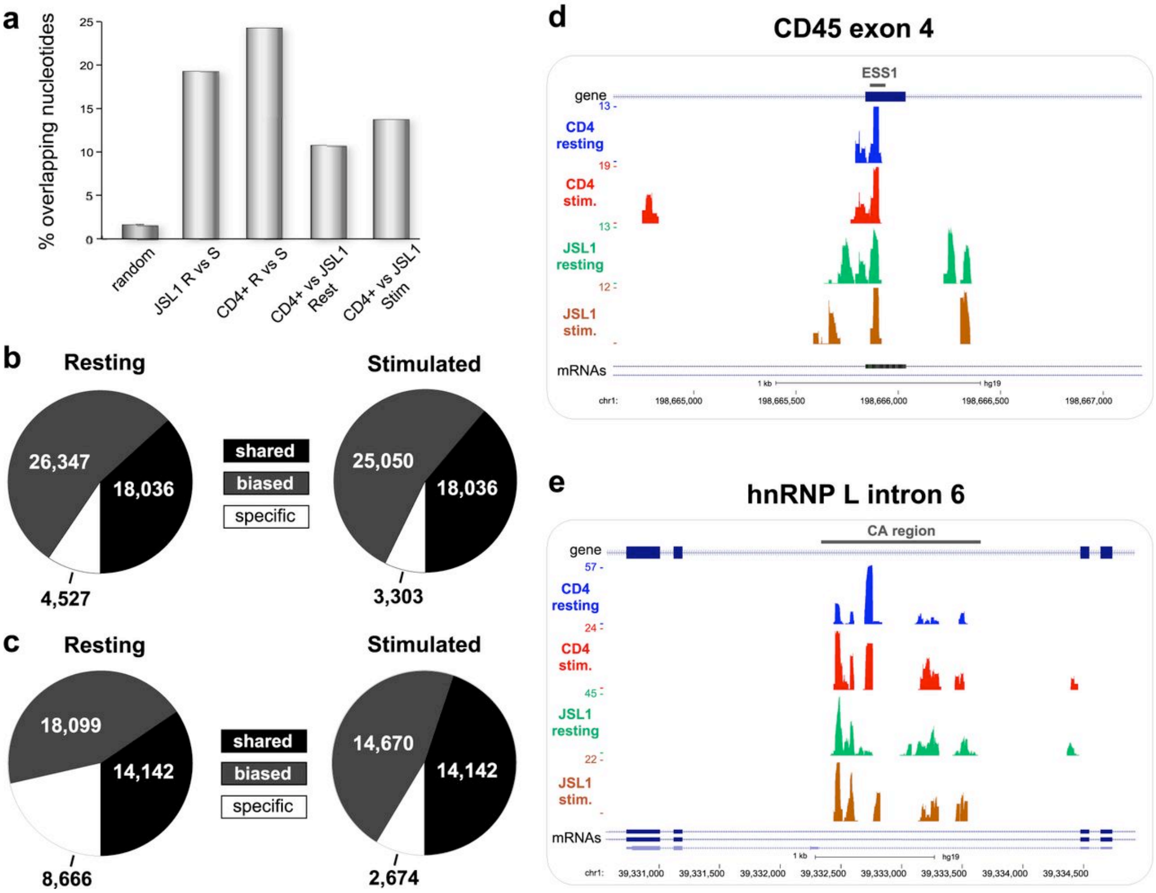
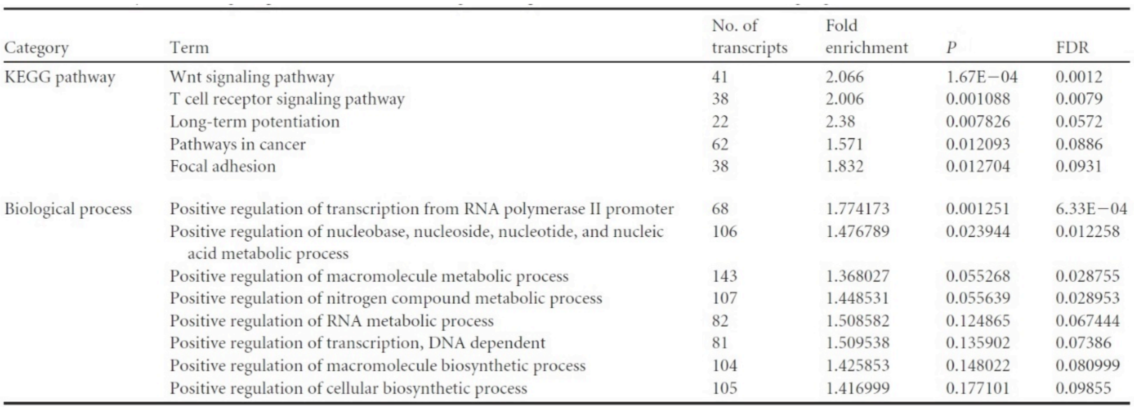


Figure 1.5. CLIP-seq identifies common hnRNP L-RNA interactions among primary and cultured T cells. (a) The percentages of overlapping nucleotides for different binding profiles were computed transcriptome-wide. The P value was ~0 for all pairwise overlaps of the data compared to the overlap of 100 permutations of resting and stimulated CD4+ binding profiles randomized within bound transcripts (control). (b and c) Total binding sites in resting and stimulated binding profiles for CD4+ (b) and JSL1 (c) cells were classified as shared, biased, or condition specific as described in Materials and Methods and in Results. (d and e) UCSC Genome Browser view of CD45 exon 4 (d) or intron 6 from HNRNPL (e), showing binding profiles from four experimental conditions. Bars above the gene schematics indicate previously identified binding sites for hnRNP L (ESS1 in CD45 and CA region in HNRNPL).

hnRNP L binds transcripts from the Wnt and TCR signaling pathways

Given the presence of known targets of hnRNP L regulatory function the common binding regions, I focused on this set of 4,585 binding events to identify new functional targets of hnRNP L and to begin to understand how this protein influences T cell development and function. First, I analyzed the KEGG pathways enriched the common target genes. Genes involved in Wnt signaling (*P* = 1.67e−4) and T cell receptor (TCR) signaling (*P* = 0.0011) are in the most overrepresented pathways among hnRNP L-bound transcripts (Table 1.4). Importantly, Wnt signaling is critical for thymic development[109], while TCR signaling is essential for both the development and the function of T cells[110]. I also analyzed biological process GO terms with DAVID, which revealed a strong enrichment of terms related to transcription and RNA-based gene regulation among common hnRNP L-bound transcripts (Table 1.4). Together, these analyses suggest that hnRNP L may broadly affect T cell function both directly, by regulating key signaling pathways, and indirectly, by altering the expression of other DNA- and RNA-binding proteins that control gene expression.



Category	Term	No. of transcripts	Fold enrichment	*P*	FDR
KEGG pathway	Wnt signaling pathway	41	2.066	1.67E−04	0.0012
	T cell receptor signaling pathway	38	2.006	0.001088	0.0079
	Long-term potentiation	22	2.38	0.007826	0.0572
	Pathways in cancer	62	1.571	0.012093	0.0886
	Focal adhesion	38	1.832	0.012704	0.0931
Biological process	Positive regulation of transcription from RNA polymerase II promoter	68	1.774173	0.001251	6.33E−04
	Positive regulation of nucleobase, nucleoside, nucleotide, and nucleic acid metabolic process	106	1.476789	0.023944	0.012258
	Positive regulation of macromolecule metabolic process	143	1.368027	0.055268	0.028755
	Positive regulation of nitrogen compound metabolic process	107	1.448531	0.055639	0.028953
	Positive regulation of RNA metabolic process	82	1.508582	0.124865	0.067444
	Positive regulation of transcription, DNA dependent	81	1.509538	0.135902	0.07386
	Positive regulation of macromolecule biosynthetic process	104	1.425853	0.148022	0.080999
	Positive regulation of cellular biosynthetic process	105	1.416999	0.177101	0.09855

Table 1.4. Transcripts with common hnRNP L binding sites were extracted from hnRNP L binding profiles of both cell types, from both cellular conditions. DAVID was used to analyze cellular pathways (KEGG) and biological processes (GOTERM_BP_FAT) overrepresented among common hnRNP L targets, at an FDR cutoff of 0.1. All significantly enriched targets are reported. P values were adjusted for multiple hypothesis testing by the Bonferroni correction.

Novel targets of hnRNP L-dependent splicing regulation

There are numerous mechanisms by which the binding of hnRNP L to a transcript may influence its expression, including regulation of transcription, stability, and efficiency of processing. Because hnRNP L is best characterized as a splicing regulatory protein, I focused on determining new targets of hnRNP L splicing regulation. I first identified several instances in which common hnRNP L binding regions (as defined above) were located in introns flanking known alternative exons, then I and others assayed the inclusion of these exons in JSL1 cells depleted of hnRNP L (Figure 1.6a) by semiquantitative radioactive RT-PCR. In agreement with the prediction from Table 1 that hnRNP L regulates genes involved in TCR signaling, T cell development, and RNA synthesis and processing, I found that hnRNP L depletion significantly alters the inclusion of known variable exons in the genes encoding the RNA-binding protein PUM2 (Figure 1.6b) and the transcription factors NFAT, BCL11A, and TCF3, which are involved in T cell developmental and activation pathways[111,112,112,113,113] (Figure 1.6c to e). I also observed hnRNP L-dependent alternative splicing of the mitogen-activated protein (MAP) kinase TAK1 and the GTPase ACAP1, which regulate NF-κB signaling upon immune signaling[114,115,115], and of CCAR1, a coactivator required for Wnt-dependent gene activation[116] (Figure 1.6f to h). For all these genes, inclusion of the variable exon either regulates overall protein expression (NFAT5 and CCAR1) or alters the domain structure of the protein (PUM2, BCL11A, TCF3, TAK1, and ACAP1) (see Discussion). Therefore, hnRNP L-regulated splicing of these genes is likely to impact T cell development and signaling, in agreement with the prediction from Table 1 and the phenotype of hnRNP L thymic deletion mice[97].

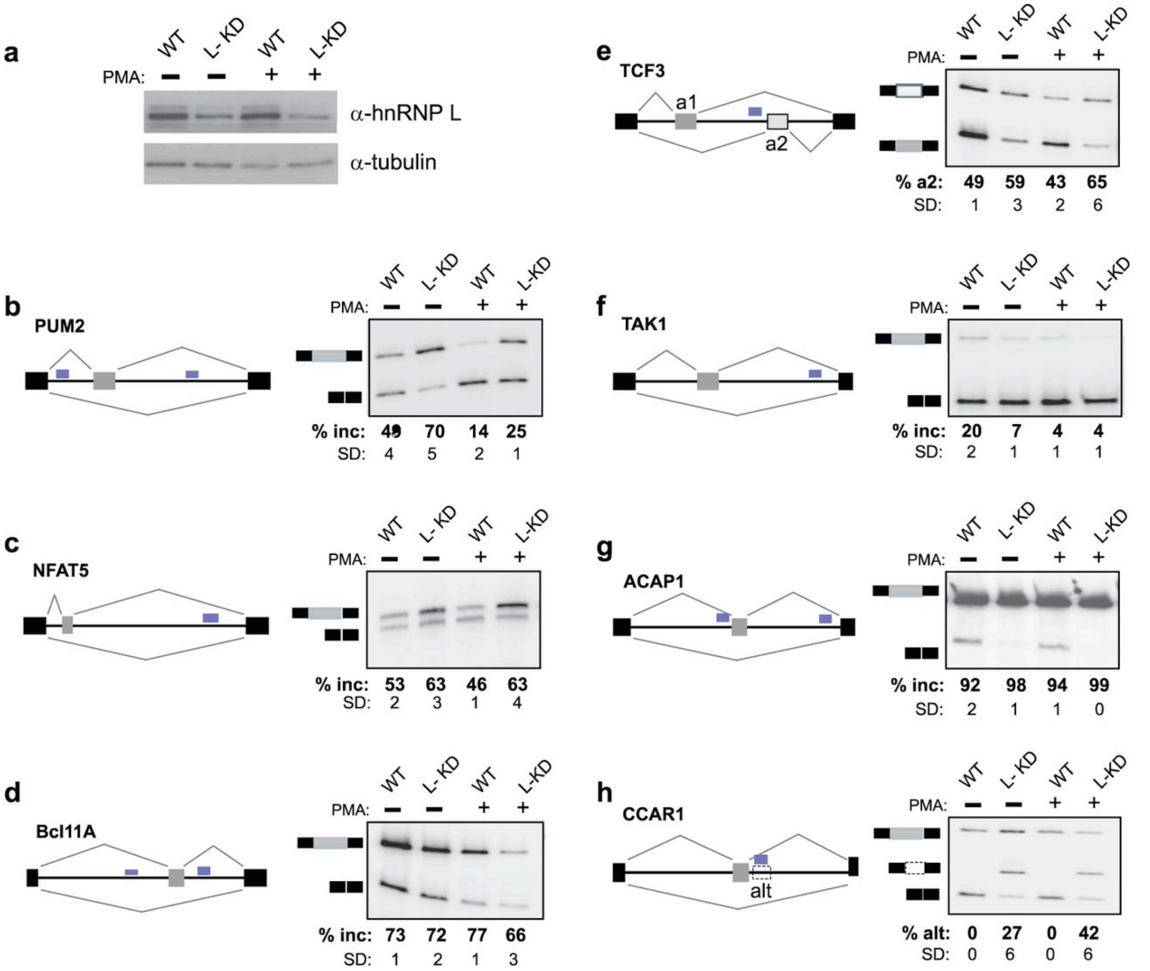


Figure 1.6. HnRNP L regulates exon inclusion of transcripts important to T cell development and signaling. (a) Lysates from wild-type cells and from cells stably transfected with a lentivirus carrying shRNA targeted to hnRNP L (L-KD) were immunoblotted using antibodies against hnRNP L or tubulin to assess loading. (b to h) Representative RT-PCR analyses of the indicated genes. Gray and black boxes represent the variable and constitutive exons, respectively, while the black line represents introns. Blue boxes represent the hnRNP L binding sites (see Fig. S3 in the supplemental material for an expanded browser view of CLIP data). The percentages of inclusion (% Inc) of the variable exons are averages for at least three independent experiments; standard deviations (SD) are shown. (e) a1 and a2 represent mutually exclusive exons. (h) The dashed box denotes the poison exon, while % alt represents the percentage of inclusion of the poison exon relative to the three isoforms.

The case of CCAR1 is particularly interesting, since Dr. Shankarling and I discovered that the binding of hnRNP L is in fact not in an intron but rather in an unannotated poison exon (i.e., an exon containing a stop codon). The fact that hnRNP L strongly represses this CCAR1 poison exon, together with our previous data on hnRNP L-mediated repression of CD45 exon 4[94], suggests that although binding of hnRNP L to exons is rare (Figure 1.2c and 1.3b), these events represent robust repressive activity of hnRNP L. Consistently, I identified ~60 genes that contain common hnRNP L binding sites within or overlapping an exon. For five of these hnRNP L-bound exons tested, the variable exon is markedly upregulated upon hnRNP L depletion (Figure 1.7). Importantly, these hnRNP L-regulated exons include those in genes encoding splicing factors (ZRANB2), cell surface receptors (SPG11, IL2RG), intracellular signaling proteins (ARAP1), and a transcription coactivator (SS18), all of which have potential roles in T cell biology.

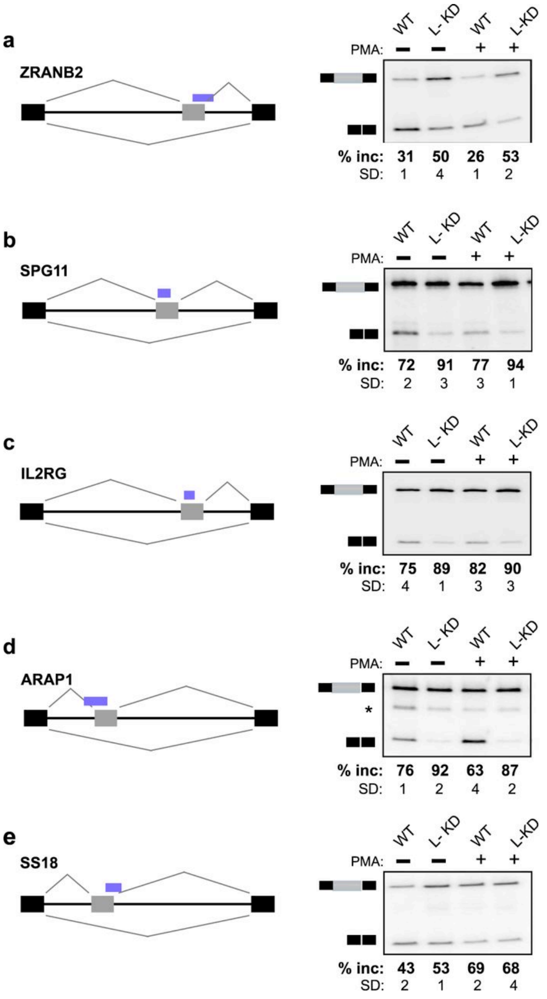


Figure 1.7. Binding of hnRNP L within exons represses exon inclusion. (a to e) Representative RT-PCR analyses of the indicated genes, as described in the legend to Fig. 6. The percentages of inclusion of the variable exons are averages from at least three independent experiments; standard deviations (SD) are shown. The asterisk in panel d indicates a nonspecific PCR product.

5′ splice site (5′ss) strength is a determinant of hnRNP L function

In addition to their functional implications, the newly identified targets of hnRNP L-mediated splicing regulation presented in Fig. 6 and 7 demonstrate the breadth of the mechanism of hnRNP L function. While exonic binding appears to correlate with hnRNP L-dependent repression (Figure 1.7), I observed no clear correlation between intron binding and hnRNP L-dependent splicing regulation. For instance, reduction of hnRNP L levels increases the inclusion of the variable exon of PUM2, whereas it decreases the inclusion of the variable exon in BCL11A, despite binding on either side of the exon in both instances. Conversely, hnRNP L appears to enhance variable exon inclusion whether it is bound to the upstream (NFAT5) or the downstream (TAK1) intron. Moreover, ~50% of exons containing or flanked by common hnRNP L binding sites that I and others in the lab tested for splicing displayed no change in inclusion in response to hnRNP L depletion. This lack of defined correlation between binding location and function is consistent both with our previous studies demonstrating that factors in addition to the location of hnRNP L binding determine its functional impact on splicing[44] and with other studies that have revealed that CLIP-defined binding sites for hnRNPs are not strong predictors of splicing regulation[117,118,118].

To determine if I could increase our ability to utilize the CLIP-defined hnRNP L binding sites to identify novel targets of hnRNP L-mediated splicing regulation, I grouped the 27 exons tested by a variety of parameters, such as intron length, position of the CLIP site, and splice site strength. Strikingly, I find that hnRNP L-dependent splicing regulation correlates best with the strength of the 5′ splice site of the alternative exon. Specifically, no alternative exons with 5′ss scores of 10 or greater (maxEntScan[119]) were regulated by hnRNP L, even when multiple common binding sites were detected close to the variable exon. In contrast, all of the hnRNP L-regulated exons had 5′ss scores less than 9.5, and 70% of the alternative exons with scores less than 9.5 exhibited hnRNP L-dependent regulation. Notably, no other single feature encompassed all of the 14 validated hnRNP L regulatory events with a positive predictive value of 70% or more.

To further validate the relevance of 5′ splice site strength, I and others in the lab tested an additional 14 exons in functionally important genes for hnRNP L-dependent splicing regulation. These exons were chosen with a range of 5′ss scores, including two in the window between 9.5 and 10 that was not represented in our initial exon set. In agreement with our predictions, I find that neither exon with a 5′ss score above 9.9 exhibits changes in splicing upon depletion of hnRNP L, while 8 of the 12 exons with 5′ss scores less than 9.9 are regulated by hnRNP L (Figure 1.8). Therefore, I conclude that 5′ss strength is an important criterion in determining regulation by hnRNP L and can be applied to CLIP-identified physical targets to increase the discovery power of functional targets of hnRNP L-regulated splicing. Importantly, using these criteria, I and others in the lab have identified a total of 20 previously unrecognized targets of hnRNP L-mediated splicing regulation, all of which are genes implicated in critical signaling and gene expression pathways in T cells, thus providing further insight into the functional role of hnRNP L in T cell biology.

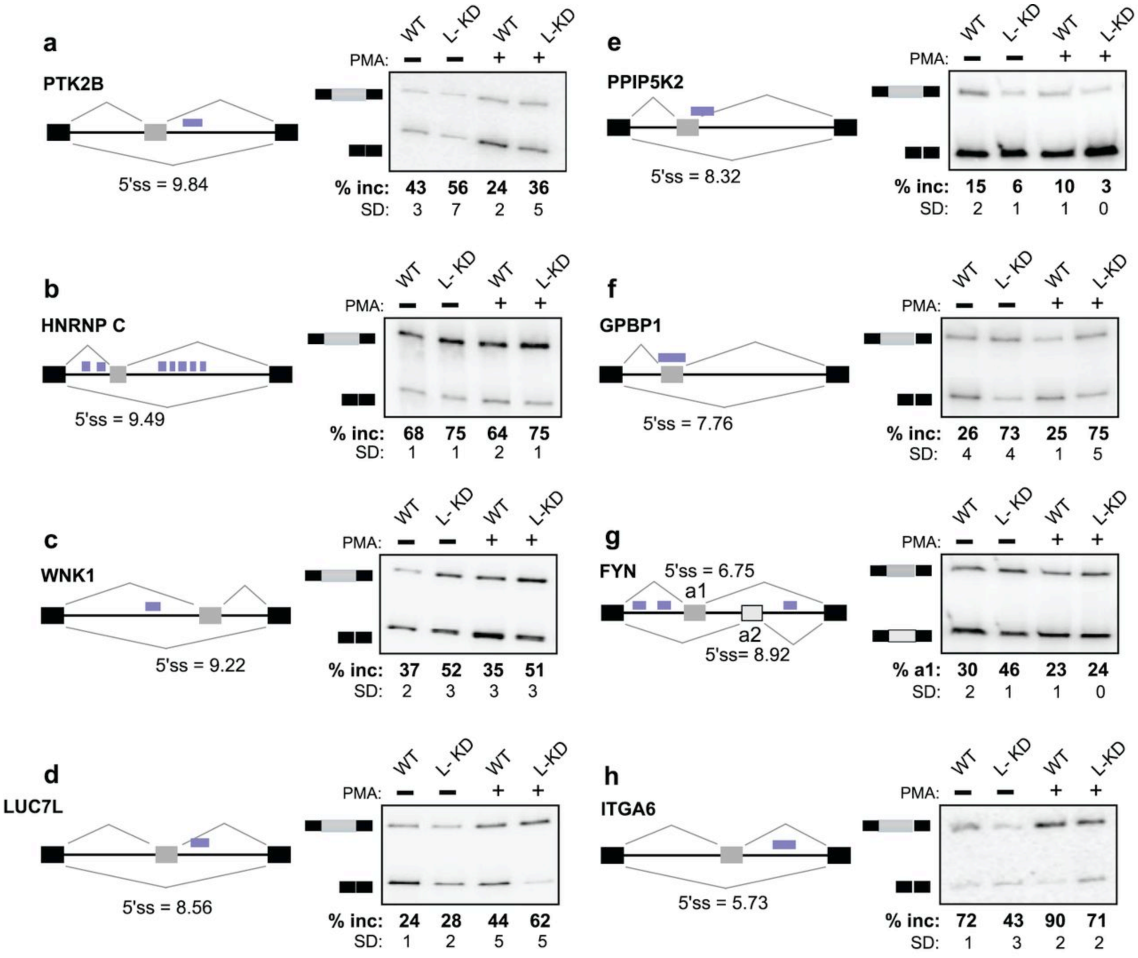


Fig 1.8. Validation of hnRNP L targets based on 5′ splice site strength. (a to h) Representative RT-PCR analysis of the indicated genes, as described in the legend to Fig. 6. 5′ss scores, as calculated by MaxEntScan, are shown for the alternative exons. The percentages of inclusion of the variable exons are averages from at least three independent experiments, and standard deviations (SD) are shown.

Condition specificity of hnRNP L binding

My analysis of the transcriptome-wide binding of hnRNP L has thus far been focused on the binding sites that are present in all four T cell populations tested, since these reveal much about the ubiquitous role of hnRNP L in T cell biology. However, as mentioned above, I did identify a subset of hnRNP L RNA interactions in both cell types that are condition specific, occurring either entirely in resting samples or entirely in stimulated samples, with no reads observed under the opposite condition (Figure 1.5b and c). To further investigate the nature of these condition-specific events, I analyzed changes in gene expression for these resting-state-specific and stimulated-state-specific binding sites, using gene expression data that our lab had obtained previously for JSL1 cells. I found that the majority of condition-specific sites are in genes whose expression does not differ significantly between resting and stimulated samples, demonstrating that the difference in association with hnRNP L is not a secondary consequence of differential gene expression (Figure 9a and b). I also found that these condition-specific binding sites maintain the general bias toward CA repeats that is seen in the common sites (Figure 1.9c and d), although this bias is less dramatic, particularly within the stimulation-specific peaks. While the possibility of direct condition-specific regulation of hnRNP L binding is not inconsistent with previous studies in T cells, there are no data to directly support such a model. Moreover, I found that the discovery of condition-specific peaks is diminished with increasing requirement for biological replication of a binding site (Table 1.5). Therefore, it remains possible that only a minor subset of the condition-specific peaks I have defined here truly represent signal-regulated changes in the binding of hnRNP L, while the majority reflect false positives due to limited local sequencing depth and biological noise.

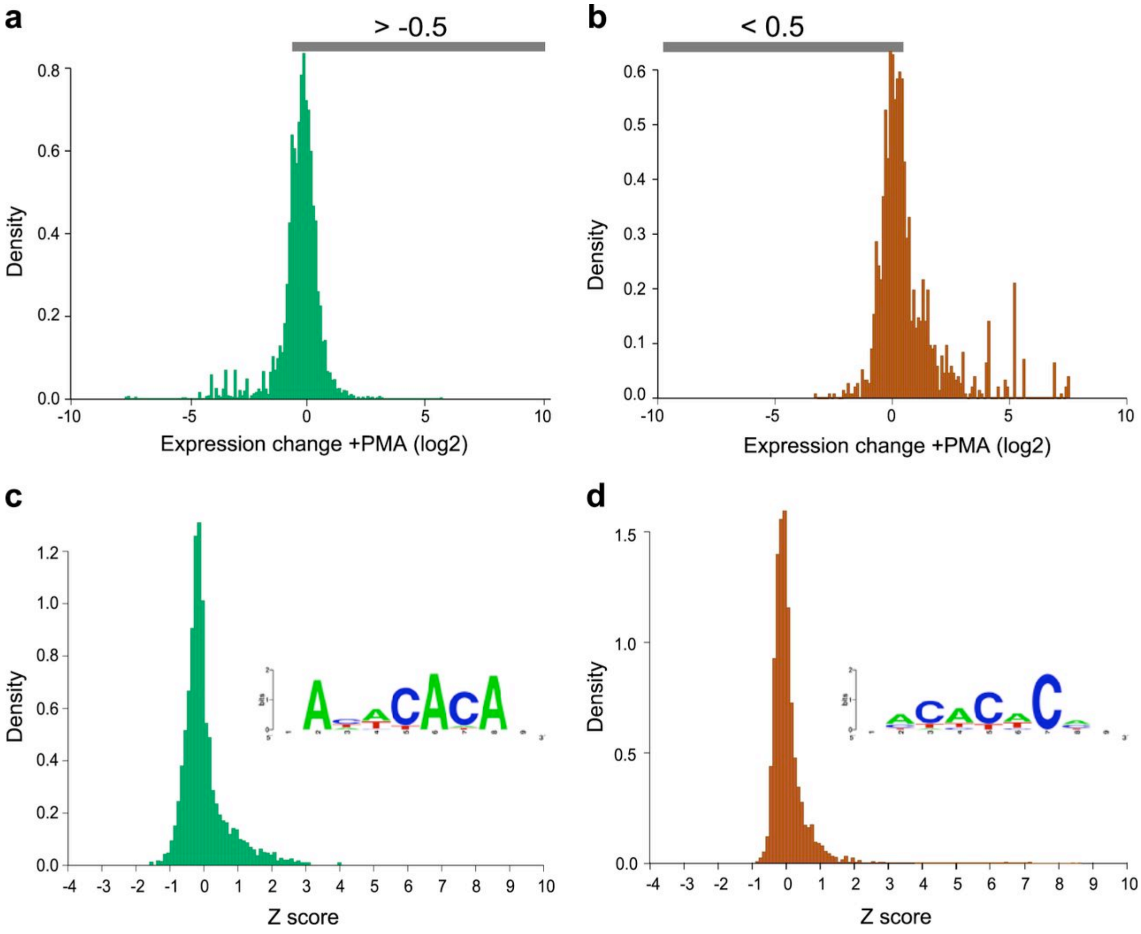
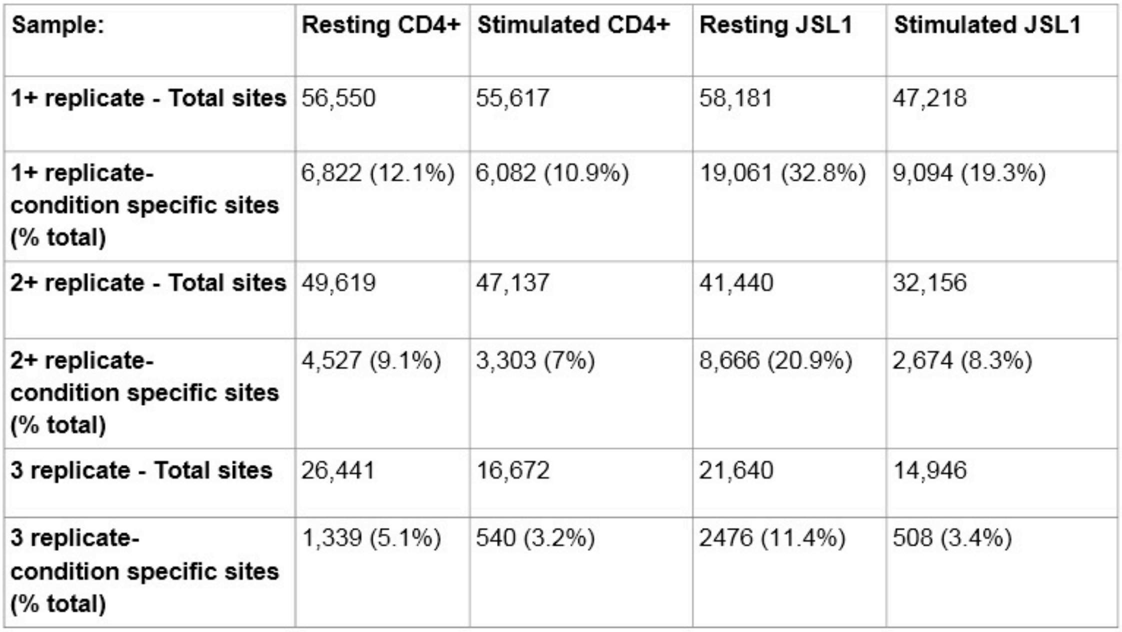


Fig 1.9. Condition-specific binding sites in JSL1 cells are not due to changes in transcript expression. (a and b) The difference in the gene expression level (expressed as the number of RNA-Seq reads per kilobase of transcript per million reads [RPKM]) between resting and stimulated JSL1 cells was calculated as $\log_2$(RPKM for stimulated cells/RPKM for resting cells) from preexisting data (24) and was plotted for all transcripts bearing resting-state-specific (a) or stimulated-state-specific (b) binding sites in JSL1 cells. (c) Hexamer enrichment for all resting-state-specific sites that are not in genes with a ≤−0.5 change in gene expression (as indicated by the gray bar in panel a). (Inset) Sequence logo generated by multiple alignment of the top 20 hexamers. (d) Hexamer enrichment for all stimulated-state-specific sites that are not in genes with a ≥0.5 change in gene expression (as indicated in panel b). (Inset) Sequence logo generated by multiple alignment of the top 20 hexamers.



Sample:	Resting CD4+	Stimulated CD4+	Resting JSL1	Stimulated JSL1
1+ replicate - Total sites	56,550	55,617	58,181	47,218
1+ replicate-condition specific sites (% total)	6,822 (12.1%)	6,082 (10.9%)	19,061 (32.8%)	9,094 (19.3%)
2+ replicate - Total sites	49,619	47,137	41,440	32,156
2+ replicate-condition specific sites (% total)	4,527 (9.1%)	3,303 (7%)	8,666 (20.9%)	2,674 (8.3%)
3 replicate - Total sites	26,441	16,672	21,640	14,946
3 replicate-condition specific sites (% total)	1,339 (5.1%)	540 (3.2%)	2476 (11.4%)	508 (3.4%)

Table 1.5. Condition-specific binding sites at various replicate stringencies. Total binding sites and condition specific sites when requiring support from 1, 2, or 3 biologic replicates. Support from two biologic replicates is the threshold used for all the data in this document.

Discussion

hnRNP L has been shown to be necessary for thymic maturation[97], suggesting that this protein plays a widespread role in shaping the proteomes of developing and mature T cells. Here we utilize CLIP-seq to identify hnRNP L binding targets within human CD4+ T cells and within a cell line commonly used for mechanistic studies of T cell biology. Importantly, the data I present here provide the first transcriptome-wide analysis of the RNA targets of hnRNP L in primary human lymphoid cells and offer novel insight into functional targets of hnRNP L in T cells.

Because the primary goal of this study was to identify novel targets of hnRNP L activity relevant to T cell function, I focused on the most conserved of the hnRNP L binding events in protein-coding genes. Using these sites, I have identified 20 new targets of hnRNP L splicing regulation. These targets include genes required for T cell signaling, such as the genes for PTK2B[120], FYN[121], NFAT5[112], and TAK1[114], genes required for T cell development (the genes for TCF3[122], Bcl11A[111,123,123], and NFAT5[112]), and the WNT signaling pathway mediator CCAR1[116]. Additional hnRNP L targets include other receptor and signaling proteins (SPG11, IL2RG, ACAP1, ARAP1, WNK1, PPIP5K2, and ITGA6), transcription factors (GPBP1, SS18), and RNA binding proteins (PUM2, ZRANB2, HNRNPC, and LUC7L), all of which may broadly influence signaling and gene expression patterns in T cells. These validated targets are consistent with the enrichment of common hnRNP L binding regions in genes involved in TCR and Wnt signaling pathways and proteins involved in transcription and RNA processing.

Of particular interest is the hnRNP L-dependent regulation of TCF3, PTK2B, and FYN, since these proteins are known to be essential for the proper development and function of T cells. In the case of FYN, we show that hnRNP L is responsible for promoting the preferential inclusion of the second mutually exclusive exon relative to the first (Figure 1.6g). Inclusion of the second exon gives rise to the FynT isoform, which is preferentially expressed in hematopoietic cells and displays altered catalytic activity relative to FynB (including the first alternative exon)[121]. Mice that specifically lack the FynT isoform have a marked defect in T cell signaling during thymic development[124]. Similarly, hnRNP L promotes the expression of the hematopoiesis-specific smaller PTK2B isoform, which exhibits a substrate profile distinct from that of the larger isoform[120]. Like FYN, PTK2B is required for appropriate T cell activation by promoting signaling through the interleukin 2 (IL-2) and LFA-1 receptors[125,126,126]. Lastly, the TCF3 gene encodes the E12 and E47 E-box transcription factors through alternative inclusion of the mutually exclusive exons[127]. These data demonstrate that hnRNP L modulates the relative expression of these factors, favoring the E12 isoform. Interestingly, ectopic overexpression of E47, as would be predicted to occur upon depletion of hnRNP L, has been shown to cause inappropriate activation of the immunoglobulin locus in pre-T cells, which would inhibit normal T cell development[128]. Therefore, while the exact contributions of FYN, PTK2B, and TCF3 misregulation to the phenotype of hnRNP L-deficient mice remain to be tested, changes in the splicing of any of these proteins upon depletion of hnRNP L in thymocytes could be sufficient to explain the developmental defects observed in vivo[41].

Finally, in addition to the identification of new targets of hnRNP L-dependent splicing regulation, I also find enrichment of 3′ UTRs among the hnRNP L binding sites, suggesting that hnRNP L may play a more widespread role in the regulation of 3′-end processing or miRNA binding than was suggested by the few instances reported previously[54,54,90]. I also observe binding of hnRNP L outside of protein-coding genes. While the majority of these interactions are isolated events, such binding may indicate additional activities of hnRNP L in the maturation of noncoding RNAs or the control of antisense transcription. In sum, the spectrum of binding events we identify here by CLIP-seq is fully consistent with known and predicted activities of hnRNP L, has identified several new targets of hnRNP L splicing regulation among genes critical for T cell development and function, and underscores the scope of the functional interactions of this abundant protein with a diverse repertoire of RNAs in T cells.

Because T cell activation by antigens is an essential component in T cell physiology, we analyzed both the binding and splicing activities of hnRNP L in both resting and activated T cell states. Proper protein expression in these two cell states is critical for maintaining appropriate functioning of the immune system. Aberrant protein expression in resting cells can lead to hyperproliferation and autoimmunity, while incorrect protein expression in activated T cells hinders the body's ability to respond to foreign antigens. Previously, our lab has identified ~180 exons for which inclusion is significantly regulated upon T cell stimulation[107]. While there is no evidence that the activity of hnRNP L is altered in response to T cell activation or directly drives these activation-induced changes in splicing, this protein has been shown to critically influence the expression of at least three of these exons (CD45 exons 4 to 6) in both resting and activated T cells. Furthermore, loss of hnRNP L-dependent repression of these exons contributes to autoimmune defects[129,130,130].

Importantly, I find common sites for hnRNP L binding around CD45 (PTPRC) exons 4 to 6 under all four cell conditions tested here (Figure 1.3d). I also observe common hnRNP L binding sites in 25 other signal-regulated genes, including 4 for which we have validated the function of hnRNP L in regulating exon inclusion in at least one cell state (the genes for TAK1, PTK2B, LUC7L, and FYN [Figure 1.4 and 1.6]). Interestingly, in three of these cases (TAK1, LUC7L, and FYN), depletion of hnRNP L is observed to influence splicing only under one cell condition, despite the fact that robust binding is observed under both conditions. Such condition-specific function was also observed for hnRNP L-dependent regulation of Bcl11A and SS18 despite the presence of common binding sites. Importantly, condition-specific effects of hnRNP L depletion are an expected result due to the combinatorial regulation of splicing. In other words, most splicing events are determined by the interplay of multiple regulatory proteins. Therefore, the requirement for any one protein is influenced by the presence or absence of other proteins. For instance, the stimulation-specific requirement for hnRNP L in repressing the LUC7L exon likely reflects the presence of a more efficient repressor protein that specifically associates with LUC7L in resting cells and compensates for the loss of hnRNP L under resting conditions. Alternatively, condition-specific effects of hnRNP L might reflect regulation of the intrinsic activity of hnRNP L upon T cell activation, although such regulation has not been described and would have to be gene specific.

Finally, in addition to the correlation of common binding sites with condition-specific function in some cases, I also detect a subset of binding sites that are apparent only in resting or stimulated T cells and cannot be explained solely by differences in the availability of transcripts. Notably, there are ~40 genes with condition-specific binding events among the previously defined signal-responsive splicing targets. While further study will be required to determine the biological relevance of these and other apparently condition specific binding sites, I note that a subset of hexamers enriched among the JSL1 stimulation-specific binding sites are distinct from the typical CA repeat element and are not enriched in the resting-state-specific or total binding site sets. Interestingly, these stimulation-specific hexamers contain motifs, such as TCT repeats and poly(C) elements, similar to those of known binding sites of other hnRNPs, such as PTB (hnRNP I) and hnRNP K and hnRNP E2, respectively[89]. Therefore, it is possible that hnRNP I, K, or E2 directs at least a subset of hnRNP L binding events in stimulated cells. I also note the possibility that stimulation of T cells results in a posttranslational modification(s) of hnRNP L that alters its binding affinity and/or specificity. While such regulation of hnRNP L binding has not been reported in T cells, at least two reports have suggested that phosphorylation of hnRNP L in other cell types can alter its ability to recognize specific RNA target sequences[131,132,132]. I emphasize, however, that less than 10% of the total binding events detected for hnRNP L appear to be condition specific, and this number decreases further with increased stringency of peak calling. Therefore, whatever mechanism(s) is at play to direct condition-specific binding of hnRNP L, the majority of hnRNP L interactions remain unaffected, underscoring the consistency of hnRNP L association with the transcriptome in both resting and activated T cells.

An inherent limitation of CLIP-seq analysis is that the method identifies physical interactions but provides no information regarding function. Therefore, a challenge in moving forward from such studies is how to identify which physical interactions are meaningful for any given function of interest. In some cases, “RNA maps” have been constructed to correlate binding location with splicing function; however, the construction of these maps requires knowledge of a large number of functional targets, so they are not suitable for *de novo* discovery. Furthermore, our lab and others have shown previously that hnRNP L can function as an enhancer or a repressor from similar locations within an exon[44,44,90], suggesting that location is not a primary determinant of hnRNP L splicing activity. Indeed, simply scoring for proximity of a conserved hnRNP L binding site to a known alternative exon provided only ~50% confidence of hnRNP L-dependent splicing.

As an alternative approach to better prediction of binding sites that correspond to splicing regulation, I scored a range of features of the first 28 test exons I investigated for hnRNP L-dependent splicing regulation and found that the strength of the 5′ss of the alternative exon was the strongest predictor of hnRNP L activity. Using this criterion, we then identified another eight targets of hnRNP L-regulated splicing, with a positive predictive value of ~70%. Interestingly, 3 of the 4 alternative exons that were not regulated by hnRNP L despite a low 5′ss score were flanked by introns that were each >10 kb long, whereas all of the hnRNP L-regulated exons were flanked by at least one intron of <9 kb. Therefore, intron length may provide additional predictive power in identifying targets of hnRNP L splicing regulation.

In addition to the predictive power of 5′ss strength, the fact that this feature correlated best with hnRNP L-regulated splicing has important mechanistic implications. Previously, our lab has shown that 5′ss strength influences the ability of hnRNP L to regulate a model exon and that at least one mechanism by which hnRNP L acts is remodeling of the interaction of the U1 snRNA with the 5′ss region[44,46,46]. Interestingly, I have identified 26 hnRNP L-bound exons within our CLIP data that have the sequence hallmarks of the U1 remodeling mechanism, including the exon in PUM2 that we have validated as strongly repressed by hnRNP L (Figure 1.4b). Therefore, these CLIP data provide further evidence of the importance of 5′ss identity in the mechanism by which hnRNP L regulates T cell biology, and they set the stage for further investigation of the determinants of hnRNP L binding and function.

CHAPTER 2 -DISCOVERY OF hnRNP L-REGULATED ALTERNATIVE SPLICING WITH RASL-seq and mRNA-seq

Introduction

We previously reported an analysis of transcriptome-wide hnRNP L-RNA interactions in cultured and primary human T cells[133]. These data provide detailed insights into the landscape of hnRNP L physical target pre-mRNAs, but the overlap between physical and functional targets is not complete: many pre-mRNAs with hnRNP L CLIP-seq peaks within and around alternative exons demonstrated no splicing changes upon hnRNP L depletion by RT-PCR validation experiments. Similarly, as hnRNP L cross-regulates other splicing factors[134] including the hnRNP L-regulated alternative splicing events in PUM2, ZRANB2, HNRNPC, and LUC7L which were discovered through our CLIP-seq analysis, the potential for indirect effects on splicing following hnRNP L depletion create a situation in which binding does not necessarily implicate splicing regulatory function nor vice versa. These dual, reciprocal caveats have motivated the development of integrative genomics approaches that combine CLIP-seq and mRNA-seq experiments to separate potentially direct splicing regulatory targets from potentially indirect targets[33,33,62,62,118,118,135].

The integrative genomics approach relies on genome-wide identification of splicing regulatory targets of the splicing factor under study. Early splicing regulatory studies utilized microarray technology, an early breakthrough in transcriptomics[79]. One of the major limitations of microarray studies is the inability to discover alternative splicing events that do not have oligonucleotide probes specifically designed to them. Our discovery of an unannotated poison exon in the CCAR1 pre-mRNA whose inclusion is regulated by hnRNP L[133] underscores the importance of *de novo* discovery capability in the analysis of hnRNP L splicing regulatory function. Technological advances in next-generation sequencing have made RNA sequencing affordable in recent years, and this technology has found application in comparative analysis of RBP-depleted and mock-depleted transcriptomes, spurring a wave of software development efforts aimed at applying statistical analysis to splicing changes observed between RNA-seq datasets[85,86,86]. These experiments typically involve depletion or overexpression of the protein under study followed by RNA extraction, RNA-seq library preparation, sequencing, alignment, and analysis of aligned reads. Statistical analysis of quantitative changes in splice junction utilization between two sample groups can identify alternative splicing events that are significantly responsive to protein depletion or overexpression, providing evidence that those splicing events could be under direct or indirect control by that protein.

I previously described the discovery of novel hnRNP L-dependent splicing regulation in pre-mRNAs encoding proteins with important roles in T cell biology[133]. The discovery of these events by CLIP-seq analysis coupled with the generally low overlap between binding and function observed in integrative genomics studies[62] suggests that there could exist a plethora of hnRNP L-responsive alternative splicing events which CLIP-seq will not reveal. Additionally, indirect regulation of splicing events, while incapable of revealing mechanistic insights into regulated alternative splicing by hnRNP L-RNA interactions, can unveil the role hnRNP L plays in an interconnected network of splicing regulators.

While recent software advances have made statistical comparisons of splicing between RNA-seq libraries possible, analyses still must be tailored to fit the specifics of the experiment. RNA-seq specific splicing analysis software such as MATS, the Multivariate Analysis of Transcript Splicing[86], provide many parameters which must be fine-tuned, including null hypothesis cutoffs, replicate composition, variance estimation, and scoring metrics. While little objective evidence exists to guide the optimization of these parameters, utilization of a Gold Standards RT-PCR dataset allows analysts to converge upon the set of parameters to software such as MATS that generates predictions in maximal agreement with previously generated RT-PCR results. To this end, our lab has previously generated a large dataset of RT-PCR validations for PMA-responsive splicing in JSL1 T cells[107]. I leverage the power of this PMA-responsive dataset to optimize an analysis of hnRNP L from RNA-seq datasets and apply these parameters to generate splicing predictions of high positive predictive value.

While RNA-seq provides high breadth-of-coverage across transcriptomes, spliced junctions represent a minority of the sequence space in aligned RNA-seq reads. For this reason, RNA-mediated oligonucleotide Annealing, Selection, and Ligation followed by next-generation sequencing (RASL-seq) was developed by Fu lab at UCSD to provide high sequencing depth at splicing junctions known to be alternatively utilized in various conditions such as development[136]. Our lab has employed RASL-seq to study regulated splicing events that respond to other treatments such as CELF2 depletion, demonstrating the utility of this approach in uncovering alternative splicing events responsive to depletion of RNA binding proteins in our JSL1 T cells. In this chapter, I combine the breadth of RNA-seq with the depth of RASL-seq to provide an additional dimension to the discovery of hnRNP L-regulated alternative splicing events, a process that overcomes the shortcomings of both experiments by combining their complementary advantages as discovery tools.

This complementary next-generation sequencing design uncovers a wide scope of hnRNP L-responsive alternative exon utilization in T cells with high rate of RT-PCR validation. Target transcripts are enriched for splicing factors, transcription factors, and epigenetic factors, but hnRNP L depletion does not induce global or subglobal differential gene expression. Finally, these data, coupled with the CLIP-seq analysis I previously reported[133] provide the foundation for integrative genomic analysis.

Results

To identify the impact of hnRNP L transcriptome-wide in pre-mRNA processing in these cells, our lab employed a complementary genomics approach to provide high depth- and breadth-of-coverage across the JSL1 transcriptome of hnRNP L-depleted or mock-depleted cells. I first performed hnRNP L depletion using an antisense morpholino oligonucleotide (AMO, see Materials and Methods), reducing hnRNP L protein levels by ~50% (figure 2.1b, Western blot). To generate high breadth-of-coverage sequencing data, I utilized paired-end mRNA sequencing to query splicing junctions from AMO-transfected or mock-transfected RNA extracts. The resulting aligned sequence read pairs provide transcriptome-wide coverage, facilitating discovery of previously unknown hnRNP L-responsive pre-mRNA splicing events.

A technician in our lab, Michael Mallory, subsequently developed JSL1 T cell sublines stably transduced with a lentivirus containing a doxycycline-inducible shRNA directed against the HNRNP L transcript. Using this distinct knockdown approach, he depleted hnRNP L in both unstimulated (resting) and PMA-stimulated cells, providing independent physiological conditions for the identification of hnRNP L-responsive pre-mRNA processing events (figure 2.1a). The lentiviral knockdown approach also depleted hnRNP L protein levels by ~50% (figure 2.1b), which, taken together with the AMO knockdown strategy, provided a robust experimental design with independent mechanisms of action.

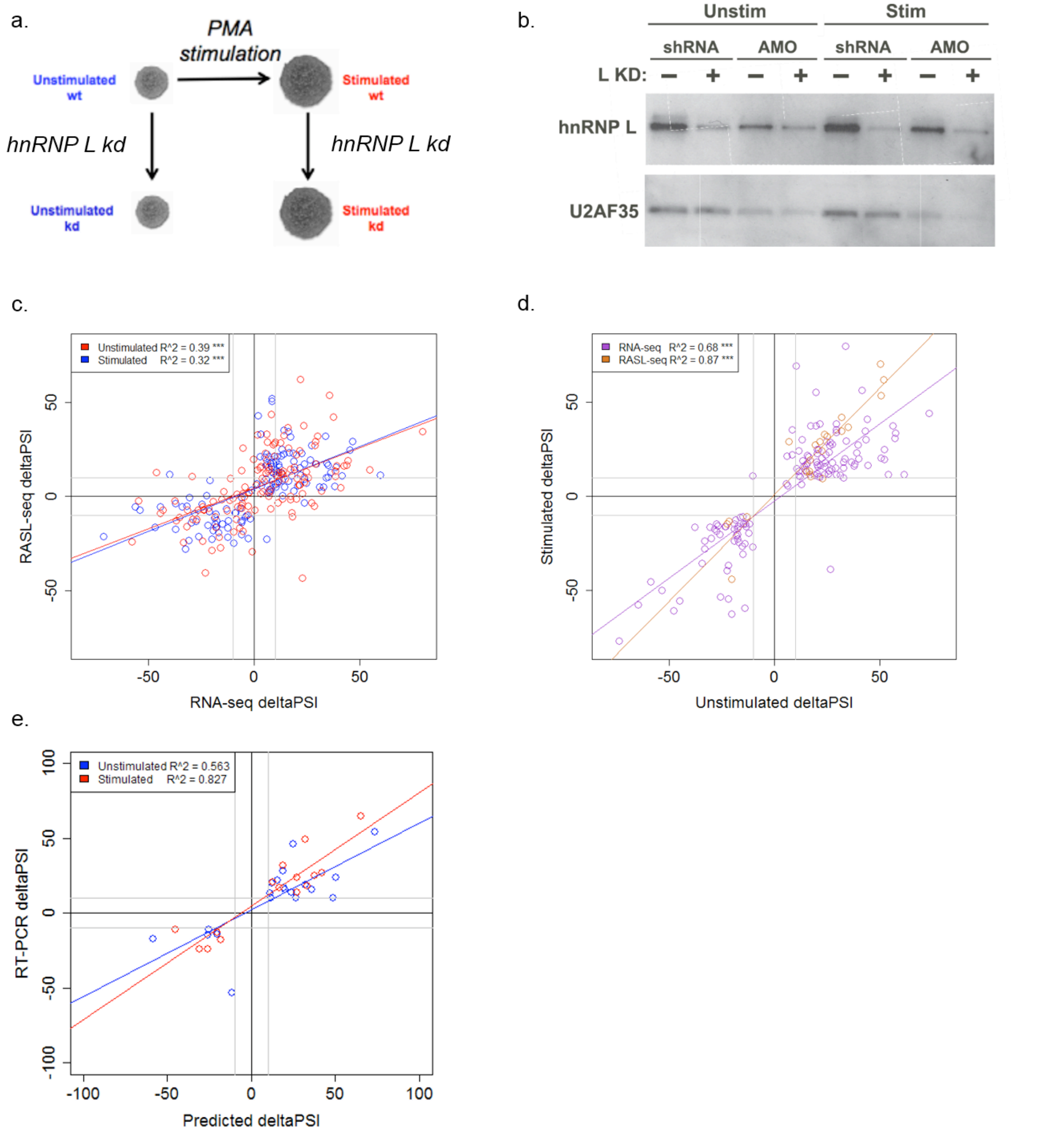


Figure 2.1. Complementary high-throughput sequencing approaches identify hnRNP L-dependent alternative splicing events in JSL1 T cells. a.) Experimental design in which unstimulated and stimulated JSL1 T cells were independently depleted of hnRNP L by prior to RNA and protein extraction. b) Western blot of hnRNP L depletion by AMO or lentiviral shRNA. c.) Regression analysis comparing cassette exon inclusion changes for significant predictions from RASL-seq and mRNA-seq. d.) Scatterplot of inclusion level changes between unstimulated and stimulated conditions by both sequencing methods. e.) RT-PCR validation of splicing predictions. (Panel b courtesy of Michael Mallory.)

To increase sequencing depth at known alternative splice junctions, we collaborated with the laboratory of Dr. Xiang-Dong Fu at UCSC, who prepared RASL-seq libraries from induced or uninduced lentiviral shRNA-transduced JSL1 cells. Taken together, the resulting datasets generated high breadth-of-coverage from the mRNA-seq aligned reads, querying over 70,000 splice junctions, and high depth-of-coverage across the splice junctions queried by RASL-seq (table 2.1).

Experiment	Total reads	Junction pairs queried	Reads per junction pair (median)	Reads per junction pair (m.a.d.)
RASL-seq	67,264,257	3,287	59	87
RNA-seq	403,942,906	70,546	13	19

Table 2.1. Sequencing depth by RASL-seq and mRNA-seq. Total reads generated for each experiment, the total count of unique junction pairs queried by analysis of aligned reads, and median with accompanying median absolute deviation (m.a.d.) are provided for each experiment. RNA-seq reads per junction pair were generated by the Tophat aligner (see Materials and Methods) and reported in the MATS output.

To analyze splicing changes from RNA-seq data, I utilized the unstimulated and stimulated conditions that were not subject to hnRNP L depletion to optimize the positive predictive value as measured by existing RT-PCR data. Specifically, the lab has previously generated 169 “gold standard” RT-PCR results for cassette exon inclusion levels in unstimulated and stimulated conditions. These gold standard RT-PCR results are performed in at least triplicate replications, and include exons that exhibit statistically significant ($p < 0.05$, T-test) and large magnitude (inclusion level change of at least 10% in either positive or negative direction) inclusion changes, as well as a cohort of negative results for exons that do not exhibit stimulation-responsive inclusion level changes. In total, there were 27 exons with stimulation-inducible exon skipping (deltaPSI <= -10), 25 exons with stimulation-inducible inclusion (deltaPSI >= 10), and 117 remaining RT-PCR results that were considered as negative (figure 2.2). Importantly, this gold standard RT-

PCR validation dataset contains a much broader set of RT-PCR results than I and others in the lab have generated for hnRNP L-responsive alternative splicing, therefore I used this dataset to optimize the computational parameters for splicing analysis from mRNA-seq data.

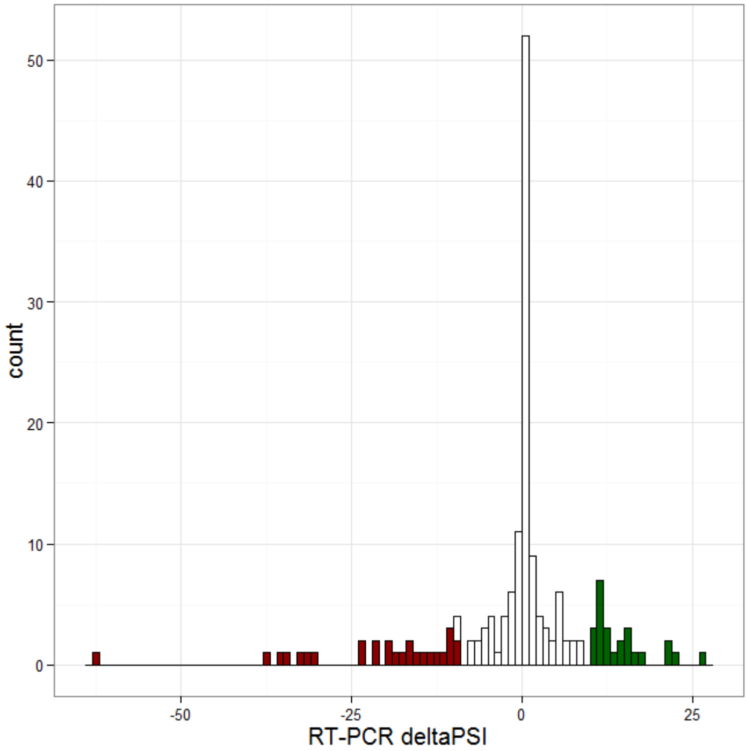
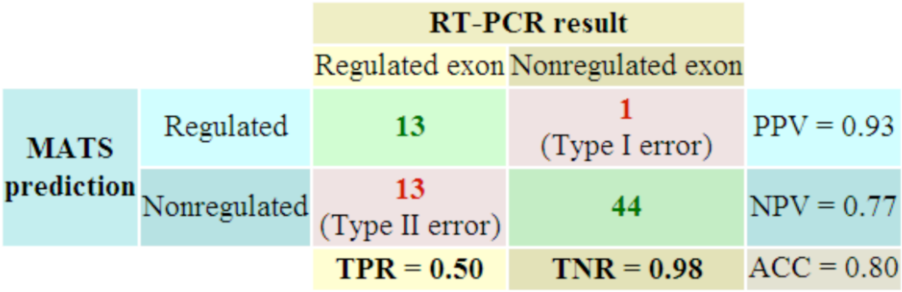


Figure 2.2. Stimulation-responsive RT-PCR results for 169 exons used as gold standards for optimization of RNA-seq detection of alternative splicing. Histogram displaying count of exons at each inclusion level change (deltaPSI) observed by RT-PCR. Red indicates the bins (width equal to 1 deltaPSI) with deltaPSI <= -10, indicating decreased inclusion upon stimulation. Green indicates the bins with deltaPSI >= 10, indicative of increased inclusion upon stimulation.

I utilized the MATS algorithm[86], a multivariate Bayesian splicing analysis program, to quantify exon inclusion changes between sample groups and their associated statistical significances. While MATS has been demonstrated to provide accurate alternative splicing predictions in other datasets, the values of parameters available for the algorithm need to be fine-tuned to a given sequencing dataset to account for variations in sequencing depth, variance between samples within sample groups and between sample groups, and experimental design. By utilizing an exhaustive sampling of combinations of different parameters that MATS uses to identify splicing targets, I generated a total of 88 different MATS analyses for stimulation-induced alternative splicing. I then utilized the 169 gold standard RT-PCR results to evaluate each of these different MATS analyses. I first extracted the positive and negative predictions from each analysis and scored these as True Positive (TP), False Positive (FP), True Negative (TP), or False Negative (FN). I then used these values to construct confusion matrices and extract confusion matrix-derived signal detection metrics, including positive predictive value (PPV, True Positive divided by the sum of True Positive and False Positive), negative predictive value (NPV, True Negative divided by the sum of True Negative and False Negative), and overall accuracy (ACC, the sum of True Positive and True Negative divided by the sum of all predictions). By comparing the 88 different MATS analyses, I identified the set of parameters (see Materials and Methods) that led to the highest PPV and ACC, achieving a positive predictive value of 93% and an overall accuracy of 80% (figure 2.3). This set of MATS parameters was then used to identify hnRNP L-responsive splicing in response to hnRNP L knockdown.



		RT-PCR result		
		Regulated exon	Nonregulated exon	
MATS prediction	Regulated	**13**	**1** (Type I error)	PPV = 0.93
	Nonregulated	**13** (Type II error)	**44**	NPV = 0.77
		TPR = **0.50**	**TNR** = **0.98**	ACC = 0.80

Figure 2.3. Confusion matrix for the MATS invocation that led to the highest PPV and ACC. These data indicate that if this MATS analysis had been used to generate positive predictions that were then tested by RT-PCR, a validation rate of 93% would be achieved.

Applying the optimized MATS parameters to the analysis of hnRNP L-depleted and mock-depleted sample groups in both unstimulated and stimulated growth conditions allows discovery of transcriptome-wide alternative splicing mediated by hnRNP L. While the degree to which these optimized parameters allows generalization across different RNA-seq analyses, the fact that the parameters were optimized for detection of alternative splicing within the RNA-seq samples I generated and yielded the highest degree of Positive Predictive Value as established by RT-PCR validation within our lab, this set of parameters is evidence that these parameters are indeed optimal for the detection of hnRNP L-responsive alternative splicing. I applied inclusion level and significance filters to subset alternative exons that exhibit an inclusion level change of at least 10% in either positive or negative directions upon hnRNP L depletion, with an accompanying p-value less than 0.05. This analysis identifies 814 and 630 hnRNP L-responsive cassette exons in unstimulated and stimulated cells, respectively. Additionally, I extracted a subset of cassette exons that do not respond to hnRNP L depletion by applying maximal inclusion level change constraints of 3% and excluding any exons with a p-value less than 0.05. This analysis identifies a set of 33,489 and 26,792 cassette exons in unstimulated and stimulated conditions, respectively. These results demonstrate that hnRNP L regulates inclusion of a specific subset of exons, with a much greater population of exons exhibiting no significant changes in inclusion.

In order to complement the high breadth-of-coverage provided by mRNA-seq data, we analyzed RASL-seq data generated by our collaborators at UCSD in the Fu lab (table 2.1). RASL-seq utilizes a custom-designed pool of splice-junction directed oligonucleotide probes that anneal across utilized junctions in RNA samples.

Subsequent capture of annealed oligonucleotides followed by high-throughput sequencing allows accurate quantification of relative junction utilization across matched probe pairs. Utilizing this technology, the Fu lab quantified over 5,500 junction pairs using a probe pool they have previously designed to known alternative splicing junctions. The Fu lab provided initial processing of the sequence reads, extracting junction pair read counts. We then further analyzed these data by excluding junction pairs that had fewer than 10 reads on average across replicates, resulting in 3,286 junction pairs to analyze. I applied identical deltaPSI and p-value constraints to these junction pairs, generating a total of 111 and 77 cassette exons with significant hnRNP L-responsive inclusion level changes in unstimulated and stimulated JSL1 cells, respectively.

I then developed a bioinformatics pipeline to integrate splicing data from RASL-seq and the PPV-optimized MATS RNA-seq analysis to identify cassette exons that exhibit hnRNP L depletion-responsive inclusion changes of at least 10 deltaPSI, where positive deltaPSI is evidence of increased exon inclusion in hnRNP L-depleted versus mock-depleted transcriptomes. Importantly, deltaPSI estimates obtained from both sequencing studies are included for exons queried by both experiments. Statistically significant alternative splicing predictions from both experiments were well correlated in both unstimulated (p=7.45e-16) and stimulated (p=7.41e-13) conditions (figure 2.1c), despite the technical differences between RASL-seq and mRNA-seq. I observed an even higher degree of correlation between the two cellular conditions within each experiment (figure 2.1d), providing evidence that hnRNP L-regulation of alternative cassette exon splicing is largely shared between conditions.

RT-PCR validation of 47 novel predictions of hnRNP L-responsive alternative splicing generated an overall 72.34% validation rate, with experimentally determined inclusion level changes well correlated with predictions (R^2 = 0.56 and 0.83 for unstimulated and stimulated cells, respectively, figure 2.1e). Taken together, these results demonstrate the high confidence of our splicing predictions and their utility in identifying novel instances of hnRNP L-mediated alternative splicing.

hnRNP L regulates exon inclusion in transcription factors, epigenetic regulators, and splicing factors

To determine the impact of hnRNP L-regulated alternative splicing in JSL1 cells, I used GO analysis to identify functional categories enriched within the set of genes that contains repressed exons and the set of genes within enhanced exons (table 2.2). I observed an enrichment for RNA binding proteins among genes containing hnRNP L-enhanced cassette exons, and genes harboring repressed exons were strongly enriched for transcription factors and chromatin modifiers.

Regulation	Category	Term	Count	P value	Fold Enrichment
Enhanced	MF	RNA binding	15	0.0161	1.98
Repressed	BP	Transcription	54	8.17E-4	1.54
	BP	Chromatin modification	15	0.00175	2.61
	BP	Regulation of transcription	59	0.00378	1.40
	MF	Transcription regulator activity	37	0.0116	1.49
	BP	Chromosome organization	18	0.0132	1.90

Table 2.2. GO terms enriched in mRNAs harboring hnRNP L-enhanced or – repressed exons compared to mRNAs expressed in JSL1 cells.

Upon observing that hnRNP L represses exon splicing in transcripts encoding DNA-binding proteins, I next tested the hypothesis that hnRNP L depletion results in differential gene expression. Using mRNA-seq aligned reads, I employed a normalized linear model implemented in the limma package for the R statistical language to identify transcripts that exhibit at least a 1.5 log2 expression difference between hnRNP L-depleted and mock-depleted conditions with an accompanying p-value below 0.05 (figure 2.4b, hnRNP L depletion in unstimulated cells; figure 2.4c, hnRNP L depletion in stimulated cells). As a positive control, I used the same analysis to identify gene expression changes between unstimulated and stimulated cells. In agreement with prior studies, I found a strong signature of upregulation of genes involved in T cell activation (figure 2.5), with 4.56% of expressed genes exhibiting significant expression changes of at least 1.5 log2 (figure 2.4a). I then performed subsequent validation of differential gene expression estimates from RNA-seq by qRT-PCR, which demonstrated excellent agreement between fold changes estimated by RNA-seq and fold changes observed by qRT-PCR (R^2 = 0.7, figure 2.4d), confirming the validity of this gene expression change analysis and the ability of qRT-PCR to confirm these changes.

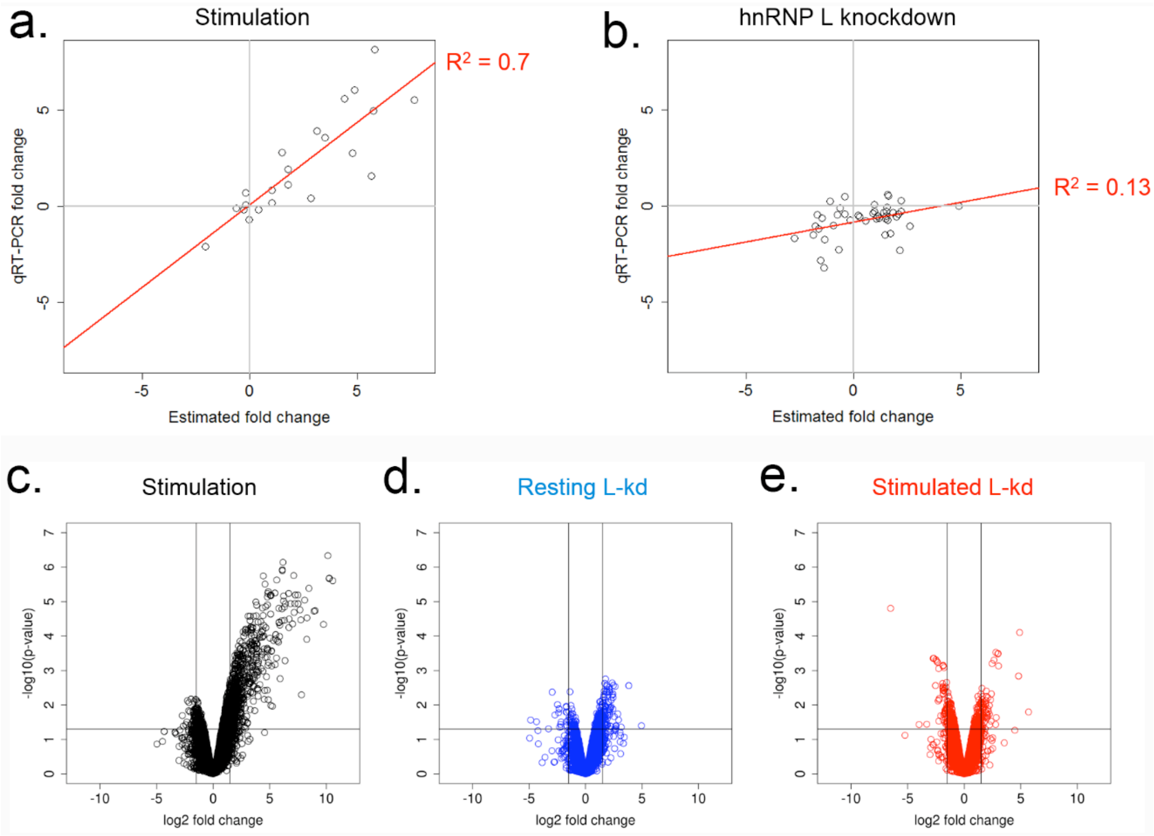


Figure 2.4. Gene expression analysis of hnRNP L-responsive differential gene expression. a.) Volcano plot of differential mRNA expression following PMA stimulation of JSL1 revealed by mRNA-seq analysis. b.) Volcano plot of hnRNP L depletion-induced differential mRNA expression in unstimulated JSL1 cells. c.) Volcano plot of hnRNP L depletion-induced mRNA gene expression in PMA-stimulated cells. d.) qPCR validation of PMA-induced differential transcript levels with linear regression. e.) qPCR validation of hnRNP L depletion-induced differential transcript levels normalized to actin.

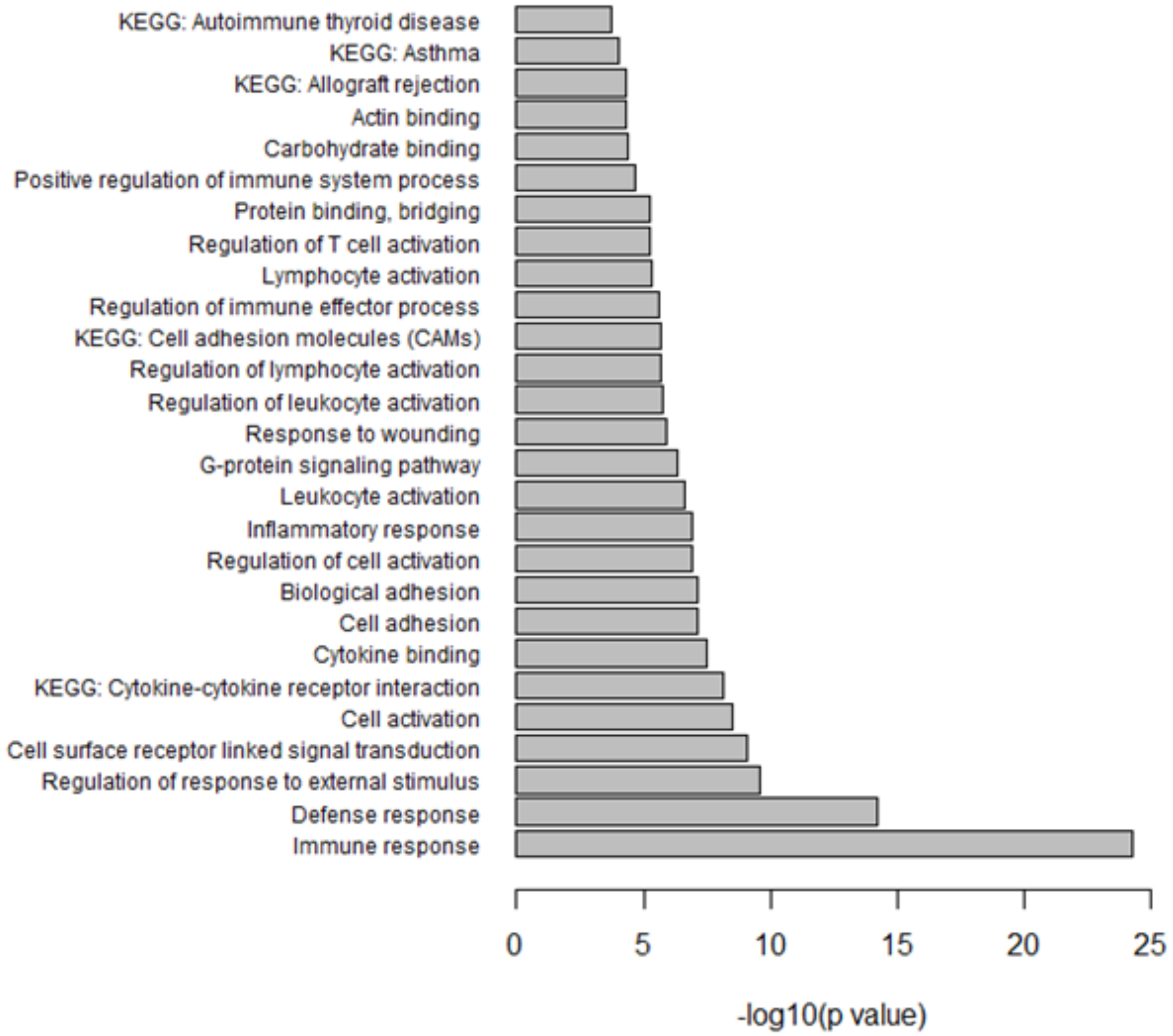


Figure 2.5. GO terms enriched in PMA-induced genes. Genes with expression changes of at least 1.5 log2 that had a p-value less than 0.05 were analyzed for enrichment of functional categories using DAVID software.

In contrast to stimulation-induced gene expression changes, applying the same analysis to hnRNP L depletion-induced differential gene expression reveals significant expression changes only a small subset of genes in unstimulated and stimulated conditions (0.63% and 0.76% of expressed genes, figure 2.4b and 2.4c, respectively), and these genes exhibit fold changes of greatly reduced ranges (compare to figure

2.4a). Subsequent qRT-PCR analysis demonstrated little correlation between fold change estimated from RNA-seq data and fold change observed by qRT-PCR (R^2 = 0.13, figure 2.4e), which elicited gene expression change values close to zero. From this analysis I concluded that while the differential gene expression analysis using mRNA-seq data is capable of sensitively and specifically identifying gene expression changes as demonstrated by the stimulation-responsive genes, hnRNP L depletion does not result in gene expression changes of high magnitude, and qRT-PCR validation demonstrates that gene expression changes are not discernible from zero.

Discussion

I report here an analysis of hnRNP L-responsive alternative splicing in JSL1 cells that combines the high breadth-of-coverage of mRNA-seq with the high depth-of-coverage of RASL-seq. Importantly, mRNA-seq analysis was optimized for confusion matrix-derived signal detection metrics utilizing a set of gold standard RT-PCR data that were generated in our laboratory from the same cell line. I find that RASL-seq and RNA-seq predictions are well correlated with each other in unstimulated and stimulated cells, and each experiment is very highly correlated between cell states. These results demonstrate that hnRNP L splicing regulation is highly consistent between unstimulated and stimulated T cells, in agreement with my previous finding that the majority of hnRNP L localization is not altered upon stimulation as measured by CLIP-seq.

I identify hnRNP L-responsive alternative splicing events in transcripts encoding RNA binding proteins, a result that is consistent with prior transcriptomics studies of splicing factors, which often report significant cross-regulation of splicing factors[62]. Our results extend existing observations of hnRNP L-responsive splicing in the pre-mRNAs encoding other splicing factors[134] and add hnRNP L to the growing list of splicing factors that exist not in isolation, but within a network of interconnected splicing events that work together to control the splicing of the transcriptome.

In addition to RNA binding proteins, DNA binding proteins are enriched among hnRNP L functional targets, including transcription factors, chromatin modifiers, and other epigenetic regulators. This observation led me to test the hypothesis that hnRNP L might be involved in control of gene expression. This hypothesis is supported by prior studies linking hnRNP L to the mediator complex[137], to miRNA regulation[54] and to alternative splicing in poison exons, as we discovered for CCAR1[133]. Interestingly, while a computational analysis of PMA-induced differential gene expression identifies a broad pattern of upregulated genes enriched for activation-related functional categories, the same analysis applied to hnRNP L depletion-induced differential gene expression identifies comparatively few genes with statistically significant gene expression changes greater than 1.5 log2. Subsequent qRT-PCR validation of PMA-induced expression changes provides high correlation between expression changes obtained from RNA-seq and from qRT-PCR analyses, confirming the accuracy of the RNA-seq and qRT-PCR analyses. However, qRT-PCR validation of hnRNP L-responsive gene expression targets revealed little correlation between RNA-seq and qRT-PCR expression changes, indicating the possibility that the few genes with hnRNP L-responsive gene expression changes in the RNA-seq analysis were due to sequencing noise and/or chance variation instead of *bona fide* hnRNP L regulation of gene expression.

Another possible explanation for the lack of widespread gene expression changes in my experiments is timecourse. PMA stimulation results in engagement of membrane-proximal cell signaling pathways that in turn result in post-translational modification changes, such as decreased phosphorylation of ERK1 and ERK2[138], which results in differential gene expression by activity changes induced by post-translation modifications. In contrast, hnRNP L-responsive alternative splicing in transcription factors and epigenetic regulators could take considerably more time to result in gene expression changes, as the shift in isoforms induced by hnRNP L knockdown requires nuclear export and translation into protein before these isoform shifts even manifest at the level of the proteome. For this reason, I cannot rule out the potential for hnRNP L regulation of gene expression in T cells, however for experimental reasons, hnRNP L depletion cannot be carried out over a longer time course with the current technology.

Complementary next-generation sequencing approaches here reveal a broad set of hnRNP L functional targets in T cells, however the mechanism(s) by which hnRNP L regulates alternative splicing is only known for a handful of cases. Importantly, our lab has previously described that hnRNP L can regulate alternative splicing in both directions: enhancement and repression, and that a balance of multiple factors, including co-associated proteins, splice site strengths, and location of binding may work together to determine hnRNP L splicing regulatory activity. Knowledge of the functional targets of hnRNP L in T cells sets the stage for integrative genomic analysis, opening new avenues for computational dissection of the combinatorial control of pre-mRNA splicing by hnRNP L.

CHAPTER 3 - INTEGRATIVE GENOMIC ANALYSIS OF hnRNP L SPLICING REGULATORY FUNCTION

Introduction

I have previously reported transcriptome-wide analyses of hnRNP L physical and functional targets. CLIP-seq analysis revealed the landscape of hnRNP L-RNA interactions, and RNA-seq and RASL-seq analysis has identified hundreds of exons whose inclusion level exhibits significant changes upon hnRNP L depletion. Integrative genomic analysis aims to combine binding and function data to provide insights into the positional dependence of RNA binding protein occupancy on splicing regulation.

Integrative genomics techniques have been previously applied to several other splicing factors, including members of the hnRNP and SR protein families[62]. This analysis often features graphics known as RNA maps which relate the fraction of regulated cassette exons containing a CLIP-seq peak to each nucleotide within and around enhanced, repressed, and unresponsive exons. RNA maps are a useful exploratory tool that provide insight into the positions of RNA binding protein interaction from which direct regulation of splicing is likely to be achieved. Even before the development of the CLIP-seq experiment, RNA mapping was used to relate the position of RNA binding protein motifs to enhanced and repressed cassettes. This approach was used to demonstrate a strategy used by the hnRNP protein PTB to repress splicing from upstream or exonic positions and to activate splicing from downstream positions[135]. These analyses demonstrate the importance of determining the positional dependence of protein-RNA interactions on splicing regulation.

Several important limitations to the integrative genomics approach exist. First, the overlap between binding and function is generally low. As a consequence of this, the majority of exons with hnRNP L CLIP-seq peaks either within or adjacent to the exon are unresponsive to hnRNP L depletion. Conversely, the fraction of hnRNP-responsive exons containing CLIP-seq peaks within 300nt of the exons is typically 5-10%. This relatively low degree of overlap between binding and function is mysterious, but is likely to arise from a combination of factors including indirect regulation of splicing, false negative CLIP-seq sites due to low transcript expression or low mappability, false positive functional targets due to chance variation, or other factors.

Even with the caveat that the overlap between binding and function is typically low in integrative genomic studies, mechanistic insights can be empirically derived through these analyses. Several important examples exist in the literature. First, integrative genomic analysis revealed that hnRNP A1 has statistically enriched binding within repressed exons, consistent with prior *in vitro* and *in vivo* studies of hnRNP A1 splicing regulation, which highlight a direct repressive role[62]. In contrast, another hnRNP protein, hnRNP A2/B1, does not display enrichment for exonic interactions, but instead binds on both sides of the exon. Two other hnRNP proteins, hnRNP F and hnRNP U, do not display any enrichment for binding within or around repressed exons, but instead have enrichment upstream of enhanced exons. These binding patterns are summarized in table 3.1.

Protein:	Repressor binding pattern:	Enhancer Binding pattern:
hnRNP A1	Exonic	none
hnRNP A2/B1	Flanking	Far upstream?
hnRNP F	None	75nt upstream
hnRNP H1	Upstream of C2?	Upstream, within, and downstream
hnRNP M	Downstream?	Upstream, within, and downstream
hnRNP U	None	Upstream and downstream

Table 3.1. Mechanistic hypothesis for splicing regulation by hnRNP proteins derived from prior integrative genomics studies. RNA maps from studies that overlaid CLIP-seq binding data with transcriptome-wide functional data provide empirical insights into the locations within and around regulated exons that display increased binding relative to unresponsive exons.

Importantly, mechanistic hypotheses can be derived from these studies. First, hnRNP A1 is thought to act as a splicing repressor through exonic interactions, as is the case for CD45 exon 4[46]. In contrast, hnRNP H1 may enhance splicing through exonic interactions. hnRNP A2/B1 displays enriched binding upstream of and downstream of repressed exons but not within the exons, consistent with a loop-out model in which hemophilic protein interactions bring the upstream and downstream RNA regions into close proximity, occluding the exon and its splice sites in a loop. hnRNP M displays markedly increased binding upstream of, within, and downstream of repressed exons (table 1), indicative of a mechanism in which the protein first binds to a high-affinity site, then through hemophilic protein-protein interactions spreads across neighboring RNA regions to occlude splice sites. While these hypotheses are purely empirical, they provide valuable insight that may guide detailed biochemical studies.

While other hnRNP proteins have been studied with integrative genomic analysis, important features such as splice site strengths are rarely included. Prior studies by our lab have implicated splice site strengths as important determinants of hnRNP L splicing regulation. In this chapter, I extract additional combinatorial features such as splice site strengths and exon/intron lengths to provide insights into the mechanisms by which hnRNP L positively and negatively regulates alternative splicing in a combinatorial manner.

We have previously described dozens of validated cases of hnRNP L-regulated alternative cassette exon splicing, however the mechanisms by which hnRNP L directly or indirectly regulates these functional targets remain unknown. Additionally, prior studies have demonstrated that location of interactions, splice site strengths, and co-associated proteins establish combinatorial control of splicing by hnRNP L. In this chapter, I use integrative genomic analysis to combine binding and functional data to generate mechanistic hypotheses about how hnRNP L positively and negatively regulates exon inclusion.

Importantly, I find that hnRNP L is enriched for binding within, upstream of, and downstream of repressed exons. In contrast, hnRNP L-enhanced exons do not display enrichment of hnRNP L interactions, suggesting indirect regulation. Importantly, hnRNP L-enhanced exons are flanked by short, GC-rich introns and are characterized by decreased nucleosome occupancy. These results indicate a possible epigenetic mechanism by which hnRNP L enhances splicing.

Results

Like other splicing regulatory proteins, hnRNP L can both enhance and repress alternative exon splicing, but the mechanisms by which these opposing regulatory functions may be effected by hnRNP L remain unclear. To investigate the features that distinguish enhanced from repressed exons in our splicing predictions, I first compiled an RNA map of hnRNP L-enhanced and –repressed exons, comparing to a stringently-defined set of unresponsive exons that meet the requirement of an inclusion level change (deltaPSI) of less than 3% in either direction as well as a p-value greater than or equal to 0.05. To provide an additional level of stringency, the unresponsive exons were required to meet these cutoffs in both unstimulated (resting) and stimulated cells, and additionally these exons must have been queried by both RASL-seq and RNA-seq, meeting the stringent criteria in both experiment. This represents the highest level of stringency I can apply to define hnRNP L-unresponsive exons, and these requirements resulted in a set of 250 unresponsive exons.

Prior analyses have suggested that splice site strength plays a role in hnRNP L-regulated alternative splicing[44,133,133]. To compare splice site strengths among responsive and unresponsive cassettes, I extracted splice site scores using the MaxEntScan algorithm for each of the four splice sites in cassettes (Figure 3.1). I observed that both repressed and enhanced exons have weaker 3' splice sites (3'ss) than those found in unresponsive cassettes (repressed p=0.0293, enhanced p=6.51e-6, t-test), and that enhanced exons have even weaker 3'ss than repressed exons (p=0.013). This analysis demonstrates a critical role for 3'ss in determination of both the responsiveness and the directionality of response to hnRNP L depletion, confirming prior studies.

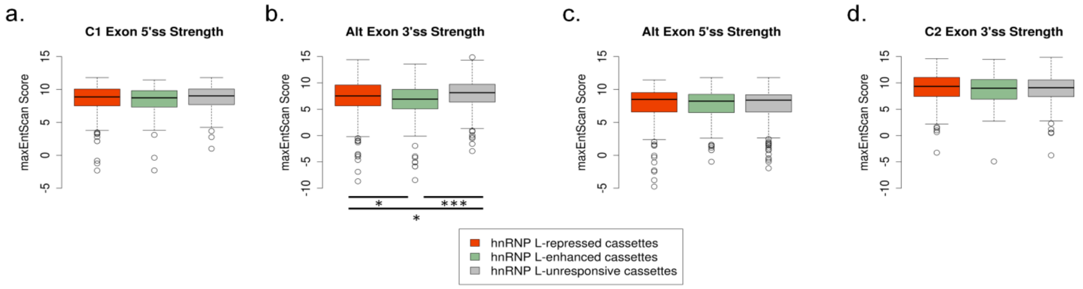


Figure 3.1. Splice site strengths in hnRNP L-responsive and –unresponsive cassettes. Splice site sequences for the four splice sites in hnRNP L-responsive or –unresponsive cassettes were scored with the maxEntScan method. Plots are arranged from left to right in 5' to 3' order: a) the 3' splice site of the C1 exon, b) the 5' splice site of the alternative exon, c) the 5' splice site of the alternative exon, and d) the 3' splice site of the C2 exon.

Given our prior observation that hnRNP L can repress and enhance splicing of exons that have CLIP-seq binding sites upstream and/or downstream of the exon, I next investigated the possibility of positional dependence of hnRNP L occupancy on splicing outcomes by overlaying hnRNP L CLIP-seq data within and around splice sites within cassettes, a computational technique known as RNA mapping (figure 3.2). By computing the fraction of repressed exons occupied by hnRNP L CLIP-seq binding sites at single-nucleotide resolution and comparing to unresponsive cassettes, I observed a marked increase in hnRNP L occupancy 100nt upstream and 40nt downstream of alternative exons' 3'ss. I also observed increased hnRNP L occupancy downstream of the exons' 5'ss, demonstrating that hnRNP L repression is associated with protein-RNA interactions in exonic and exon-proximal intervals. A similar comparison of enhanced exons to unresponsive exons demonstrates that a lower total fraction of enhanced cassettes contain hnRNP L CLIP-seq binding sites at any given position, and the overall pattern is similar to unresponsive exons. This analysis provides evidence that hnRNP L directly represses alternative exon inclusion through exonic and exon-proximal protein-RNA interactions, and that enhancement is not associated with an enrichment of hnRNP L occupancy within these RNA regions.

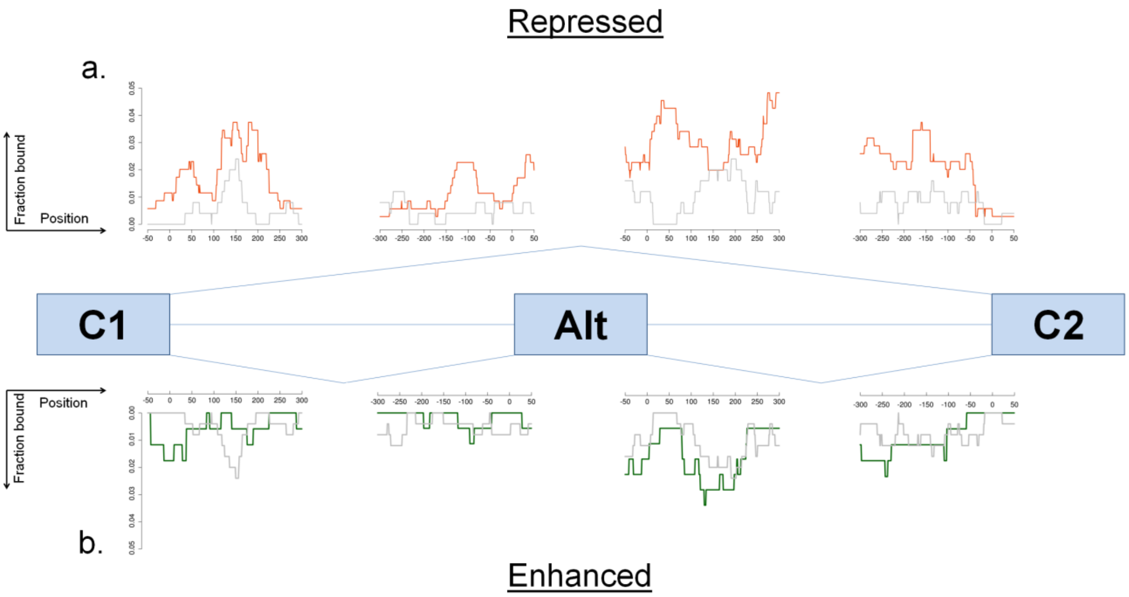


Figure 3.2. Positional dependence of hnRNP L splicing regulation: the hnRNP L RNA map. The fraction of hnRNP L-responsive and –unresponsive cassettes containing hnRNP L CLIP-seq peaks is plotted at nucleotide resolution separately for (a) hnRNP L-repressed cassettes and for (b) hnRNP L-enhanced cassettes. Unresponsive cassettes (gray) are plotted as a negative control.

As exon and intron length have both been implicated in exon inclusion and alternative splicing[139], I compared intron and exon lengths in hnRNP L-responsive and –unresponsive cassettes (Figure 3.3). Surprisingly, I observed that the introns upstream (I1) or downstream (I2) of enhanced exons are significantly shorter than those flanking repressed exons (I1 p=0.00077, I2 p=0.03924, t-test). Additionally, both types of regulated cassettes have longer alternative exons than unresponsive exons (repressed p=0.0157, enhanced p=0.04451). These data indicate alternative exon length as a potential feature involved in hnRNP L regulation of alternative exon inclusion and implicate intron length as a feature that differentiates repressed from enhanced cassettes.

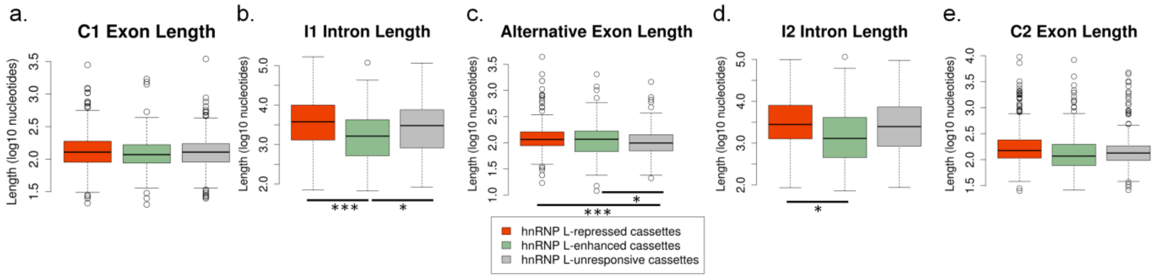


Figure 3.3. Exon and intron lengths in hnRNP L-responsive and –unresponsive cassettes. Lengths of exons and introns (intron lengths expressed as log10) for the 5 exons and introns are plotted from left to right in 5’ to 3’ order: a) C1 exon length, b) I1 intron length, c) alternative exon length, d) I2 intron length, and e) C2 exon length. Statistical hypotheses were tested using non-log-transformed lengths.

Having observed a spatial hnRNP L binding signal within and around repressed but not enhanced exons, I applied a statistical analysis to the upstream, exonic, and downstream intervals around hnRNP L-responsive and –unresponsive exons (figure 3.4). I observed a statistically significant increase in the fraction of L-repressed exons containing at least one CLIP-seq site within the exon or within the exon-proximal upstream or downstream 300nt regions when compared to unresponsive or enhanced cassettes (p < 0.001 for all comparisons between repressed and any other sample group). In contrast, enhanced cassettes are not enriched for hnRNP L occupancy in any of these regions, even when the interval is widened to the entire flanking introns. These data provide further evidence for a direct repressive role for hnRNP L from exonic or proximal intronic positions on either side of the regulated exon.

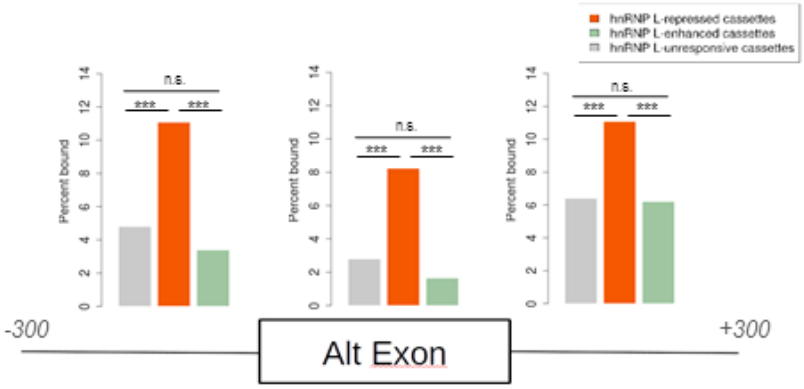


Figure 3.4. Repression of splicing by hnRNP L is associated with exonic and periexonic interactions. The fraction of upstream 300nt intervals, exonic intervals, and downstream 300nt intervals containing at least one hnRNP L CLIP-seq peak is plotted for three sample groups: hnRNP L-repressed exons, hnRNP L-enhanced exons, and unresponsive exons.

Upon observing that hnRNP L occupancy is not enriched within enhanced cassettes, I next hypothesized that another splicing factor with hnRNP L-dependent expression and/or activity might mediate indirect enhancement of splicing. To examine this possibility, I developed an exon-directed *de novo* motif enrichment strategy to elicit sequence features enriched within and around exons after partitioning for hnRNP L occupancy (see Materials and Methods). Importantly, this motif enrichment analysis is specifically designed to identify motifs enriched in indirect splicing targets. First, I extracted potential indirect splicing targets of hnRNP L by partitioning cassettes into bound or unbound based on the presence or absence of any hnRNP L CLIP-seq binding site within the cassette. I then compared hexanucleotide sequences enriched in intervals upstream of, downstream of, or within the enhanced or repressed exons against background sequences extracted from the same regions (upstream, exonic, or downstream) from all refSeq internal exons (see Materials and Methods). Subsequent statistical analysis allows elicitation of potential *cis*-regulatory motifs that could provide insight into the regulation of hnRNP L-responsive exons that do not have hnRNP L CLIP-seq sites within exonic or periexonic regions of the pre-mRNA.

I found no sequences to be significantly enriched upstream of, within, or downstream of hnRNP L-repressed exons that are not bound by hnRNP L (figure 3.5a). In contrast, a GC-rich sequence feature was found to be enriched upstream of and within enhanced and unbound exons (most significant hexamer is CCGCGG, logo of all significant hexamers is displayed). A further comparison of the fraction of all hnRNP L-repressed, -enhanced, and –unresponsive cassettes that have the GC-rich motif within or upstream of the alternative exon demonstrates that this sequence feature is significantly depleted in repressed cassettes.

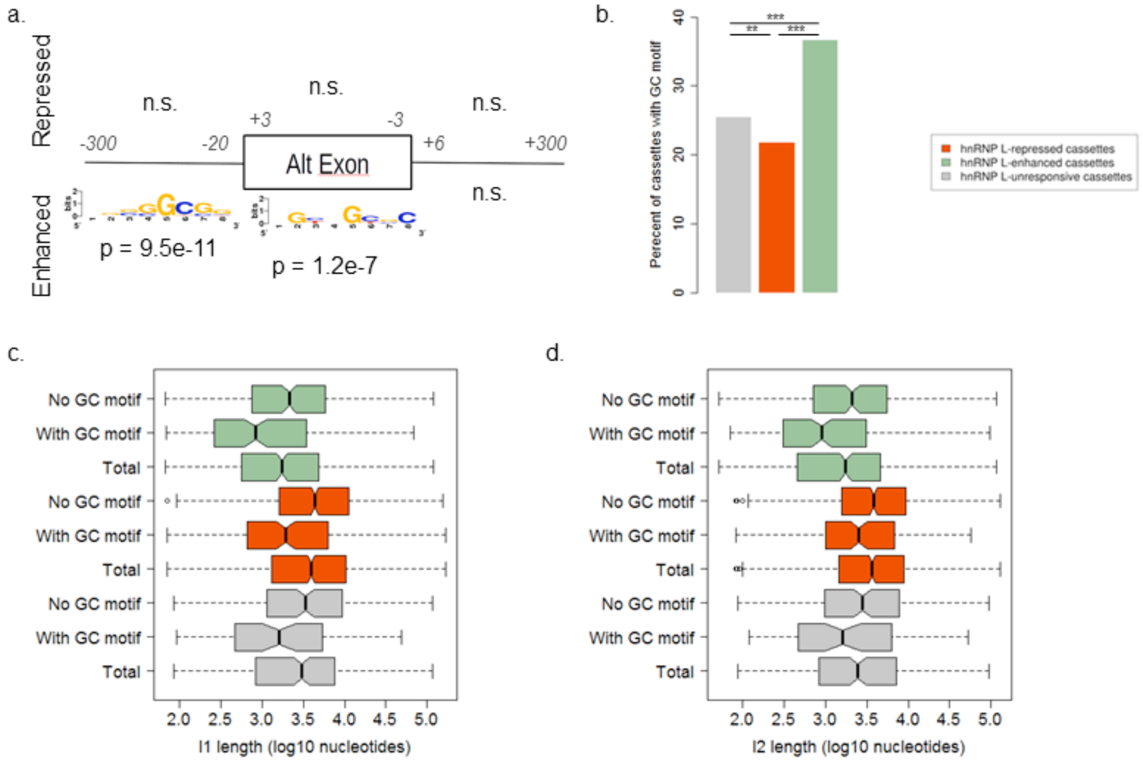


Figure 3.5. A GC-rich motif is enriched upstream of and within indirectly hnRNP L-enhanced exons. To investigate possible mechanisms of indirect enhancement of splicing by hnRNP L, the cassette exons enhanced by L with no CLIP-seq peaks in the entire cassette were first extracted (a). Sequences from upstream (-300 to -20nt), exonic (+3 to -3nt), and downstream (+6 to +300nt) were extracted, avoiding the splice site sequences themselves. Binomial comparison of the fraction of sequences containing at least one occurrence of each *k*mer of length 6nt was performed, using cognate intervals from all refSeq internal exons as a background. All significant hexamers were then aligned and a motif logo is presented, with the lowest p-value from the hexamers in the mutliple sequence alignment displayed. (b) The fraction of hnRNP L-repressed, -enhanced, and –unresponsive exons containing any of the significant hexamers displayed in the motif logos in (a) within the upstream and exonic intervals are plotted. Unlike in (a), the entire sets of exons are investigated, demonstrating global enrichment/depletion of the GC-rich motifs. (c) The hnRNP L-repressed, -enhanced, and –unresponsive exons were partitioned into by occurrence of any of the significant hexamers displayed in (a) within the upstream and exonic intervals, and flanking intron lengths were plotted on a log10 scale.

These results demonstrate that a GC-rich sequence feature is significantly enriched in hnRNP L-enhanced exons and significantly depleted in hnRNP L-repressed exons when compared to unresponsive exons (figure 3.5b). Another feature that strongly differentiates hnRNP L-enhanced exons in the shortness of the flanking introns.

I next investigated the possibility that these two features co-occurred within the hnRNP L-enhanced exons. I first partitioned the hnRNP L-enhanced exons by the occurrence of any of the significantly enriched hexamers. The set of enhanced exons that contain the GC-rich motif either upstream or within the exon indeed have shorter upstream and downstream introns than the total enhanced exons or the enhanced exons that do not have the motif (figure 3.5c upstream intron, figure 3.5d downstream intron). This finding suggests an association between short introns and the GC-rich motif, as has been previously described on a transcriptome-wide level[139,140,140]. In support of a global association between this set of GC-rich motifs and short flanking introns, identical partitioning of hnRNP L-repressed and –unresponsive cassettes also results in the subset of both classes of cassettes that contain the motif displaying shorter flanking introns. In sum, a GC-rich motif is enriched within and upstream of hnRNP L-enhanced exons, and this GC-rich motif is associated with short flanking introns across all sets of exons investigated, suggesting the motif and the shortness of introns are globally associated in a manner that is not specific to hnRNP L-enhanced exons.

These results, combined with prior global observations of two distinct classes of exons based on intron length and GC content[139,140,140], suggest the possibility that there is a fundamental mechanistic difference between the manner in which hnRNP L-enhanced exons are recognized by the splicing machinery, and it is this difference that might explain the manner in which hnRNP L may enhance splicing of exons that are not subject to direct physical interaction.

An alternative hypothesis for indirect enhancement of splicing by hnRNP L is via another splicing factor that engages the GC-rich motif enriched within and upstream of enhanced exons. Recent technological advancements have enabled the *in vitro* characterization of RNA binding protein recognition specificities[141,142,142]. I conducted a literature search for potential RNA binding proteins that might engage the GC-rich motif identified upstream of and within hnRNP L-enhanced alternative exons, identifying 4 candidate proteins: SRSF2 (SC35), RBM4, Y14, and FUS. A technician in our lab, Michael Mallory, then used protein extracts from hnRNP L-depleted or mock-depleted JSL1 cells to test for hnRNP L-responsive changes in protein level or migration. Importantly, no consistent changes were observed in any of the 4 proteins tested, suggesting that hnRNP L depletion does not induce changes in the concentration of any of these RNA binding proteins.

Prior transcriptome-wide studies have identified nucleosome occupancy as a demarcating factor for exons that are flanked by long, GC-poor introns. This type of exon is common in the human transcriptome. However, exons flanked by short, GC-rich introns do not display a marked increase in nucleosome occupancy when compared to the flanking introns. To investigate the possibility that hnRNP L-enhanced exons and –repressed exons have different patterns of exonic nucleosome occupancy, I extracted the average %GC at single-nucleotide resolution for the same intervals examined in the hnRNP L RNA map (figure 3.6). Consistent with the *de novo* motif enrichment results, the hnRNP L-enhanced exons display increased GC content upstream of, within, and downstream of the alternative exons. However, compared to hnRNP L-repressed exons and to hnRNP L-unresponsive exons, hnRNP L-enhanced exons display a reduction in the GC-content differential between the exonic and perixonic regions for the 5' splice site (figure 3.6a) and the 3' splice site (figure 3.6b).

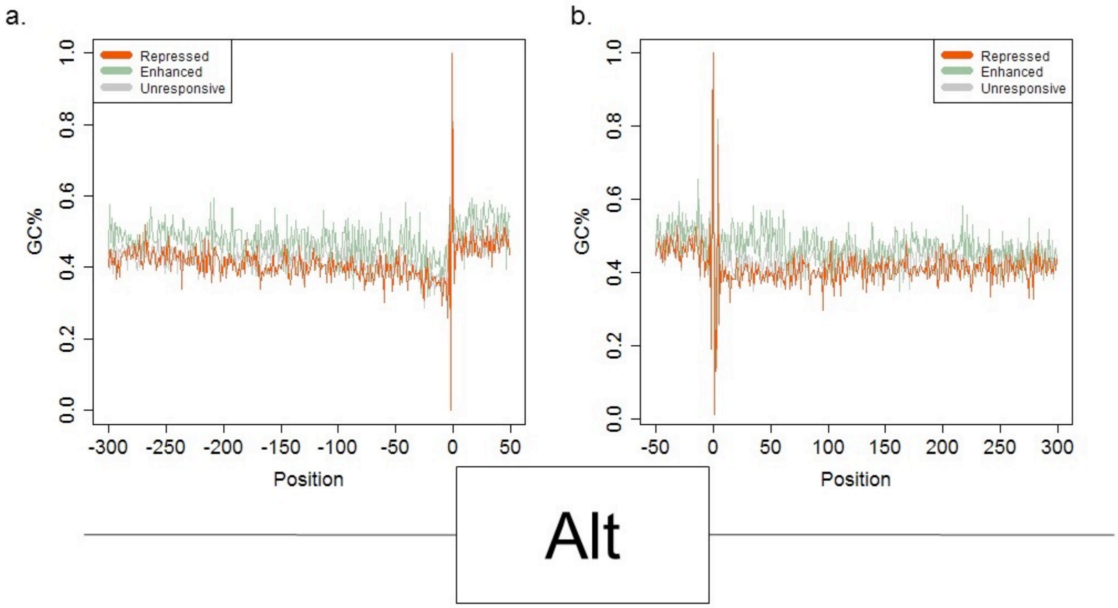


Figure 3.6. GC architecture for hnRNP L-responsive and –unresponsive exons. Average fraction of nucleotides that are G or C at each nucleotide for a 300nt window containing 50nt of exonic sequence and 300nt of flanking intronic sequence were separately computed for hnRNP L-enhanced, hnRNP L-unresponsive, and hnRNP L-repressed exons. The splice donor and acceptor nucleotides are demarcated by zero and 1 values for the GT..AG dinucleotide sequences that are core features of the respective splice sites.

I subsequently quantified the mean %GC for equal-sized 50nt intervals on either side of the two splice sites for enhanced, repressed, and unresponsive exons and computed the difference between exonic %GC and intronic %GC (figure 3.7). Importantly, I defined the downstream GC differential as mean %GC of the last 50 nucleotides of the exon (up to but not including the final 3nt of the exon that are part of the 5' splice site) minus the mean %GC of the first 50 nucleotides of the downstream intron (excluding the initial 6nt of the intron that are part of the 5' splice site). The downstream GC differential for hnRNP L-enhanced exons is much lower than that for repressed or unresponsive exons (1.3% versus 6.8% and 5.6%, respectively). This suggests that hnRNP L-enhanced exons might display reduced nucleosome occupancy, as has been observed globally for exons flanked by short introns. It is worth noting that the upstream GC differential was not as notably deficient for hnRNP L-enhanced exons, although this can potentially be explained by sequence constraints imposed by polypyrimidine tracts located upstream of the exons.

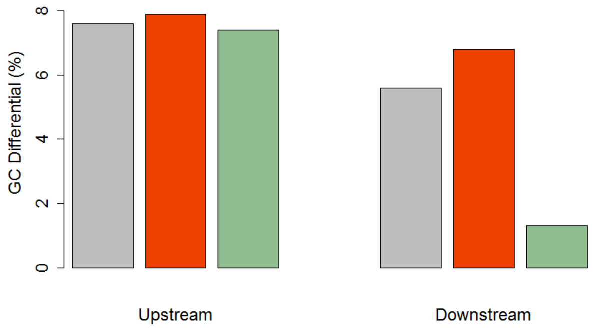


Figure 3.7. Upstream and downstream GC content differentials across splice sites in hnRNP L-responsive and –unresponsive cassettes. Mean %GC for 50nt intervals on both sides of both splice sites for hnRNP L-unresponsive exons (gray), repressed exons (red), and enhanced exons (green) were computed as exonic minus intronic mean %GC.

Nucleosome occupancy is a demarcating feature of human exons flanked by long introns. Compared to flanking intronic regions, the exonic DNA of human genes displays an increase in nucleosome occupancy as evidenced by nucleosome-sensitive DNA sequencing methodologies such as bisulfite sequencing. This increased nucleosome occupancy is thought to enhance spliceosomal assembly at exons that are buried in long introns by slowing transcription. This increase in nucleosome occupancy is associated with a pronounced GC-content cliff at the two splice sites of the exons: the GC content of the exon is higher than its neighboring introns.

To investigate the hypothesis that hnRNP L-enhanced exons display reduced nucleosome occupancy relative to their flanking intronic regions, I extracted nucleosome occupancy scores for hnRNP L-enhanced, hnRNP L-repressed, and hnRNPL-unresponsive exons from ENCODE MNase-seq data (K562 cells). In agreement with previous studies of nucleosome occupancy, hnRNP L-unresponsive exons display elevated nucleosome occupancy within the exon relative to the surrounding periexonic intervals (figure 3.8). Similarly, hnRNP L-repressed exons are demarcated by increased nucleosome occupancy, even when partitioned for absence of hnRNP L CLIP-seq peaks (unbound). In contrast, I observed a reduction in the degree of nucleosome occupancy in hnRNP L-enhanced exons, especially visible at the 5' splice site, around which the GC content differential was strikingly low for enhanced exons (figure 3.8b).

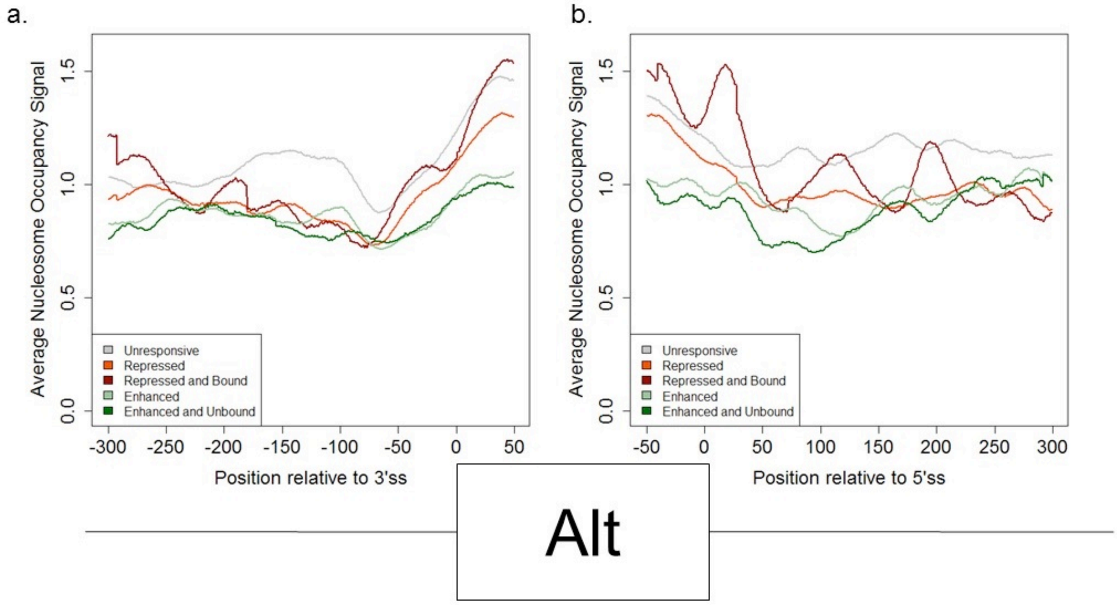


Figure 3.8. Nucleosome occupancy map of hnRNP L splicing regulation. Average nucleosome occupancy signals at each nucleotide for hnRNP L-repressed, hnRNP L-enhanced, and hnRNP L-unresponsive exons are plotted. Additional series for enhanced exons that have no hnRNP L CLIP-seq peaks in the entire cassette and for repressed exons that do have at least one hnRNP L CLIP-seq peak in the entire cassette are plotted. Nucleosome occupancy data were extracted from ENCODE K562 cells.

Discussion

In this chapter, I used integrative genomic analysis to explore the features that characterize hnRNP L repressed and enhanced exons. Importantly, I identify enrichment of hnRNP L-RNA interactions within, upstream of, and downstream of repressed exons. This analysis suggests that hnRNP L represses splicing through both splice sites, potentially by blocking access to splice sites by the splicing machinery, preventing the early steps of spliceosome assembly.

In contrast to other splicing factors that have been studied by integrative genomics techniques, hnRNP L does not demonstrate enrichment for interactions within or around enhanced exons. This demonstrates that the majority of splicing enhancement by hnRNP L is likely indirect. A motif enrichment approach identified a GC-rich motif within and upstream of L-enhanced exons. While this finding initially raised the possibility of secondary effects via another splicing factor that is itself responsive to hnRNP L depletion, subsequent analysis demonstrated that this hypothesis in unsupported. The technician in our lab, Michael Mallory, used western blotting to investigate protein level of RBM4, RBM8a, SC35, and FUS in response to hnRNP L knockdown. These candidate proteins were identified based on affinity studies such as RNAcompete and RNA Bind-N-Seq[141,142,142]. Importantly, none of these proteins demonstrated hnRNP L-responsive changes in protein level.

Transcriptome-wide analysis by Gil Ast and colleagues has identified an association between short flanking introns and a leveled GC-architecture[139]. This class of exons was found to be depleted in nucleosome occupancy when compared to exons with long flanking introns and a well-defined GC content differential between exon and introns. All of these features are associated with hnRNP L-enhanced exons. This finding raised the possibility that hnRNP L-enhanced exons are more susceptible to alterations in chromatin because they are already poorly defined by nucleosome occupancy. Nucleosome occupancy and the epigenetic modifications of histone proteins play an important role in splicing outcomes, and the dynamic interrelationship between RNA and chromatin is a subject of increasing appreciation.

Recent work in the Reinberg group physically and functionally links hnRNP L to histone methylation. The human Set2 complex, also known as the KMT3a complex, is responsible for trimethylation of lysine 36 on the histone H3 protein (H3K36me3). This epigenetic mark is associated with actively transcribed regions and is known to recruit the histone deacetylase Rpd3 in yeast[143]. Subsequent deacetylation of open reading frames protects against internal transcription initiation[144]. In humans, hnRNP L copurifies with the C-terminal half of the KMT3a complex and is required for its H3K36me3 activity in vivo[145]. Importantly, this requirement is likely physical as hnRNP L knockdown does not deplete the KMT3a complex.

The H3K36me3 modification is enriched at exon-intron boundaries in humans, suggesting that this modification marks exons within gene bodies[146]. The finding that hnRNP L depletion globally reduces the H3K36me3 modification suggests a functional link between transcription of nascent pre-mRNA and the H3K36me3 mark that demarcates exons by hnRNP L. I hypothesize that hnRNP L knockdown in JSL1 T cells reduces the H3K36me3 mark in a global manner and that hnRNP L-enhanced exons are particularly sensitive to reduction in H3K36me3 because they are poorly demarcated by nucleosomes in the first place and have short flanking introns. Transcription through these exons rapidly exposes competing downstream splice sites, which are preferentially utilized by the spliceosome upon hnRNP L knockdown.

In sum, integrative genomic analysis provides support for a model in which hnRNP L directly represses exon inclusion through interactions within or near exons. In contrast, hnRNP L enhanced splicing indirectly through a potential epigenetic mechanism. These results significantly expand our knowledge of splicing control by hnRNP L and raise interesting hypotheses about the interplay between RNA and chromatin.

APPENDIX 1: INVESTIGATING DDX17-RNA INTERACTIONS IN RIFT VALLEY FEVER VIRUS INFECTION

Introduction

In previous chapters I related how Ganesh Shankarling and I cooperated to perform and analyze hnRNP L CLIP-seq. Specifially, my contribution was in the analysis of the hnRNP L CLIP-seq data, including developing software pipelines for peak calling and the use of parallel execution on grid-based compute clusters. While my focus was on hnRNP L, it is important to note that the pipeline I developed is generalizable to other CLIP-seq experiments. The first test and demonstration of the ability to generalize my pipeline for the analysis of other proteins came from a collaboration with the group of Sara Cherry to study the RNA binding of a DEAD-box RNA helicase, DDX17, that was found in a knockdown screen to restrict the replication of an RNA virus, Rift Valley Fever Virus (RVFV).

The CLIP library for DDX17 was prepared by Ryan Moy and Ganesh Shankarling from human cells that had been infected with RFVF. Once sequencing of the library was complete I carried out alignment of the reads to a metagenome index containing the chromosomes of the human genome and the RNA segments of the viral genome, and completed the subsequent bioinformatic analysis. To my knowledge, this is the first demonstration of metagenomic CLIP-seq. The success of this analysis provides strong evidence that CLIP-seq is a useful tool to investigate host-pathogen interactions, expanding the utility of CLIP-seq as an experimental protocol.

In this appendix, I provide a report of computational analysis of CLIP-seq data that I carried out. This appendix, combined with the second appendix to this thesis, provide a valuable set of comparisons: three proteins that were subjected to CLIP-seq library preparation by the same individual, Ganesh Shankarling, and were analyzed by the same individual, myself, with minor variations in the analysis as required by the details of each experiment. In the concluding chapter of my thesis, I provide a comparison between the results of the computational analysis of these three CLIP-seq experiments with the aim of identifying key similarities and differences.

Results

Because DEAD-box helicases function as RNA-binding proteins, the Cherry lab hypothesized that DDX17 may directly bind RVFV RNAs to inhibit viral replication. To determine the specific RNAs bound to endogenous DDX17, Ryan Moy, graduate student with Sara Cherry, performed CLIP-seq. Briefly, uninfected or RVFV-infected U2OS cells were UV-irradiated, and endogenous DDX17-bound RNAs were digested to ~100 nt fragments, immunoprecipitated from cell lysates with anti-DDX17 or anti-FLAG as a control, and radiolabeled for visualization. DDX17 was efficiently depleted from the lysates with anti-DDX17 but not anti-FLAG antibodies (Figure A1.1a). Autoradiography of RNA-protein complexes revealed extensive signal for anti-DDX17 but not anti-FLAG immunoprecipitations, suggesting enrichment for DDX17-bound RNAs (Figure A1.1b). cDNA libraries were then generated from purified RNAs and submitted for Illumina deep-sequencing.

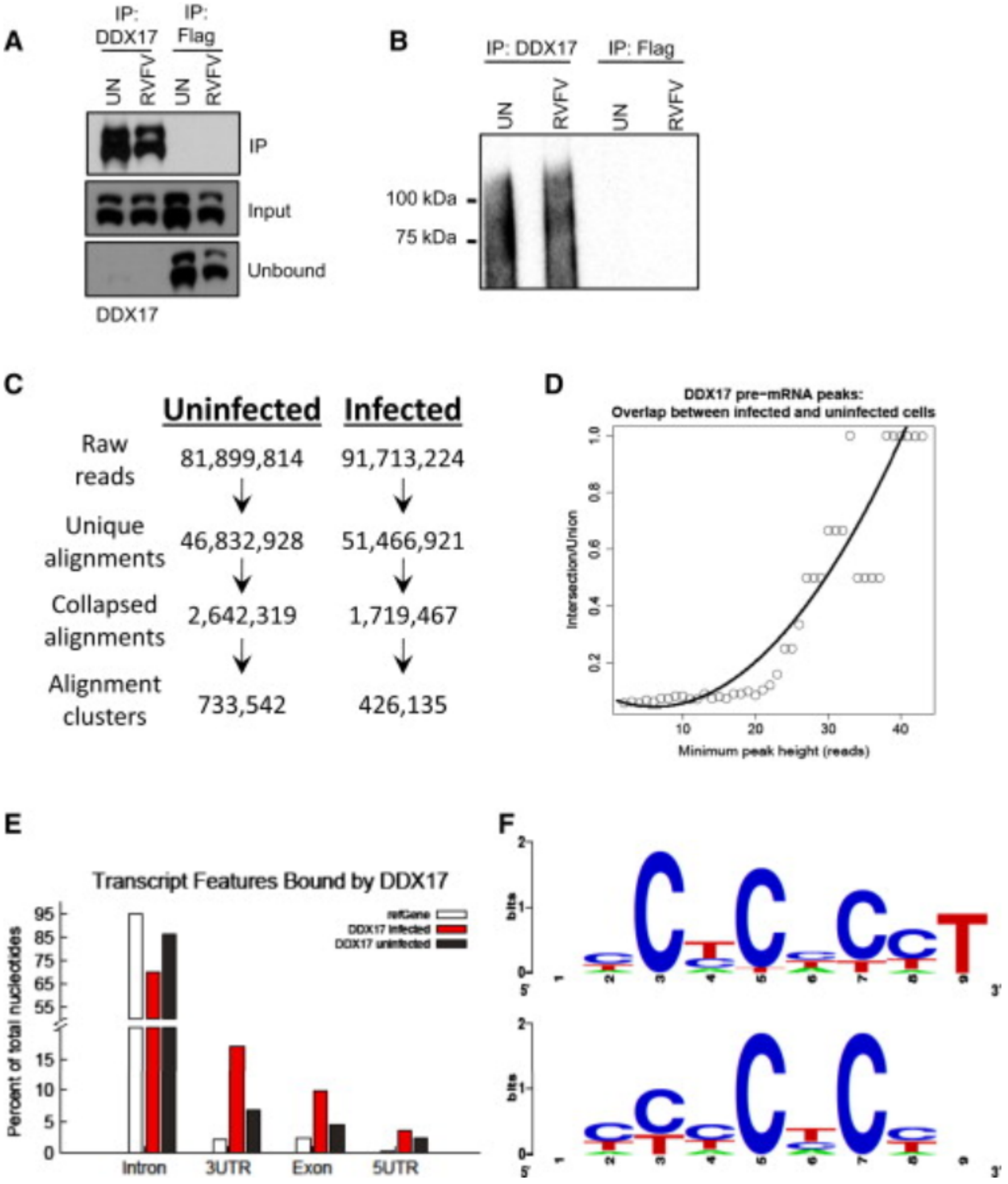


Figure A1.1. CLIP-Seq Analysis of DDX17-Bound RNAs from Uninfected and RVFV-Infected U2OS Cells. (a) Immunoblot of DDX17 from uninfected or RVFV-infected U2OS cells with immunoprecipitation (IP) using anti-DDX17 or anti-FLAG (control). Input, IP, and unbound fractions are shown, with high efficiency of DDX17 IP. (b) Autoradiograph of immunopurified and 32P-labeled DDX17-RNA complexes transferred to nitrocellulose membrane. Immunoprecipitation with anti-FLAG as a control shows high specificity of the DDX17-RNA signal. (c) Flowchart of CLIP-seq alignment and processing pipeline, resulting in alignment clusters. (d) Alignment clusters overlapping annotated regions of the genome (refSeq) were further searched for significant peaks, and the overlap between infected and uninfected DDX17 significant CLIP-seq peaks (FDR < 0.001) in protein-coding genes from refSeq at increasing peak height is plotted. R2 = 0.88. (e) Percentage of total nucleotides under significant CLIP-seq peaks within refSeq protein-coding genes broken down into transcript feature types extracted from refSeq. (f) Composite motif logo of the multiple sequence alignment of the 20 most enriched hexamers under significant CLIP-seq peaks within protein-coding genes as identified by Z score, comparing hexamer frequencies to 100 permutations of binding-site locations within bound transcripts for uninfected (top) or infected (bottom) cells.

From three pooled DDX17-CLIP experiments, ~80 million raw reads and ~90 million raw reads were obtained from uninfected and infected cells, respectively (Figure A1.1c). To process these DDX17 CLIP-seq reads, I first generated a composite genome index incorporating the hg19 human genome and three genomic segments of RVFV (L, M, S), with over 55% of reads aligning unambiguously to the composite genome (unique alignments). Collapsed alignments were obtained by removing PCR duplicates and retaining only one alignment for each 5′ coordinate. Genomic intervals with at least two overlapping alignments were clustered together generating the alignment clusters. This yielded 733,542 clusters for uninfected cells and 426,135 clusters from RVFV-infected cells. Alignment clusters within human pre-mRNA loci were further searched for significant peaks (false discovery rate [FDR] < 0.001) using an empirical algorithm[147]. This analytical approach which separates significant DDX17 binding sites in human pre-mRNAs (peaks) from potential interaction sites that are not within pre-mRNAs (alignment clusters) is required to identify DDX17-RNA interactions that occur outside of annotated transcripts, for instance to intergenic miRNA loci without annotated pri-miRNA transcripts or intracellular RVFV RNAs.

DDX17 pre-mRNA peaks showed strong overlap between uninfected and infected cells (Figure A1.1d), indicating that the overall profile of DDX17-bound cellular RNAs is similar during infection. Next, we determined the transcript features of DDX17 pre-mRNA peaks (Figure A1.1e). Interestingly, DDX17 peaks were enriched in coding exons, 5′ UTRs, and 3′ UTRs, suggesting that DDX17 preferentially binds mature mRNA. Hexamer enrichment analysis of CLIP-seq peaks within protein-coding genes showed a bias for CT- and CA-repeat elements (Figure A1.1f). Together, these data indicate both location and sequence preference for DDX17 binding to mRNAs.

To understand the functional targets of DDX17, I used DAVID to identify KEGG GO terms enriched among protein-coding genes associated with DDX17 CLIP-seq peaks. I observed enrichment for cell adhesion as well as several cellular signaling pathways (Figure A1.2a). Intriguingly, one of the most overrepresented KEGG pathways was mitogen-activated protein kinase (MAPK) signaling (Figure A1.2b). Previous data suggest that MAPK-activated protein kinase 2 (MK2) physically interacts with DDX5 to control its localization, and that DDX5/DDX17 regulate splicing of p38 MAPK[148,149,149]. Thus, DDX17-bound RNAs identified in our experiments overlap with known targets in MAPK signaling, suggesting that the CLIP-seq peaks reflect the biological activity of DDX17.

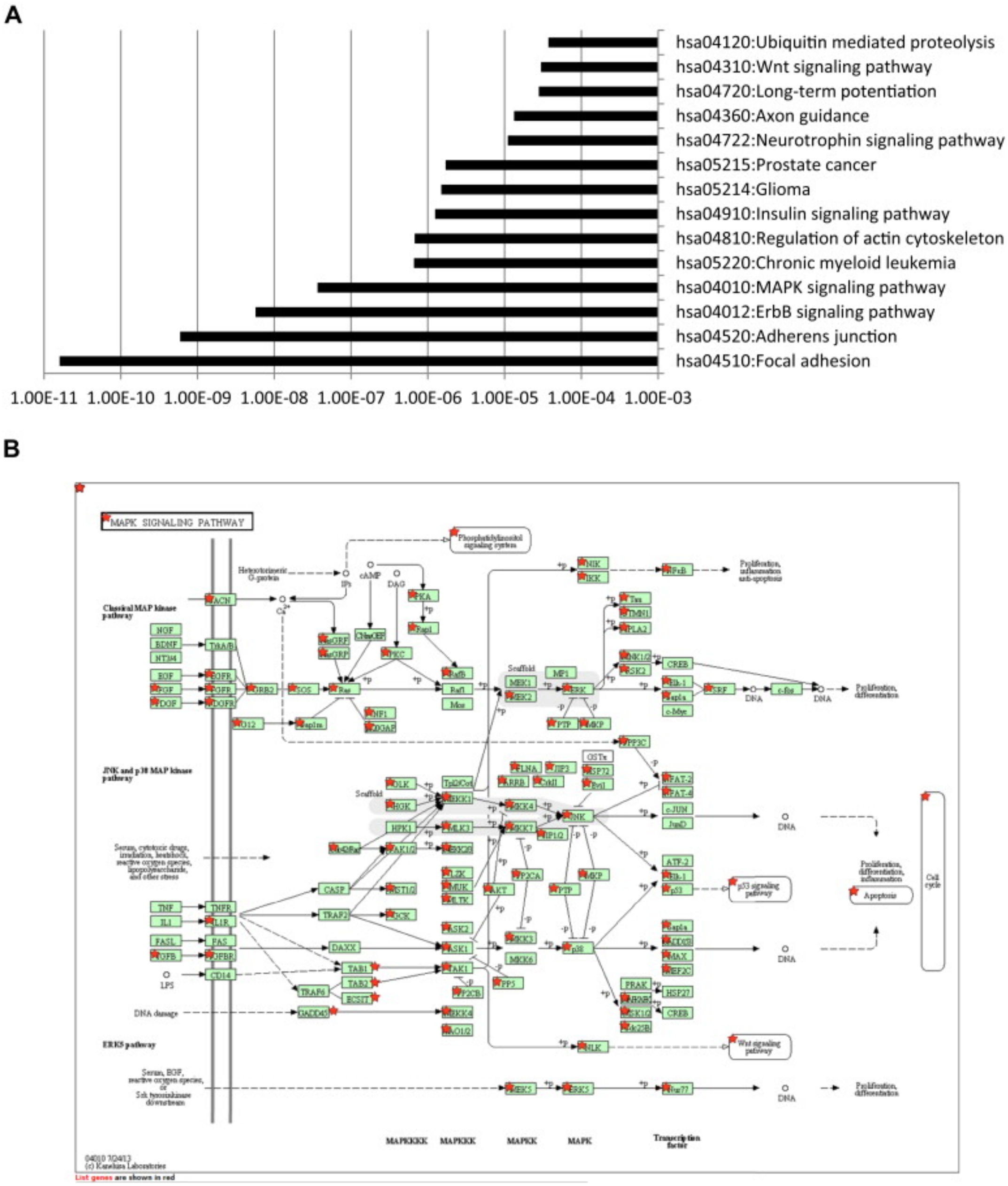


Figure A1.2. KEGG and GO Term and Pathway Analysis of DDX17-Bound RNAs. (a) Plot of p values for enriched KEGG GO terms using DAVID of protein-coding genes bound by significant DDX17 CLIP-seq peaks in either infected or uninfected cells. (b) KEGG pathway diagram of the MAPK signaling pathway genes intersecting significant DDX17 CLIP-seq peaks.

In addition to roles in transcriptional regulation and alternative splicing, DDX17 has been linked to miRNA biogenesis. DDX5 and DDX17 are components of the

Microprocessor complex, which processes the pri-miRNA transcript into the 60–70 nt stem-loop intermediate known as the pre-miRNA[150,151,151,152,152,153,153]. Loss of DDX17 results in decreased expression of a subset but not all miRNAs[152]. Therefore, as further validation of our CLIP-seq data, I also analyzed the intersection of DDX17 CLIP signal with annotated miRNA stem loops.

I observed 160 pri-miRNA loci that were associated with DDX17 CLIP clusters. There was strong correlation in normalized CLIP signal within pri-miRNAs from uninfected and RVFV-infected samples (Figure A1.3a), suggesting that similar pri-miRNAs are bound by DDX17 in uninfected and infected cells. In contrast, I found no correlation between CLIP-seq signal and level of miRNA expression reported in a previous study of small RNAs in U2OS cells, indicating that DDX17 clusters represent bias for certain miRNAs independent of expression level (Figure A1.3b). Among DDX17-bound miRNAs, miR-663a, miR-99b, and miR-6087 were some of the most highly represented miRNAs (Figure A1.3c). Analysis of DDX17 CLIP signal in relation to the predicted pri-miRNA stem loop showed that DDX17 clusters were preferentially localized immediately 5′ and 3′ to the center of the loop (Figure A1.3d). These data suggest that DDX17 interactions are strongest with the stem region of the miRNA hairpin rather than the loop. Analysis of overrepresented hexamers in DDX17-associated miRNAs did not show any enrichment of the CA- or CT-repeat elements found with the DDX17 mRNA peaks. Furthermore, *de novo* analysis of the bound pri-miRNAs identified no significantly enriched motifs compared to total pri-miRNA background. Thus, the interaction of DDX17 with pri-miRNAs is likely determined by RNA secondary structure.

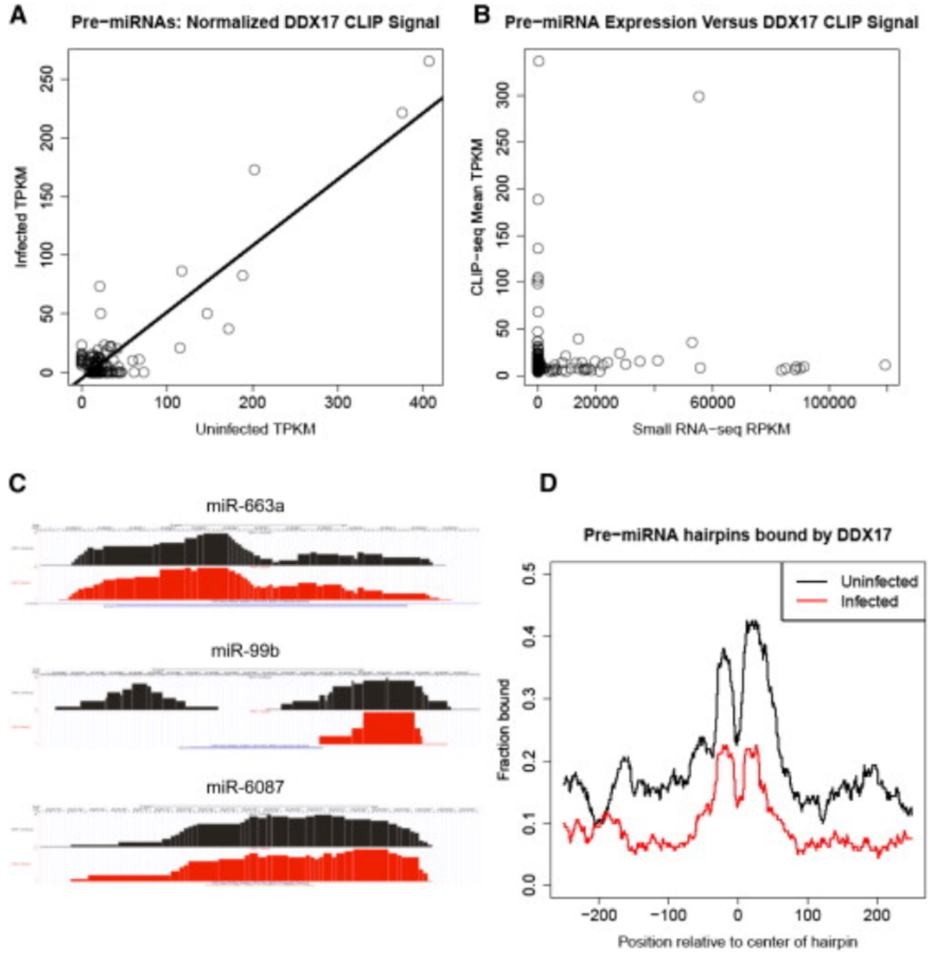


Figure A1.3. DDX17 Directly Binds miRNA Stem Loops in Human U2OS Cells. (a) Normalized CLIP-seq signal (TPKM, tags per kilobase of pre-miRNA per million CLIP-seq reads) in pre-miRNA hairpin loci with CLIP signal extracted from miRBase. Linear regression of infected TPKM on uninfected TPKM is plotted, R2 = 0.79. (b) Scatterplot of miRNAs that are bound; normalized pre-miRNA expression (RPKM) from small RNA-seq and the mean of normalized CLIP-seq signal (TPKM) between infected and uninfected U2OS cells are plotted, R2 = 0.001. (c) Alignment clusters overlapping miRBase pre-miRNA hairpin loci on the UCSC genome browser with uninfected cells colored black and infected cells colored red. (d) RNA map of DDX17 CLIP signal in pre-miRNA hairpins. Fraction of 160 hairpins bound is plotted at single-nucleotide resolution relative to the center of the stem loop.

To determine whether DDX17 regulation of miRNA biogenesis is directly involved in antiviral defense, Ryan Moy then silenced the Microprocessor component Drosha in U2OS cells. Loss of Drosha had no impact on RVFV replication, suggesting that the antiviral mechanism of DDX17 is independent of Drosha and the canonical miRNA pathway. Using luciferase reporter assays as previously described[154], Ryan also found that Rm62 is not required for siRNA- or miRNA-mediated silencing in Drosophila cells. These data indicate that DDX17 does not act through RNAi to restrict RVFV infection.

Next, I tested whether DDX17 directly interacts with viral RNA by analyzing the overlap of DDX17 CLIP clusters with the RVFV genome. I observed multiple DDX17 clusters, with the highest signal on the M and S segments (Figure A1.4a). These data suggest that DDX17 binds RVFV RNA in infected U2OS cells. In addition, DDX17 viral clustersdid not overlap with CA- and CT-repeat motifs, suggesting that DDX17-viral interactions are not dependent on these elements.

Because viral RNAs are often highly structured and DDX17 was enriched at the stem region of pri-miRNA hairpins, we hypothesized that DDX17 may recognize structured elements in RVFV RNAs. Indeed, we observed a prominent CLIP cluster within the intergenic region (IGR) on the S segment (between N and NSs). The IGR on other ambisense bunyaviruses has been shown to form a highly complementary sequence that folds into a hairpin to control transcription termination[155]. This IGR in the RVFV antigenome similarly forms a hairpin that generates the majority of virus-derived siRNAs in infected Drosophila and mosquito cells[156]. We defined a 75 nt RNA that overlaps the largest S segment DDX17 CLIP cluster within the IGR on the genome strand, which is predicted to form a hairpin structure that resembles miRNA stem loops (Figure A1.4b). Ryan Moy synthesized this RNA in vitro using T7 RNA polymerase to test whether it is bound by DDX17. Biotinylated DDX17 peak RVFV RNA efficiently precipitated DDX17 from U2OS cell lysates in a dose-dependent manner, demonstrating that DDX17 physically interacts with RVFV RNA and validating our CLIP-seq results

(Figure A1.4c). In contrast, a nonspecific control from RVFV RNA not bound in our CLIP-seq data set did not precipitate DDX17 (Figure A1.4d).

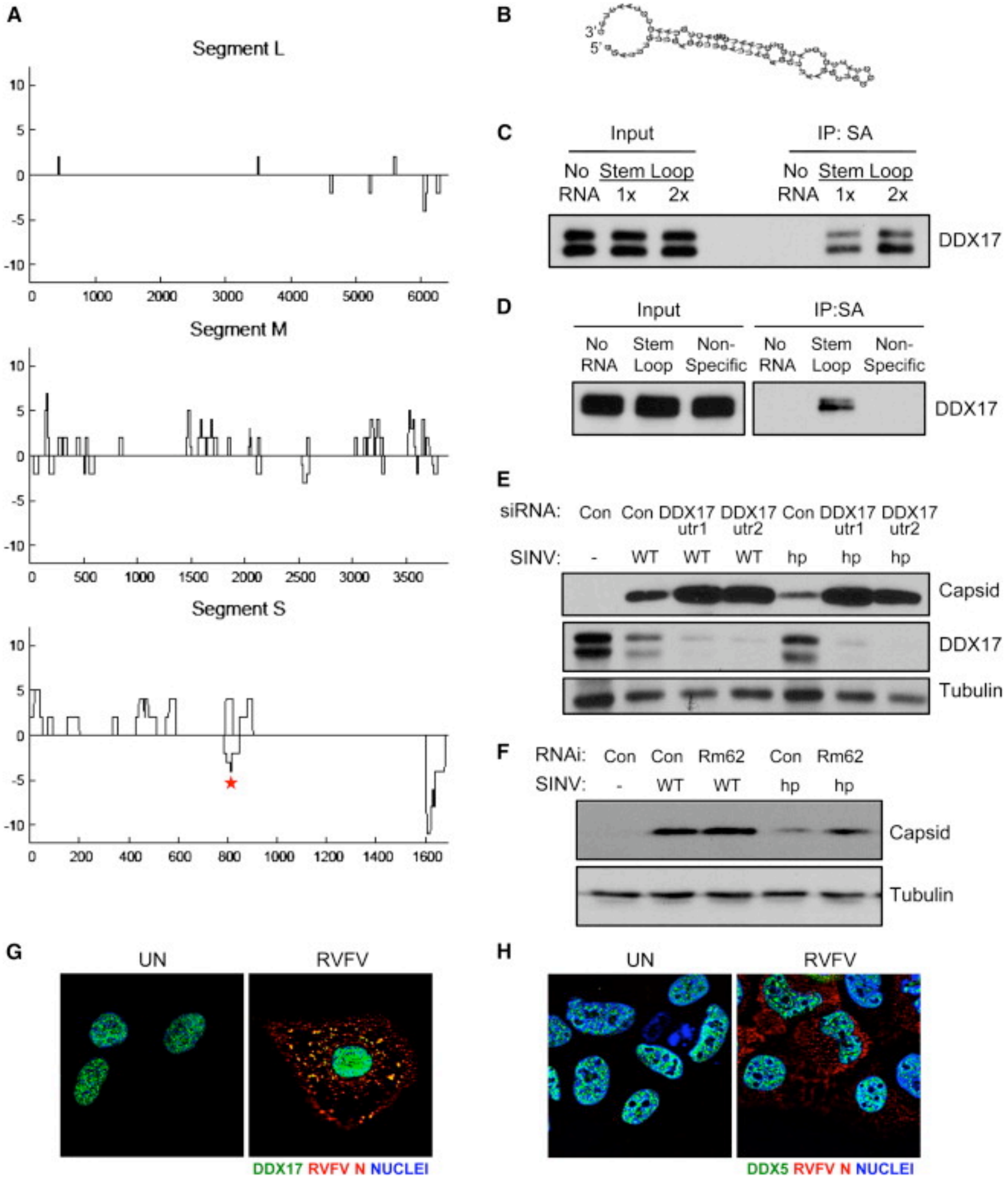


Figure A1.4. DDX17 Binds RVFV RNA to Restrict Viral Infection. (a) DDX17 CLIP-seq clusters aligned to the RVFV tripartite genome, plotted 3′ to 5′ (genome orientation) along the x

axis. Binding sites that map to the genome are below and to the antigenome are above the line. CLIP-seq signal intensity (black) is measured in total overlapping reads at each nucleotide position. (b) Predicted secondary structure of a 75 nt RNA from DDX17 CLIP peak on the RVFV S segment between N and NSs as determined by RNA fold (asterisk in A). (c) The 75 nt DDX17 CLIP peak RNA from (B) was synthesized by T7 in vitro transcription and biotinylated. Biotinylated RVFV RNA was incubated with U2OS cell protein lysates and immunoprecipitated, and DDX17-RVFV RNA complexes were analyzed by immunoblot. (d) RNA-protein interaction assays were performed as in (C) using the biotinylated RVFV stem loop and nonspecific control RNA from RVFV not bound in the DDX17 CLIP-seq data set. (e) Representative immunoblot of U2OS cells transfected with the indicated siRNAs and infected with SINV WT or SINV encoding the RVFV hairpin (SINV-hp) 8 hpi. (f) Representative immunoblot of Drosophila cells treated with control (β-gal) or Rm62 dsRNA and infected with SINV WT or SINV-hp 24 hpi (moi = 0.3). (g) Representative IF images of DDX17 and RVFV N from uninfected or infected U2OS cells 12 hpi (helicase, green; RVFV N,red; nuclei, blue). (h) Representative IF images of DDX5 and RVFV N from uninfected or infected U2OS cells 12 hpi (helicase, green; RVFV N, red; nuclei, blue). (All panels except for A and B courtesy of Ryan Moy.)

To determine whether DDX17 binding on viral RNA can directly restrict viral infection, the lab of Dr. Ben tenOever cloned the RVFV DDX17 hairpin into the 3′ UTR of SINV under the control of a subgenomic promoter (SINV-hp). This same strategy has been previously shown to tolerate the insertion of noncoding hairpin RNAs (Shapiro et al., 2010). We found that control cells supported substantially less infection of SINV-hp compared to wild-type (WT) SINV (Figure A1.4e). Furthermore, whereas depletion of DDX17 led to modest increases in SINV capsid production of WT virus, loss of DDX17 led to large increases in capsid production from SINV-hp virus (Figure A1.4e). In addition, we tested whether this RVFV hairpin also impacted SINV replication in Drosophila cells. WT SINV was unaffected by the loss of Rm62 (Figure A1.4f). Moreover, as we found in human cells, control RNAi-treated cells supported less infection of SINV-hp than WT SINV, and depletion of Rm62 led to a large increase in SINV capsid production from SINV-hp virus (Figure A1.4f). A second DDX17 peak at the 5′ end of the S genomic segment was also predicted to form a hairpin, and cloning this hairpin into SINV (SINV-5′hp) also sensitized the virus to DDX17 restriction in Drosophila and human cells. Together, these data demonstrate that the presence of a DDX17-

binding site on viral RNA is restrictive and that this repression can be alleviated by loss of DDX17 across hosts.

The Cherry lab next assessed the localization of DDX17 and DDX5 during infection by immunofluorescence, as RVFV and SINV RNA replication occur exclusively in the cytoplasm. RNAi was used to validate the specificity of these antibodies for immunofluorescence. As previously reported[157], DDX17 was found in the nucleus in uninfected cells (Figure A1.4g). At 12 hpi, however, we observed some DDX17 staining in cytosolic puncta that colocalized with RVFV nucleocapsid protein N, which coats viral RNA and facilitates replication (Figure A1.4g). In contrast, DDX5 remained in the nucleus in the presence and absence of infection (Figure A1.4h), suggesting a distinct localization pattern for DDX17. Collectively, these data suggest that DDX17 may gain access to cytosolic RVFV replication complexes during infection and bind viral RNA to antagonize viral replication.

Discussion

This study represents one of the first applications of CLIP-seq to study RNA helicase-RNA interactions. In this study, DDX17 was found to restrict replication of an RNA virus through direct interactions. Importantly, we demonstrate that CLIP-seq can be used to study interactions between a host RNA binding protein and the RNAs of an intracellular pathogen.

Much of the computational analysis presented in this appendix utilized core components that I had previously developed to analyze hnRNP L CLIP-seq data. One important difference is the alignment of reads. In this experiment, reads were aligned against a composite metagenome index (see Materials and Methods) containing the chromosomes of the human genome plus the three RNA segments of the Rift Valley Fever Virus genome. This alignment strategy was developed to most closely recapitulate the available RNA substrates for DDX17 interactions within the cell. Notably, this alignment strategy was critical to allow discovery of the DDX17-RVFV interaction site that, when cloned into a virus that is not restricted by DDX17, confers restrictivity. The biological relevance of this interaction site therefore underscores the importance of the metagenomic index technique for host-pathogen studies using CLIP-seq.

As expected based on prior studies of DDX17-miRNA interactions, a subset of expressed miRNAs were targets of DDX17 interactions in the U2OS cells under study. Interestingly, DDX17 appears to engage the double-stranded regions of the miRNAs preferentially over the loop region (figure A1.3d), a finding consistent with DDX17's RNA helicase activity in the human miRNA biogenesis pathway. This finding suggests that the data from this study could be useful to the scientific community beyond the context of virology and intracellular immunology. Indeed, recent studies have highlighted DEAD box RNA helicases, including DDX17, in tumorigenesis and tumor migration.

In analyzing the human pre-mRNA features that are occupied by DDX17 CLIP-seq peaks, I observed an increase in 3'UTR occupancy upon infection with RVFV (figure A1.1e). This finding is particularly interesting because the functional significance of DDX17-3'UTR interactions is unknown. Two hypotheses might explain this phenomenon. First, infection with RVFV could alter DDX17-mRNA interactions such that an increase in DDX17-3'UTR interactions results. Alternatively, infection with RVFV could alter the expression of genes and/or the length of 3'UTRs such that a broader expanse of DDX17-3'UTR interaction sites is made available. While this study utilized mRNA-seq data from U2OS cells, only uninfected cell data was available, and investigating the potential for RVFV-dependent alterations in the host cell transcriptome is not currently possible. However, this finding presents an exciting opportunity for further study.

In conclusion, this study represents a fruitful collaboration between the Lynch and Cherry labs that allowed both labs to broaden their scientific horizons. In the Conclusion chapter of this thesis, I synthesize the CLIP-seq analyses of hnRNP L, DDX17, and CELF2.

APPENDIX 2: CLIP-SEQ ANALYSIS OF CELF2-RNA INTERACTIONS IN T CELLS

Introduction

We have previously demonstrated the use of CLIP-seq to analyze protein-RNA interactions for hnRNP L and DDX17. While hnRNP L expression and activity were found to be consistent between unstimulated and stimulated T cells, our lab has previously described a broad pattern of stimulation-induced alternative splicing in T cells using transcriptome sequencing[107], suggesting that other splicing factors may respond to T cell stimulation with altered activity, thus controlling the stimulation-responsive splicing phenotype. In support of this, I previously performed motif enrichment analysis of stimulation-responsive exons and found multiple sequence features to be enriched[107], suggestive of a network of splicing factors that control activation-induced alternative splicing. However, the contribution of individual splicing factors to T cell stimulation-induced alternative splicing is largely unknown.

One example of activation-induced alternative splicing that our lab has studied using biochemical techniques is LEF1 exon 6[158]. LEF1 is a transcription factor that drives expression of the T cell receptor (TCR) alpha subunit[159]. Upon T cell stimulation, LEF1 exon 6 inclusion is increased. Importantly, this exon encodes a portion of the context-regulatory domain (CRD) of LEF1 that is required for optimal gene-expression enhancer activity[160]. Our lab has previously demonstrated that the stimulation-induced alternative splicing of LEF1 exon 6 is driven by the splicing factor CELF2[158]. CELF2 binds to UG-rich, evolutionarily conserved sequences on both sides of LEF1 exon 6.

CELF2 expression increases upon stimulation, resulting in higher levels of CELF2 binding in these periexonic sequences. Importantly, CELF2 is an enhancer of LEF1 exon 6 inclusion because knockdown of CELF2 reduced exon inclusion.

These studies of LEF1 exon 6 suggest that CELF2 is one splicing factor that links T cell stimulation to alternative splicing. However, the transcriptome-wide occupancy of CELF2 is unknown. While the study of CELF2 is the chief focus of other students in the lab, this project provided the basis for another productive collaboration in which I was able to utilize software tools developed for hnRNP L analysis on a new protein, and one with expression and activity that are altered by T cell stimulation.

Importantly, the post-doc in our lab that performed the CLIP-seq library preparations for hnRNP L and DDX17 also prepared CLIP-seq libraries for CELF2 in unstimulated and stimulated JSL1 cells. The resulting CLIP-seq reads were processed by the same computational pipeline that was developed for hnRNP L and DDX17, providing an ideal opportunity to compare the results of these three CLIP-seq experiments and derive insights into common features of the CLIP-seq findings and also unique, distinguishing characteristics of each study.

Results

To identify transcriptome-wide CELF2-RNA interactions in T cells, Ganesh Shankarling prepared CLIP-seq libraries from unstimulated and stimulated JSL1 cells. Importantly, these conditions are identical to those utilized in the CLIP-seq analysis of hnRNP L. The library preparation was identical to that utilized in the CLIP-seq analysis of hnRNP L and DDX17 except for one modification. One key step in the computational analysis of CLIP-seq reads is to remove PCR duplicates. When more than one CLIP-seq read aligns to the same genomic locus, it is not possible to discern whether multiple oligonucleotide fragments were present in the immunoprecipitated RNA or whether a single fragment gave rise to multiple reads by PCR duplication. For this reason, duplicates are entirely removed and only one alignment is allowed at each position in the genome. This removal of PCR duplicates results in considerable reduction of the size of the aligned reads relative to the size of the raw reads.

To compensate for this, Ganesh employed a barcoding strategy which allows discrimination of PCR duplicates from multiple distinct RNA fragments. Before the PCR reactions are performed, each CLIP sample is split into three aliquots and unique hexanucleotide barcodes are ligated to each (figure A2.1). After sequencing of the resulting libraries, each of the three barcoded aliquots from the same CLIP sample are aligned separately to the genome and PCR duplicates are removed. Then, the three aliquots from each sample are combined, allowing a maximum of three reads to align to the same genomic position within each CLIP sample. This strategy aims to increase the sensitivity of the CLIP-seq peak caller.

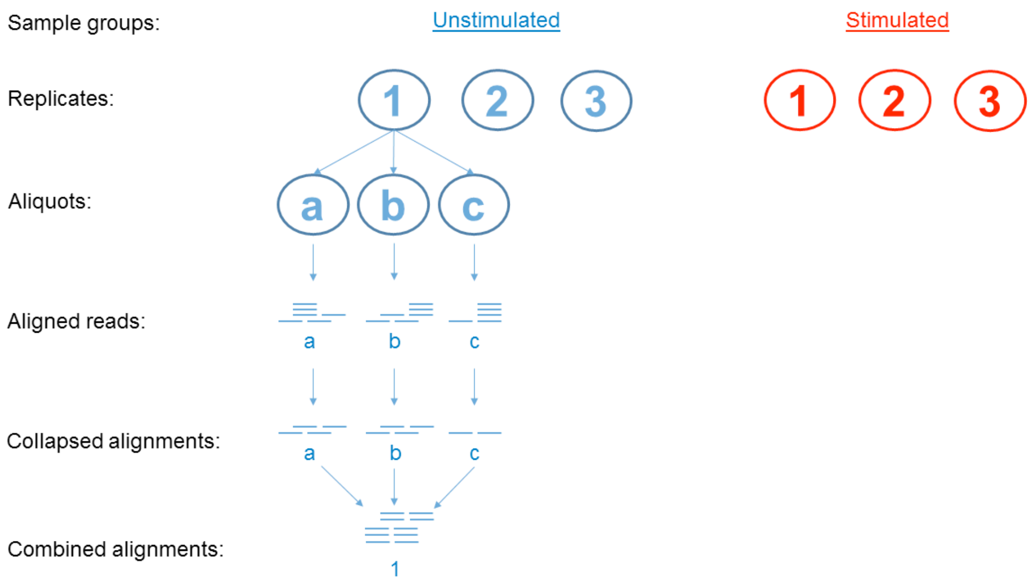


Figure A2.1. Triplicate barcoding strategy for CELF2 CLIP-seq. Unstimulated and stimulated sample groups of JSL1 cells were subjected to UV crosslinking (top) in triplicate. After immunoprecipitation of CELF2-RNA complexes and isolation of fragmented RNAs, each replicate was split into three identical aliquots before PCR amplification (only one sample is diagrammed). Each aliquot was ligated to a different oligonucleotide barcode. Reads from each uniquely-barcoded aliquot were separately aligned to the genome and PCR duplicates removed, generating "collapsed alignments" (middle). Finally, each aliquot's unique alignments were combined to recreate the individual replicates, a process which allows a maximum of three reads to align to the same genomic position (bottom).

In total, 277 million raw reads were generated, from which 122 million were mapped to the human genome (figure A2.2a). Removal of PCR duplicates left 7.8 million alignments remaining, suggesting a high degree of duplication in the aligned reads. Importantly, although only a small fraction of the human genome is contained within the refSeq annotation, almost all CELF2 CLIP-seq reads aligned to portions of the human genome contained within the refSeq annotations. Subsequent combining of the aligned reads (as demonstrated in Figure A2.1) allowed me to identify 49,962 significant CELF2 peaks in unstimulated JSL1 cells and 52,249 peaks in stimulated cells. These peaks were subjected to further analysis using software tools developed for hnRNP L CLIP-seq.

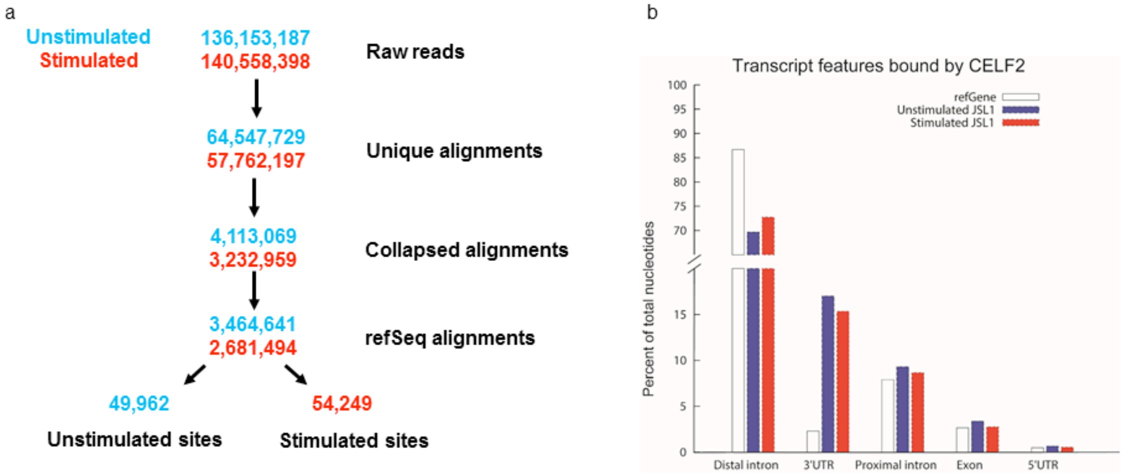


Figure A2.2. Summary of CELF2 CLIP-seq analysis. a) Total counts of raw reads, unique alignments (reads that aligned to one and only one position in the hg19 build of the human genome), collapsed alignments (unique alignments that have had PCR duplicates removed), and collapsed alignments within refSeq transcriptome annotation are displayed. Total numbers of CLIP-seq peaks identified by an identical algorithm to that used in the analysis of hnRNP L CLIP-seq experiments are also displayed (see Materials and Methods). b) Barplot of the fraction of nucleotides covered by CELF2 CLIP-seq peaks in unstimulated and stimulated JSL1 T cells occupying each of five categories of annotation.

First, I examined the fraction of the total genomic footprint covered by CELF2 CLIP-seq peaks is annotated as 5'UTR, 3'UTR, exon, proximal intron (within 300nt of an exon), or distal intron. For this analysis, I used the refSeq transcriptome annotation, as was performed for hnRNP L CLIP-seq analysis. Importantly, the CLIP-seq peak caller only searches refSeq transcripts for peaks, so every nucleotide covered by CELF2 CLIP-seq peaks may be uniquely classified under these five categories. By comparing the fraction of nucleotides under CELF2 CLIP-seq peaks within each of these five categories to the fraction of the total refSeq transcriptome annotation that is composed of each category, a relative enrichment for 3'UTR interactions is evident (figure A2.2b). This enrichment of CELF2 for 3'UTR interactions is potential evidence of a broader role for CELF2 in 3' end processing or splicing within 3' UTRs. Interestingly, CELF1, a related RNA binding protein from the same Elav-like factor family (CELF family), binds to GU-rich elements in the 3'UTRs of human mRNAs to trigger mRNA decay[161]. A recent CLIP-seq analysis of CELF1 in mouse cardiac tissue similarly identified enrichment of CELF1 CLIP-seq peaks within 3'UTRs[162]. Taken together, these results suggest that CELF2, similar to CELF1, is enriched for 3'UTR interactions, and suggests a possible interplay between CELF2 and mRNA stability. This is particularly interesting given that CELF2 is strictly nuclear in localization, while CELF1 resides in both cytoplasmic and nuclear protein fractions in JSL1 cells.

To identify enriched motifs within CELF2 CLIP-seq peaks, I used the Z-score motif enrichment algorithm directed at *k*mers of length 6 (hexamers). Histograms of Z scores for hexamer enrichment display a long right tail, indicating a subpopulation of hexamers are enriched relative to permuted background (figure A2.3). I used multiple sequence alignment to generate a motif logo of the top 20 most enriched hexamers (inset). The resulting motifs display a marked bias toward UGU trinucleotides in both unstimulated (figure A2.3a) and stimulated (figure A2.3b) CLIP-seq peaks. This motif preference is in agreement with the motif preference of CELF1 as identified by SELEX[163], providing further evidence of the similarity between CELF1 and CELF2 at the level of motif preference. Interestingly, the motif preference of CELF2 is not drastically altered by cell stimulation, despite the fact that the splicing regulatory activity of CELF2 changes upon stimulation, as is the case with LEF1 exon 6[158].

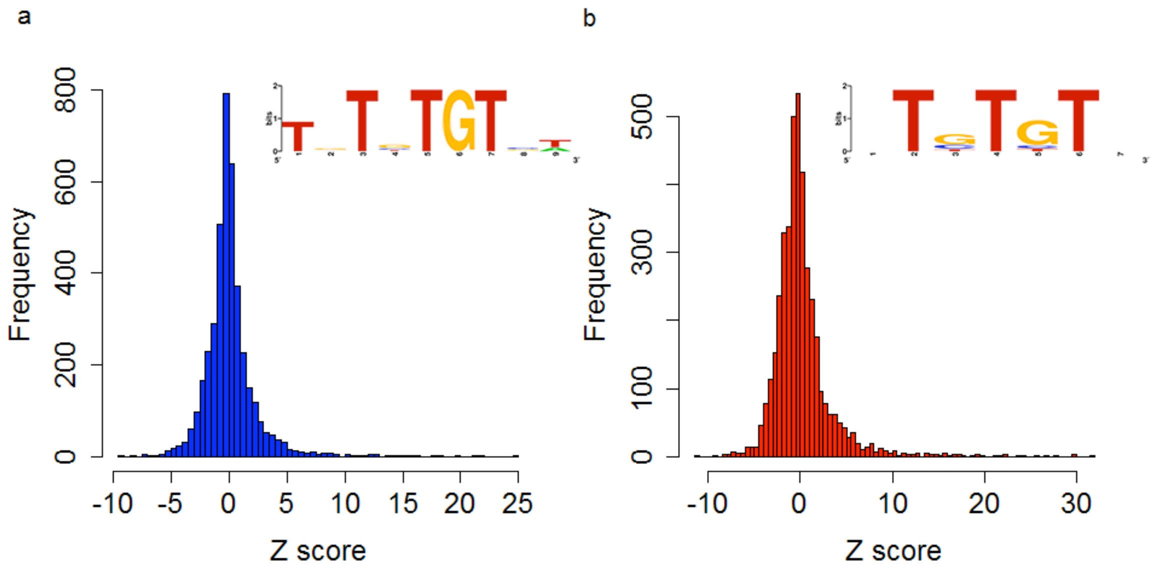


Figure A2.3. CELF2 CLIP-seq motif enrichment analysis. Unstimulated (a) and stimulated (b) CELF2 CLIP-seq peaks were permuted 100 times within the refSeq transcripts to which they align and hexamer frequencies within the actual CLIP-seq peaks were compared to the mean and standard deviation for that hexamer across the 100 iterations of independent permutations. The Z-score is reported as the number of standard deviations away from the mean permuted frequency, with positive values denoting enrichment and negative values depletion. Inset: a composite motif logo generated from multiple sequence alignment of the top 20 hexamers by Z score.

Finally, other individuals in our lab have performed functional studies of CELF2 splicing regulation using RASL-seq. While these results are not detailed here, one interesting hypothesis that arose from the results of CELF2 functional studies is the possibility of a regulatory interplay between CELF2 and RBFOX, another splicing regulator. One hypothesis that these results raised is that CEFL2 and RBFOX coregulate splicing by colocalization. To explore this hypothesis further, I examined the extent and the significance of overlap between CELF2 and RBFOX CLIP-seq sites. For this analysis, I used RBFOX CLIP-seq samples from mouse brain tissue. To control for differences in the software used to process RBFOX CLIP-seq data, I reprocessed the data from this study with the same pipeline used to process hnRNP L and CELF2 CLIP-seq experiments.

To compare the overlap between CELF2 and RBFOX, I first computed the overlap between CELF2 CLIP-seq peaks in unstimulated and stimulated JSL1 cells (table A2.1). More than 62% of CELF2 CLIP-seq peaks in unstimulated JSL1 cells have some degree of overlap with CELF2 CLIP-seq peaks in stimulated cells. To assess the significance of this overlap, I permuted the unstimulated CELF2 CLIP-seq peaks within the transcripts that contain them. This permutation process estimates the degree of overlap between two sets of CLIP-seq peaks due to random chance, given the size and number of CLIP-seq peaks within each transcript that was a physical target of the protein under study. When unstimulated CELF2 JSL1 peaks are permuted in this manner, the fraction of peaks that overlap stimulated CELF2 peaks falls below 5%. As a negative control, I examined the overlap between CELF2 and hnRNP L and hnRNP A1 CLIP-seq peaks, for which I have no evidence of a functional relationship. The degree of overlap between CELF2 and hnRNP L is less than 4%, and the overlap between CELF2 and hnRNP A1 is not more than 1%, indicating that only a small fraction of CELF2-RNA interaction sites are also physical targets of these two hnRNP proteins.

Finally, I compared the overlap between CELF2, hnRNP A1, hnRNP L, and RBFOX. Despite the fact that RBFOX CLIP-seq was performed in a different tissue and in a different organism, there is a greater than 3x higher overlap between CELF2 and RBFOX CLIP-seq peaks than expected by randomization. Notably, this is not true for the overlap between hnRNP L and RBFOX or for the overlap between hnRNP A1 and RBFOX. While the total degree of overlap between CELF2 and RBFOX is low, at least some of this might be attributable to the imperfect conversion between the mouse genome coordinates to human genome coordinates (liftOver). Additionally, the fact that the species and the tissue are both different begs a large measure of caution when interpreting these results.

	Unstimulated CELF2 JSL1		Stimulated CELF2 JSL1		hnRNP A1		Mouse Brain RBFOX	
	Percent overlap:	Percent randomized overlap:	Percent overlap:	Percent randomized overlap:	Percent overlap:	Percent randomized overlap:	Percent overlap:	Percent randomized overlap:
hnRNP L	3.24	2.20	2.86	2.02	0.17	0.12	0.77	0.62
Unstimulated CELF2 JSL1			62.41	4.75	1.00	0.32	3.65	1.11
Stimulated CELF2 JSL1					0.42	0.22	3.09	0.96
hnRNP A1							2.33	1.25

Table A2.1. CELF2 CLIP-seq overlap matrix. The fraction of CELF2 CLIP-seq peaks in unstimulated and stimulated JSL1 cells was compared to a panel of other CLIP-seq studies. Each CLIP-seq study was reprocessed by the identical pipeline as used to generate CELF2 peaks. One of the two sets of CLIP-seq peaks was then permuted within the transcripts in which they occur and the fraction of these permuted peaks that overlap the other set of CLIP-seq peaks (e.g. CELF2 versus hnRNP A1) was computed.

Discussion

CELF2 is a splicing regulator that has differential regulatory function in unstimulated and stimulated JSL1 cells. Because of this, CELF2 might play a broad role in reshaping the transcriptome upon stimulation. The mechanisms by which signal-inducible alternative splicing events are regulated remain unknown.

Here, I present an analysis of CELF2-RNA interactions using CLIP-seq. While the functional impact of CELF2 on the T cell transcriptome is the subject of study for other individuals in the lab, Ganesh Shankarling prepared CELF2 CLIP-seq libraries in unstimulated and stimulated cells using a similar approach to that employed for hnRNP L CLIP-seq library preparation. This provides an important level of experimental control when comparing hnRNP L and CELF2 CLIP-seq datasets. However, one modification was added to the library preparation protocol: multiplexed barcoding within individual samples (figure A2.1). To provide the ability to detect multicopy CLIP fragments within the background of PCR duplication, Ganesh split each CLIP sample into three aliquots and ligated a unique barcode onto each aliquot. This allowed me to retain up to three copies of alignments that map to the same genomic position because the finding of these three copies cannot be attributable to PCR duplication.

Motif enrichment analysis of CELF2 identifies UGU trinucleotide repeats. Notably, the most enriched motifs in unstimulated CELF2 CLIP-seq peaks are visibly similar to those in stimulated CELF2 CLIP-seq peaks, suggesting that cell stimulation does not alter the sequence specificity of this RNA binding protein. One possible interpretation of this result is that the difference in CELF2's splicing regulatory function that results as a consequence of cell stimulation is not attributable to a difference in RNA recognition; instead, in both cell states, CELF2 engages UG-rich and UGU-containing sequences.

Another hypothesis for how CELF2 can have differential function in unstimulated and stimulated cells is an alteration in the RNA sites occupied by CELF2 that is attributable to differences in CELF2 protein expression. For instance, CELF2 mRNA and protein levels are higher in stimulated cells than in unstimulated cells, and this could result in an increase in the number of sites bound by CELF2 in stimulated cells. However, an analysis of the degree of overlap between CLIP-seq peaks in unstimulated and stimulated cells found that a majority of the CLIP-seq peaks overlap between the two conditions. Notably, this overlap is more than 13 times higher than that expected by randomization. This, combined with the finding of similar total numbers of CELF2 CLIP-seq peaks between the two conditions, suggests that the majority of CELF2-RNA interaction sites are physical targets in both unstimulated and stimulated cells.

These results motivate other hypotheses that might explain the differences in CELF2 splicing regulatory function between unstimulated and stimulated cells. One possibility is that CELF2 is differentially modified at the protein level between unstimulated and stimulated cells and that these modifications can account for the functional differences. While CELF2 largely engages the same RNA sites in both cell conditions, the consequences of that engagement may be altered by post-translational modifications. For example, CELF2 might recruit snRNP proteins or splicing coactivators in unstimulated cells, but inhibit snRNP/coactivator recruitment in stimulated cells due to post-translational modifications. Other students in the lab have utilized mass spectrometry to investigate this interesting hypothesis.

Finally, I have demonstrated the use of a permutation algorithm to estimate the significance of the degree of overlap between two CLIP-seq samples. Indeed, the overlap between CELF2 and RBFOX CLIP-seq sites is three times higher than that expected based on randomization. This analysis raises the hypothesis that CELF2 and RBFOX could establish functional interplay by interacting with the same or neighboring RNA sites. This hypothesis is particularly interesting because a recent analysis of CELF1 splicing regulation in muscle tissue identified a functional interplay between CELF1 and MBNL1, another splicing regulatory protein. In this case, CELF1 and MBNL1 are mutually antagonistic and interact with neighboring or overlapping RNA sites. This style of coregulation could be at play in JSL1 cells in the case of the functional interplay between CELF2 and RBFOX.

In sum, CLIP-seq analysis of CELF2 raises several functional hypotheses about the stimulation-responsive alterations in splicing regulation. These results provide a foundation for further biochemical studies.

CONCLUSION

The advent of high-throughput genetic sequencing technology has ushered in the post-genomic era. One of the first major discoveries that accompanied informatic studies of the first draft of the human genome was a puzzling paucity of gene count relative to protein count. Early transcriptome annotations identified ~25,000 recognizable genes within the human genome, and while subsequent efforts have expanded our appreciation of the coding potential of the human genome, including species such as lincRNAs and upstream antisense transcripts, the human proteome appears to be contain many more proteins than the number of genes in the human genome could directly account for.

One of the mechanisms by which humans generate rich proteomic complexity from the human genome is by alternative pre-mRNA processing such as alternative splicing. High-throughput sequencing studies of expressed transcripts in a diversity of human cell types has expanded our appreciation of the degree and extent to which human pre-mRNAs are alternatively processed to generate multiple isoforms with distinct coding potential. Alternative pre-mRNA processing is also subject to control by development and intercellular signaling, adding a dynamic dimension to the coding potential of the human genome.

Alternative pre-mRNA processing is regulated by the combinatorial activities of cis- and trans-acting features which work in concert to dictate the processing of a pre-mRNA. While biochemical studies have provided mechanistic insights into alternative splicing events at the single-gene level, the extent to which these insights apply to the rest of the transcriptome is unclear. In this thesis, I demonstrate the use of integrative genomics to study alternative splicing on the transcriptome-wide scale. In particular, I utilize CLIP-seq and bioinformatics analysis to identify hnRNP L-RNA interactions in human T lymphocytes. I then use RNA sequencing to identify alternative splicing events that respond to hnRNP L depletion. Finally, I use integrative genomic analysis to examine the pattern of features associated with hnRNP L-responsive splicing.

CLIP-seq identifies hnRNP L-RNA interactions in JSL1 and CD4+ human T lymphocytes

Our lab has previously used biochemical methods and molecular biology to study hnRNP L-responsive alternative splicing events in T cells. Importantly, our lab has identified instances in which hnRNP L interacts directly with exonic RNA to repress exon inclusion. Splice site strength and co-associated proteins impact this regulation, suggesting a combinatorial code or splicing control by hnRNP L. However, the extent to which the conclusions of these studies generalize to the rest of the T cell transcriptome is unclear. To assess the transcriptomic impact of hnRNP L, we first utilized CLIP-seq to identify hnRNP L-RNA interaction sites. The identification of hnRNP L-RNA interaction sites is a crucial component of the transcriptome-wide analysis of hnRNP L splicing regulatory function.

Ganesh Shankarling prepared hnRNP L CLIP-seq libraries in unstimulated and stimulated cells of two types. First, he utilized JSL1 cells, a monoclonal, Jurkat-derived, immortalized T lymphocyte cell line. These cells have been utilized by our lab and others to study alternative splicing in cultured cells. Additionally, he prepared hnRNP L CLIP-seq libraries using primary human CD4+ T cells purified from peripheral blood mononuclear cells. The addition of these primary human CD4+ samples allows us to compare findings from JSL1 cells and assess the extent to which experimental results from immortalized cells may generalize to primary cells. Second, unstimulated and stimulated conditions from both types of cells were utilized. While there is no prior evidence from our lab or others that the activity of hnRNP L responds to cell stimulation, the inclusion of these two physiologic conditions provides another layer of biological replication because the transcriptome is altered at the expression and processing level by stimulation[107]. The inclusion of two cell conditions in addition to two cell types therefore expands the experimental conditions of this study, adding breadth to our experimental and analytical results and conclusions.

Using bioinformatic analysis of hnRNP L CLIP-seq reads, I identified over 100,000 unique interaction sites (peaks). Importantly, I found a high degree of overlap between hnRNP L peaks identified from unstimulated and stimulated cells, suggesting that the majority of hnRNP L interactions are not altered by cell stimulation. This finding is consistent with prior biochemical studies of hnRNP L-regulated alternative splicing events in which the splicing control by hnRNP L is not altered by cell stimulation, and also consistent with the fact that hnRNP L protein levels and nucleocytoplasmic localization are not altered by stimulation. I also identified a high degree of overlap between hnRNP L-RNA interaction sites identified in JSL1 cells and those identified in primary human CD4+ cells, which supports the JSL1 cell line as a model for studying hnRNP L splicing control.

Consistent with *in vitro* studies of the consensus sequence for hnRNP L-RNA interactions, I identified a CA-rich motif as most enriched within hnRNP L CLIP-seq peaks. Notably, the two most strongly enriched hexamers within hnRNP L CLIP-seq peaks were CACACA and ACACAC, the two pure CA or AC dinucleotide repeat hexamer sequences. This motif preference is consistent between T cell types and stimulation states.

Importantly, the CLIP-seq peaks were used to identify novel instances of hnRNP L-regulated alternative splicing. I first identified exons with hnRNP L peaks within the exon or within the upstream or downstream introns. Ganesh and I, along with others in the lab, then used hnRNP L-depleted or mock-depleted RNA extracts to quantify splicing changes by RT-PCR. This approach identified 27 cases of hnRNP L-regulated splicing events. These studies of hnRNP L-RNA interaction sites form the foundation for functional genomics studies of hnRNP L splicing control.

High-throughput sequencing approaches identify hnRNP L-responsive alternative splicing events in JSL1 T cells

Having identified hnRNP L-RNA interaction sites using CLIP-seq, I next studied the splicing regulatory role of hnRNP L in JSL1 T cells using genomics approaches. Using separate knockdown techniques, I depleted hnRNP L to approximately 50% of normal levels. These dual knockdown methods provide an important control for off-target effects and also provide a degree of experimental redundancy. I then utilized complementary high-throughput sequencing techniques to identify hnRNP L-responsive splicing events. I combined the high depth-of-coverage of RASL-seq with the high breadth-of-coverage of RNA-seq, resulting in high correlation between the splicing predictions generated by both experiments. This approach increased the positive predictive value of splicing predictions when compared to either experiment alone.

I identified hundreds of exons with inclusion level changes that respond significantly to hnRNP L depletion. Through gene ontology analysis, I found that these targets are enriched for RNA and DNA-binding functional categories, as well as chromatin modifiers and other epigenetic factors. This finding suggests an interplay between hnRNP L and chromatin. An additional analysis of other types of alternative splicing, such as intron retention and alternative 5' or 3' splice site utilization, produced a much smaller volume of predictions, suggesting that the primary splicing regulatory activity of hnRNP L is cassette-type alternative exons.

A gene expression change analysis that successfully captured stimulation-responsive upregulation of immune response genes identified very few hnRNP L-responsive differential gene expression events. One possible explanation for the inability of the RNA-seq data to detect hnRNP L-responsive gene expression changes that validate by qRT-PCR is the timecourse of the study. The RNA-seq experiment utilized transient transfection of antisense morpholino oligonucleotides to deplete hnRNP L. For this reason, RNA extracts can be generated between 24 and 60 hours, before which cells have not yet recovered from electroporation, but after which the extent of knockdown begins to subside as hnRNP L protein expression returns to normal levels. It is possible that gene expression changes induced by hnRNP L depletion take longer to manifest because hnRNP L regulates splicing of transcripts that encode transcription factors and epigenetic regulators. In order for the function of these proteins to be altered as a consequence of hnRNP L depletion, the isoform shift induced by hnRNP L depletion requires translation into protein in order to generate a shift in the proteome. In contrast, cell stimulation results in intracellular signaling that results in rapid post-translational modifications on transcription factors. This process could result in gene expression changes that are observable by RNA-seq during the timecourse utilized by this study. These differences in the mechanisms by which signaling and splicing result in changes in gene expression could explain the lack of hnRNP L-responsive gene expression events in comparison to stimulation-responsive gene expression changes.

Another hypothesis for the lack of hnRNP L-responsive gene expression events even though hnRNP L regulates splicing in transcription factors and epigenetic regulators is that the consequence of this hnRNP L splicing control is not gene expression change. Recent studies have highlighted the interplay between DNA binding proteins such as CTCF and splicing. Additionally, the role that chromatin state plays in determining splicing outcomes has become appreciated. Taken together, these studies suggest that DNA and RNA existing in a regulatory interplay. These transcriptome-wide functional studies of the control of RNA by hnRNP L form the foundation for integrative genomics analysis.

Integrative genomics analysis identifies direct repressive and indirect enhancing roles for hnRNP L in exon splicing

The above analyses have generated transcriptome-wide datasets of hnRNP L binding and function. One key question in the study of hnRNP L splicing control is how the position of binding relates to splicing regulatory outcome. Integrative genomics analyses often involve computational analysis of RNA binding protein occupancy within and around regulated alternative splicing events. The resulting visualizations, sometimes referred to as RNA maps, can reveal insights into the positions of binding associated with repression and enhancement of splicing. However, this approach suffers from several major caveats. First, RNA maps are empirical analyses and any visible trends are associative, not causal. Second, evidence suggests a combinatorial code of splicing control, but features like splice site strengths, sequence composition, exon/intron lengths, exon/intron ordinality, coassociated proteins, and secondary structure are often not included in RNA maps. Finally, the overlap between binding function is incomplete: CLIP-seq identified over 100,000 hnRNP L peaks in pre-mRNAs, but not all of these peaks have hnRNP L-responsive alternative splicing events nearby. Similarly, not all hnRNP L-responsive alternative splicing events are expected to have CLIP-seq peaks nearby, based on integrative genomics studies of other hnRNP proteins.

Prior integrative genomics analyses of other splicing regulators, including those of hnRNP proteins, have also identified a relatively low overlap between binding and function. At least two possible explanations for this exist. First, the highly interconnected biology of splicing regulatory proteins could lead to high levels of indirect effects. Splicing regulatory proteins regulate splicing and thereby protein expression of other splicing factors, which in turn regulate other splicing factors. Depletion of one splicing regulator, such as hnRNP L, induces systems-level alterations in the cell that extend beyond other splicing regulators, for instance to DNA binding proteins and epigenetic factors in the case of hnRNP L. These network effects potentiate the indirection of splicing changes as a readout of the direct consequences of depletion of the splicing regulator under study. Second, the methodology of high throughput sequencing could provide partial discovery of splicing changes and/or RBP binding sites due to sequence bias in library preparation, mapping errors or unequal mappability of the genome, and low expression of mRNAs that results in low confidence of splicing changes. Taken together, these caveats urge strong caution in interpreting the results of integrative genomic analyses.

Despite these caveats, valuable information about hnRNP L splicing control can be gained by combining CLIP-seq and RNA-seq results. First, hnRNP L-repressed exons are strongly demarcated by hnRNP L peaks within the exon or in the upstream or downstream periexonic regions. This observation supports a model in which hnRNP L directly represses exon inclusion through exonic and exonic proximal interactions, potentially by blocking access of the spliceosomal subunits to the 5' and 3' splice sites, or potentially also by stabilizing snRNP/RNA interactions such that snRNP exchange is not possible. This latter mechanism was first described in CD45 exon 4, which has an exonic binding site for hnRNP L that was first identified biochemically and is successfully captured in CLIP-seq. While genomics approaches do not discriminate between these two possibilities, the strong association between exonic and periexonic hnRNP L occupancy and repression of splicing indicates a direct mechanism by which hnRNP L blocks exon inclusion.

In contrast, hnRNP L-enhanced exons are not significantly associated with hnRNP L interactions when compared to a set of exons with inclusion levels that are not altered by hnRNP L knockdown. This finding suggests indirect enhancement of splicing. While hnRNP L-enhanced exons are not associated with CLIP-seq peaks, they are flanked by short, GC-rich introns. These two features co-occur globally within the human transcriptome, and are thought to represent a distinct class of exons which are not strongly demarcated by nucleosome occupancy. In support of this, hnRNP L-enhanced exons are less occupied by nucleosomes than hnRNP L-repressed exons. These results support a model in which hnRNP L-enhanced exons are generally more sensitive to alterations in chromatin than the majority of human exons, which have well-positioned nucleosomes.

This model raises the question of how hnRNP L is linked to chromatin. Several lines of evidence support a functional connection between hnRNP L and chromatin. First, chromatin modifiers are enriched among hnRNP L splicing targets, including enzymes with histone deacetylase and demethylase activity (HDAC10 and KDM6A, respectively), proteins that recruit histone modifiers (HMG20A, e.g.), enzymes that modify DNA (DNMT3B), and histone acetyltransferase complex subunits (EP400, MORF4L2). While the alterations in the activity of these factors induced by hnRNP L depletion is difficult to infer, the enrichment of epigenetic functional categories among hnRNP L splicing targets is a surprising find.

As described in previous chapters, hnRNP L is physically associated with the KMT3a (also known as SETD2) complex in humans and is essential to its H3K36me3 activity *in vivo*. This epigenetic mark is enriched at exon-intron boundaries and is an important component of the “chromatin code” of pre-mRNA splicing. The observation that hnRNP L-enhanced exons have reduced nucleosome occupancy raises the hypothesis that they are particularly sensitive to global reductions in H3K36me3 levels upon hnRNP L depletion.

To test this hypothesis, I propose a comparative chromatin immunoprecipitation experiment to quantify H3K36me3 and total H3 levels in hnRNP L-depleted and mock-depleted JSL1 cells. I hypothesize that hnRNP L-enhanced exons have lower basal levels of H3K36me3 due to reduced nucleosome occupancy, which can be quantified by H3 ChIP-qPCR. Additionally, I propose knockdown of KMT3a and RT-PCR quantification of exon inclusion for exons that were indirectly enhanced by hnRNP L. I hypothesize that KMT3a knockdown will reduce inclusion of these exons.

Because hnRNP L regulates splicing of other chromatin modifiers, including histone deacetylases, it is possible that the effect of hnRNP L on chromatin occurs independently of H3K36me3 or via a combination of multiple mechanisms. Further exploration of chromatin modifications from ENCODE data may reveal deeper insights into the chromatin state underlying hnRNP L-enhanced exons.

In sum, these findings highlight a direct repressive role for hnRNP L in pre-mRNA splicing. In addition, an indirect role for hnRNP L in splicing enhancement highlights a potential interplay between hnRNP L and chromatin. Future experiments are necessary to test this hypothesis.

Comparison of CLIP-seq analyses

In this dissertation, I provide computational analysis of three distinct CLIP-seq studies with the same software pipeline. Minor modifications to the pipeline were made to reflect the experimental differences in each study, for example the incorporation of uniquely barcoded aliquots in the CELF2 CLIP-seq experiment and the metagenome index of Rift Valley Fever Virus-infected human cells in the DDX17 CLIP-seq study. Combined with the hnRNP L CLIP-seq analysis, these three experiments provide valuable comparisons and contrasts. Importantly, by comparing CLIP-seq studies performed and analyzed by the same group using the same tools, a bigger picture of the possible biases of CLIP-seq experiments emerges.

First, CELF2 and hnRNP L CLIP-seq libraries were prepared from the same cells in the same conditions: unstimulated and stimulated JSL1 cells. This identical setting coupled with the fact that both of these proteins are splicing factors makes a comparison particularly interesting. One major difference between the two CLIP-seq experiments was the library preparation. In contrast to hnRNP L CLIP-seq, CELF2 CLIP samples were split into three aliquots before PCR amplification, and unique barcodes were ligated onto each aliquot. This allowed the computational discrimination of PCR duplicates from true multicopy inserts. Despite this difference, the input to the CLIP-seq peak caller was the same format, the only difference in the data being that in the case of CELF2, up to three copies of a read aligned to the same genomic position were possible. The peak caller therefore was identical in both experiments.

One of the first analyses of CLIP-seq peaks is motif enrichment. For both CELF2 and hnRNP L CLIP-seq peaks, an identical permutation algorithm was used to identify enriched hexamers within CLIP-seq peaks as compared to permuted backgrounds. Although standard CLIP-seq library preparation was utilized (in contrast to photoreactive nucleoside crosslinking, or PAR-CLIP), there exists a possibility of crosslinking bias. If this type of sequence bias were inherent to CLIP-seq in general, it would complicate the interpretation of motif enrichment results.

If CLIP-seq were inherently biased towards specific nucleotides, one might expect to see enrichment for that nucleotide within CLIP-seq motifs. However, all four simple nucleotide repeat hexamers (TTTTTT, AAAAAA, CCCCC, and GGGGGG) are depleted within CELF2 CLIP-seq peaks in both unstimulated and stimulated conditions (table C1). These simple nucleotide repeat hexamers are also depleted in unstimulated and stimulated CELF2 hnRNP L peaks, except for CCCCCC which is enriched.

Because this hexamer is not enriched in CELF2 CLIP-seq peaks, it is likely that this motif is enriched in hnRNP L CLIP-seq peaks due to sequence specificity of hnRNP L-RNA interactions and not general sequence bias in UV crosslinking. Notably, this motif is similar to the most enriched hexamers within hnRNP L CLIP-seq peaks, CACACA and ACACAC. While these results demonstrate that homopolymeric hexamers are not enriched generally across these two CLIP-seq experiments from the same cell line, the possibility of systematic enrichment for shorter homopolymeric repeats still exists and would require different motif enrichment strategies than those developed for this dissertation to examine comprehensively.

Hexamer	Unstimulated hnRNP L Z-score	Stimulated hnRNP L Z-score	Unstimulated CELF2 Z-score	Stimulated CELF2 Z-score
TTTTTT	-17.53	-14.32	-8.14	-6.86
AAAAAA	-3.82	-2.24	-6.53	-7.56
CCCCCC	5.39	8.86	-1.83	-1.70
GGGGGG	-3.97	-2.86	-1.93	-1.95

Table C1. Homopolymeric hexamer enrichment analysis within hnRNP L and CELF2 CLIP-seq peaks. Z-scores for each *k*mer of length 6nt (hexamers) were computed by iterative randomization of CLIP-seq peaks within the transcripts that contain them, as described in Materials and Methods and discussed in above chapters and appendices. To investigate the possibility of systematic sequence bias in CLIP, Z-scores for each of the four possible homopolymeric hexamers are displayed.

Another important permutation-based method was used in my analysis of CLIP-seq peaks to estimate the significance of overlap between two sets of peaks. In this analysis, one set of CLIP-seq peaks is held constant while the other is iteratively permuted within the transcripts that contain the peaks, disallowing cross-transcript randomization. If iterative permutation generates overlaps equal to or greater than the extent of overlap between the actual CLIP-seq peaks, there is insufficient evidence to reject the null hypothesis that these two sets of CLIP-seq peaks are not spatially associated with one another. This analysis was first used to compare the extent of overlap between hnRNP L peaks in unstimulated and stimulated cells. I went on to use this same analysis on CELF2 CLIP-seq peaks in unstimulated and stimulated cells to investigate the possibility that cell stimulation alters CELF2-RNA interaction sites. This hypothesis was particularly interesting due to the fact that CELF2 splicing regulatory function is altered by cell stimulation. Finally, I used this analysis on DDX17 CLIP-seq peaks from uninfected and infected U2OS cells to investigate the hypothesis that infection alters DDX17-RNA interactions in the host cell transcriptome.

In all of these comparisons, the overlap between the two sets of CLIP-seq peaks under comparison was greater than the overlap generated by permutation (table C2). This analysis demonstrates that the global extent of overlap between CLIP-seq peaks is greater than that expected by permutation, a process which simulates the null hypothesis of uniform randomness in the positioning of CLIP-seq peaks. However, this analysis does not generate transcript-level information, instead only the extent of global overlap is considered. For this reason, it is still possible that protein-RNA interactions within certain transcripts are different between the two sets of peaks under comparison. Additionally, the number of reads aligning to a CLIP-seq peak is not considered, so quantitative variations in the signal intensity of a CLIP-seq peak are discarded. It is therefore also possible that global alterations in the heights of CLIP-seq peaks exist in a consistent manner that is lost in this analysis because only the footprint of a peak is utilized. Future developments in comparison of CLIP-seq datasets might provide valuable insight into the differences in protein-RNA interactions induced by cell signaling or infection with pathogens.

CLIP-seq peaks compared:	Fraction of overlapping peaks:	Fraction of permuted overlap:	Binomial p-value:
Unstimulated and stimulated hnRNP L	0.347	0.0128	< 2.2e-16
Unstimulated and stimulated CELF2	0.621	0.0448	< 2.2e-16
Uninfected and infected DDX17	0.0826	0.00342	< 2.2e-16

Table C2. Randomized overlap analysis between CLIP-seq binding profiles. A permutation analysis of the significance of overlap between CLIP-seq binding profiles for hnRNP L in unstimulated and stimulated JSL1 cells, for CELF2 in the same two conditions, and for DDX17 in uninfected and RVFV-infected U2OS cells is provided. P-values were obtained using a two-sided exact binomial test.

In all comparisons, the significance of the degree of overlap is statistically significant by a binomial test. These results indicate that cell stimulation does not globally alter either CELF2 or hnRNP L CLIP-seq sites. In addition, infection with Rift Valley Fever Virus does not appear to globally alter DDX17-RNA interaction sites. However, it is still possible that a subset of RNAs is subject to condition-specific or infection-specific interactions. Such subtle changes in CLIP-seq peaks will require more fine-grained analyses to detect.

In addition to the above information, CLIP-seq provides data on the transcript features (e.g. 3'UTR, coding exon, 5'UTR, etc.) enriched within CLIP-seq peaks. These data provide valuable insights into possible transcriptome-wide roles of an RBP in previously unappreciated processes. I compared the transcript features of hnRNP L, CELF2, and DDX17 CLIP-seq experiments (table C3). A further hnRNP L CLIP-seq experiment in CD4+ cells is omitted from this table, but the findings were similar to hnRNP L CLIP-seq in JSL1 cells (see above chapters).

Transcript feature	Unstim. hnRNP L	Stim. hnRNP L	Unstim. CELF2	Stim. CELF2	Uninf. DDX17	Inf. DDX17	Total refSeq
Distal intron	86.98	88.21	69.66	72.73	58.54	71.48	87.32
Proximal intron	6.14	5.83	9.29	8.64	11.0	14.66	7.64
3'UTR exon	5.76	4.95	17.0	15.35	16.97	6.81	2.17
Coding exon	0.98	0.87	3.39	2.74	9.9	4.58	2.46
5'UTR exon	0.13	0.14	0.67	0.53	3.59	2.48	0.41

Table C3. Transcript features within CLIP-seq peaks from hnRNP L, CELF2, and DDX17. CLIP-seq experiments from hnRNP L and CELF2 (JSL1 cells) and from DDX17 (U2OS) cells were analyzed by the same bioinformatics pipeline. The fraction of nucleotides overlapping each of 5 types of transcript feature were calculated. Proximal intron is defined as intronic sequence within 250nt of an annotated exon.

Several interesting findings arise when these experiments are examined together. First, CELF2 CLIP-seq peaks are 7-fold enriched for 3'UTR interactions when compared to the size of the 3'UTR footprint within the refSeq transcriptome. As mentioned in the above appendix, 3'UTR interactions are a property of mRNA regulation by CELF1, and this finding suggests that CELF2 may also regulate mRNA through 3'UTR interactions. Interestingly, all three proteins are enriched for 3'UTR interactions, suggesting that at least some of CELF2's 3'UTR bias might be explained by a systematic bias toward 3'UTR localization in all CLIP-seq experiments. Second, stimulation of JSL1 cells induces a slight yet noticeable shift toward distal intron interactions for hnRNP L and CELF2, with commensurate loss in other types of transcript features. Because this shift exists for both proteins, it is possible that the gene expression changes induced by cell stimulation result in a transcriptome that has a larger distal intronic footprint. Third, infection of U2OS cells with Rift Valley Fever Virus results in a similar redistribution of DDX17 CLIP-seq peaks to distal intronic regions. While striking, this result is difficult to interpret as DDX17-mRNA interactions are poorly understood because much of the focus on DDX17-RNA interactions has been centered around the role this RNA helicase plays in the biogenesis of miRNAs.

In conclusion, I have presented computational analysis of three independent CLIP-seq experiments in two cell types. By comparing motif enrichment analyses from all three CLIP-seq experiments, I find little evidence for a homopolymeric sequence bias in peaks defined these studies. This result lends credence to motif enrichment results as reflective of the sequence specificity of the protein under study. For this reason, CLIP-seq is a useful tool not only to study protein-RNA localization in the transcriptome, but also to study sequence specificity of RNA binding proteins. Several high-throughput methods to study sequence specificity of RNA binding proteins have been developed, including RNAcompete and RNA Bind-n-seq. One caveat of *in vitro* studies of sequence specificity for RNA binding proteins is that the protein has been removed from its cellular context. Numerous biological factors could influence RBP-RNA interactions *in vivo*, including posttranslational modification status, co-associated proteins, substrate abundance, subcellular localization, and others. I propose that CLIP-seq provides valuable insight into the sequence specificity of an RNA binding protein in its natural cellular context.

Additionally, I describe a permutation algorithm for the analysis of the significance of overlap between two sets of CLIP-seq peaks. While this analysis is an important component of the study of an RNA binding protein in multiple cell states, as presented here, several important caveats exist. As CLIP-seq analyses are increasingly utilized to study RNA binding protein localizations across multiple cell types at various developmental stages, software development efforts will be needed to identify biologically relevant dynamics in the midst of a larger context of overlapping binding sites. Computational challenges such as these motivate the development of new tools that are generalizable and useful in a variety of contexts.

Finally, I demonstrate a comparative analysis of the types of transcript features engaged by hnRNP L, CELF2, and DDX17 in different cell states. This analysis demonstrates the value of comparisons between CLIP-seq experiments and highlights a potential role for CLIP-seq in studying the dynamics of protein-RNA interactions in response to viral infection or cell signaling.

REFERENCES

1. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature*. 2010;463(7280):457-463.

2. Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456(7221):470-476.

3. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*. 2008;40(12):1413-1415.

4. Wahl MC, Will CL, Luhrmann R. The spliceosome: Design principles of a dynamic RNP machine. *Cell*. 2009;136(4):701-718.

5. Chen M, Manley JL. Mechanisms of alternative splicing regulation: Insights from molecular and genomics approaches. *Nature Reviews Molecular Cell Biology*. 2009;10(11):741-754.

6. Guth S, Valcarcel J. Kinetic role for mammalian SF1/BBP in spliceosome assembly and function after polypyrimidine tract recognition by U2AF. *J Biol Chem*. 2000;275(48):38059-38066.

7. Cheng Z, Menees TM. RNA splicing and debranching viewed through analysis of RNA lariats. *Molecular Genetics & Genomics: MGG*. 2011;286(5-6):395-410.

8. Hegele A, Kamburov A, Grossmann A, et al. Dynamic protein-protein interaction wiring of the human spliceosome. *Mol Cell*. 2012;45(4):567-580.

9. Matlin AJ, Moore MJ. Spliceosome assembly and composition. *Advances in Experimental Medicine & Biology*. 2007;623:14-35.

10. Inoue K, Hoshijima K, Sakamoto H, Shimura Y. Binding of the drosophila sex-lethal gene product to the alternative splice site of transformer primary transcript. *Nature*. 1990;344(6265):461-463.

11. Hoshijima K, Inoue K, Higuchi I, Sakamoto H, Shimura Y. Control of doublesex alternative splicing by transformer and transformer-2 in drosophila. *Science*. 1991;252(5007):833-836.

12. Valcarcel J, Singh R, Zamore PD, Green MR. The protein sex-lethal antagonizes the splicing factor U2AF to regulate alternative splicing of transformer pre-mRNA. *Nature*. 1993;362(6416):171-175.

13. Forch P, Merendino L, Martinez C, Valcarcel J. Modulation of msl-2 5' splice site recognition by sex-lethal. *Rna-A Publication of the Rna Society*. 2001;7(9):1185-1191.

14. Zuo P, Maniatis T. The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. *Genes Dev*. 1996;10(11):1356-1368.

15. Boucher L, Ouzounis CA, Enright AJ, Blencowe BJ. A genome-wide survey of RS domain proteins. *Rna-A Publication of the Rna Society*. 2001;7(12):1693-1701.

16. Fu XD. The superfamily of arginine/serine-rich splicing factors. *Rna-A Publication of the Rna Society*. 1995;1(7):663-680.

17. Ge H, Manley JL. A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA in vitro. *Cell*. 1990;62(1):25-34.

18. Krainer AR, Conway GC, Kozak D. The essential pre-mRNA splicing factor SF2 influences 5' splice site selection by activating proximal sites. *Cell*. 1990;62(1):35-42.

19. Zahler AM, Neugebauer KM, Lane WS, Roth MB. Distinct functions of SR proteins in alternative pre-mRNA splicing. *Science*. 1993;260(5105):219-222.

20. Zahler AM, Roth MB. Distinct functions of SR proteins in recruitment of U1 small nuclear ribonucleoprotein to alternative 5' splice sites. *Proc Natl Acad Sci U S A*. 1995;92(7):2642-2646.

21. Choi YD, Dreyfuss G. Isolation of the heterogeneous nuclear RNA-ribonucleoprotein complex (hnRNP): A unique supramolecular assembly. *Proc Natl Acad Sci U S A*. 1984;81(23):7471-7475.

22. Pinol-Roma S, Choi YD, Matunis MJ, Dreyfuss G. Immunopurification of heterogeneous nuclear ribonucleoprotein particles reveals an assortment of RNA-binding proteins. *Genes Dev*. 1988;2(2):215-227.

23. Mayeda A, Helfman DM, Krainer AR. Modulation of exon skipping and inclusion by heterogeneous nuclear ribonucleoprotein A1 and pre-mRNA splicing factor SF2/ASF. *Molecular & Cellular Biology*. 1993;13(5):2993-3001.

24. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nature Reviews Genetics*. 2002;3(4):285-298.

25. Kashima T, Rao N, David CJ, Manley JL. hnRNP A1 functions with specificity in repression of SMN2 exon 7 splicing. *Hum Mol Genet*. 2007;16(24):3149-3159.

26. David CJ, Chen M, Assanah M, Canoll P, Manley JL. HnRNP proteins controlled by c-myc deregulate pyruvate kinase mRNA splicing in cancer. *Nature*. 2010;463(7279):364-368.

27. Zhu J, Mayeda A, Krainer AR. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol Cell*. 2001;8(6):1351-1361.

28. Martinez-Contreras R, Fisette JF, Nasim FU, Madden R, Cordeau M, Chabot B. Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *Plos Biology*. 2006;4(2):e21.

29. Venables JP, Koh CS, Froehlich U, et al. Multiple and specific mRNA processing targets for the major human hnRNP proteins. *Molecular & Cellular Biology*. 2008;28(19):6033-6043.

30. Oh Hk, Lee E, Jang HN, et al. hnRNP A1 contacts exon 5 to promote exon 6 inclusion of apoptotic fas gene. *Apoptosis*. 2013;18(7):825-835.

31. Das S, Krainer AR. Emerging functions of SRSF1, splicing factor and oncoprotein, in RNA metabolism and cancer. *Mol Cancer Res*. 2014;12(9):1195-1204.

32. Cho S, Hoang A, Sinha R, et al. Interaction between the RNA binding domains of ser-arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. *Proc Natl Acad Sci U S A*. 2011;108(20):8233-8238.

33. Pandit S, Zhou Y, Shiue L, et al. Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol Cell*. 2013;50(2):223-235.

34. Lynch KW. Consequences of regulated pre-mRNA splicing in the immune system. *Nature Reviews.Immunology*. 2004;4(12):931-940.

35. Majeti R, Bilwes AM, Noel JP, Hunter T, Weiss A. Dimerization-induced inhibition of receptor protein tyrosine phosphatase function through an inhibitory wedge. *Science*. 1998;279(5347):88-91.

36. Xu Z, Weiss A. Negative regulation of CD45 by differential homodimerization of the alternatively spliced isoforms. *Nat Immunol*. 2002;3(8):764-771.

37. Lynch KW, Weiss A. A model system for activation-induced alternative splicing of CD45 pre-mRNA in T cells implicates protein kinase C and ras. *Molecular & Cellular Biology*. 2000;20(1):70-80.

38. Lynch KW, Weiss A. A CD45 polymorphism associated with multiple sclerosis disrupts an exonic splicing silencer. *J Biol Chem*. 2001;276(26):24341-24347.

39. Melton AA, Jackson J, Wang J, Lynch KW. Combinatorial control of signal-induced exon repression by hnRNP L and PSF. *Molecular & Cellular Biology*. 2007;27(19):6972-6984.

40. Topp JD, Jackson J, Melton AA, Lynch KW. A cell-based screen for splicing regulators identifies hnRNP LL as a distinct signal-induced repressor of CD45 variable exon 4. *Rna-A Publication of the Rna Society*. 2008;14(10):2038-2049.

41. Gaudreau MC, Heyd F, Bastien R, Wilhelm B, Moroy T. Alternative splicing controlled by heterogeneous nuclear ribonucleoprotein L regulates development, proliferation, and migration of thymic pre-T cells. *Journal of Immunology*. 2012;188(11):5377-5388.

42. Shankarling G, Lynch KW. Minimal functional domains of paralogues hnRNP L and hnRNP LL exhibit mechanistic differences in exonic splicing repression. *Biochem J*. 2013;453(2):271-279.

43. Oberdoerffer S, Moita LF, Neems D, Freitas RP, Hacohen N, Rao A. Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL. *Science*. 2008;321(5889):686-691.

44. Motta-Mena LB, Heyd F, Lynch KW. Context-dependent regulatory mechanism of the splicing factor hnRNP L. *Mol Cell*. 2010;37(2):223-234.

45. Tong A, Nguyen J, Lynch KW. Differential expression of CD45 isoforms is controlled by the combined activity of basal and inducible splicing-regulatory elements in each of the variable exons. *J Biol Chem*. 2005;280(46):38297-38304.

46. Chiou NT, Shankarling G, Lynch KW. hnRNP L and hnRNP A1 induce extended U1 snRNA interactions with an exon to repress spliceosome assembly. *Mol Cell*. 2013;49(5):972-982.

47. Lareau LF, Brenner SE. Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. *Mol Biol Evol*. 2015.

48. Rossbach O, Hung LH, Schreiner S, et al. Auto- and cross-regulation of the hnRNP L proteins by alternative splicing. *Molecular & Cellular Biology*. 2009;29(6):1442-1451.

49. Rosel M, Khaldoyanidi S, Zawadzki V, Zoller M. Involvement of CD44 variant isoform v10 in progenitor cell adhesion and maturation. *Exp Hematol*. 1999;27(4):698-711.

50. Loh TJ, Cho S, Moon H, et al. hnRNP L inhibits CD44 V exon splicing through interacting with its upstream intron. *Biochim Biophys Acta*. 2015.

51. Zoller M, Gupta P, Marhaba R, Vitacolonna M, Freyschmidt-Paul P. Anti-CD44-mediated blockade of leukocyte migration in skin-associated immune diseases. *J Leukoc Biol*. 2007;82(1):57-71.

52. Manten-Horst E, Danen EH, Smit L, et al. Expression of CD44 splice variants in human cutaneous melanoma and melanoma cell lines is related to tumor progression and metastatic potential. *International Journal of Cancer*. 1995;64(3):182-188.

53. Shih SC, Claffey KP. Regulation of human vascular endothelial growth factor mRNA stability in hypoxia by heterogeneous nuclear ribonucleoprotein L. *J Biol Chem*. 1999;274(3):1359-1365.

54. Jafarifar F, Yao P, Eswarappa SM, Fox PL. Repression of VEGFA by CA-rich element-binding microRNAs is modulated by hnRNP L. *EMBO J*. 2011;30(7):1324-1334.

55. Ule J, Jensen K, Mele A, Darnell RB. CLIP: A method for identifying protein-RNA interaction sites in living cells. *Methods (Duluth)*. 2005;37(4):376-386.

56. Darnell R. CLIP (cross-linking and immunoprecipitation) identification of RNAs bound by a specific protein. *Cold Spring Harbor protocols*. 2012;2012(11):1146-1160.

57. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP identifies nova-regulated RNA networks in the brain. *Science*. 2003;302(5648):1212-1215.

58. Licatalosi DD, Mele A, Fak JJ, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*. 2008;456(7221):464-469.

59. Yeo GW, Coufal NG, Liang TY, Peng GE, Fu XD, Gage FH. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature Structural & Molecular Biology*. 2009;16(2):130-137.

60. Sanford JR, Wang X, Mort M, et al. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res*. 2009;19(3):381-394.

61. Licatalosi DD, Yano M, Fak JJ, et al. Ptbp2 represses adult-specific splicing to regulate the generation of neuronal precursors in the embryonic brain. *Genes Dev*. 2012;26(14):1626-1642.

62. Huelga SC, Vu AQ, Arnold JD, et al. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Reports*. 2012;1(2):167-178.

63. Charizanis K, Lee KY, Batra R, et al. Muscleblind-like 2-mediated alternative splicing in the developing brain and dysregulation in myotonic dystrophy. *Neuron*. 2012;75(3):437-450.

64. Pandit S, Zhou Y, Shiue L, et al. Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol Cell*. 2013;50(2):223-235.

65. Weyn-Vanhentenryck SM, Mele A, Yan Q, et al. HITS-CLIP and integrative modeling define the rbfox splicing-regulatory network linked to brain development and autism. *Cell Reports*. 2014;6(6):1139-1152.

66. Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*. 2009;460(7254):479-486.

67. Zisoulis DG, Lovci MT, Wilbert ML, et al. Comprehensive discovery of endogenous argonaute binding sites in caenorhabditis elegans. *Nature Structural & Molecular Biology*. 2010;17(2):173-179.

68. Moore MJ, Zhang C, Gantman EC, Mele A, Darnell JC, Darnell RB. Mapping argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nature Protocols*. 2014;9(2):263-293.

69. Bahn JH, Ahn J, Lin X, et al. Genomic analysis of ADAR1 binding and its involvement in multiple RNA processing pathways. *Nat Commun*. 2015;6:6355.

70. Selvanathan SP, Graham GT, Erkizan HV, et al. Oncogenic fusion protein EWS-FLI1 is a network hub that regulates alternative splicing. *Proc Natl Acad Sci U S A*. 2015.

71. Hurt JA, Robertson AD, Burge CB. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res*. 2013;23(10):1636-1650.

72. Yang YC, Di C, Hu B, et al. CLIPdb: A CLIP-seq database for protein-RNA interactions. *BMC Genomics*. 2015;16(1):51.

73. Shoemaker DD, Schadt EE, Armour CD, et al. Experimental annotation of the human genome using microarray technology. *Nature*. 2001;409(6822):922-927.

74. Johnson JM, Castle J, Garrett-Engele P, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*. 2003;302(5653):2141-2144.

75. Clark TA, Sugnet CW, Ares M Jr. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*. 2002;296(5569):907-910.

76. Pan Q, Shai O, Misquitta C, et al. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell*. 2004;16(6):929-941.

77. Blanchette M, Green RE, Brenner SE, Rio DC. Global analysis of positive and negative pre-mRNA splicing regulators in drosophila. *Genes Dev*. 2005;19(11):1306-1314.

78. Sugnet CW, Srinivasan K, Clark TA, et al. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Computational Biology*. 2006;2(1):e4.

79. Blencowe BJ. Alternative splicing: New insights from global analyses. *Cell*. 2006;126(1):37-47.

80. Bentley DL. Rules of engagement: Co-transcriptional recruitment of pre-mRNA processing factors. *Curr Opin Cell Biol*. 2005;17(3):251-256.

81. Dye MJ, Gromak N, Proudfoot NJ. Exon tethering in transcription by RNA polymerase II. *Mol Cell*. 2006;21(6):849-859.

82. Wilhelm BT, Marguerat S, Watt S, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*. 2008;453(7199):1239-1243.

83. Sultan M, Schulz MH, Richard H, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*. 2008;321(5891):956-960.

84. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-seq. *Bioinformatics*. 2009;25(9):1105-1111.

85. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*. 2010;7(12):1009-1015.

86. Shen S, Park JW, Huang J, et al. MATS: A bayesian framework for flexible detection of differential alternative splicing from RNA-seq data. *Nucleic Acids Res*. 2012;40(8):e61.

87. Bomsztyk K, Denisenko O, Ostrowski J. hnRNP K: One protein multiple processes. *Bioessays*. 2004;26(6):629-638.

88. Makeyev AV, Liebhaber SA. The poly(C)-binding proteins: A multiplicity of functions and a search for mechanisms. *Rna-A Publication of the Rna Society*. 2002;8(3):265-278.

89. Martinez-Contreras R, Cloutier P, Shkreta L, Fisette JF, Revil T, Chabot B. hnRNP proteins and splicing control. *Advances in Experimental Medicine & Biology*. 2007;623:123-147.

90. Hung LH, Heiner M, Hui J, Schreiner S, Benes V, Bindereif A. Diverse roles of hnRNP L in mammalian mRNA processing: A combined microarray and RNAi analysis. *Rna-A Publication of the Rna Society*. 2008;14(2):284-296.

91. Erkelenz S, Mueller WF, Evans MS, et al. Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *Rna-A Publication of the Rna Society*. 2013;19(1):96-102.

92. Martinez NM, Lynch KW. Control of alternative splicing in immune responses: Many regulators, many predictions, much still to learn. *Immunol Rev*. 2013;253(1):216-236.

93. Ansel KM. RNA regulation of the immune system. *Immunol Rev*. 2013;253(1):5-11.

94. Rothrock CR, House AE, Lynch KW. HnRNP L represses exon splicing via a regulated exonic splicing silencer. *EMBO J*. 2005;24(15):2792-2802.

95. Topp JD, Jackson J, Melton AA, Lynch KW. A cell-based screen for splicing regulators identifies hnRNP LL as a distinct signal-induced repressor of CD45 variable exon 4. *Rna-A Publication of the Rna Society*. 2008;14(10):2038-2049.

96. Tong A, Nguyen J, Lynch KW. Differential expression of CD45 isoforms is controlled by the combined activity of basal and inducible splicing-regulatory elements in each of the variable exons. *J Biol Chem*. 2005;280(46):38297-38304.

97. Gaudreau MC, Heyd F, Bastien R, Wilhelm B, Moroy T. Alternative splicing controlled by heterogeneous nuclear ribonucleoprotein L regulates development, proliferation, and migration of thymic pre-T cells. *Journal of Immunology*. 2012;188(11):5377-5388.

98. Hui J, Hung LH, Heiner M, et al. Intronic CA-repeat and CA-rich elements: A new class of regulators of mammalian alternative splicing. *EMBO J*. 2005;24(11):1988-1998.

99. Hermiston ML, Xu Z, Majeti R, Weiss A. Reciprocal regulation of lymphocyte activation by tyrosine kinases and phosphatases. *J Clin Invest*. 2002;109(1):9-14.

100. Preussner M, Schreiner S, Hung LH, et al. HnRNP L and L-like cooperate in multiple-exon regulation of CD45 alternative splicing. *Nucleic Acids Res*. 2012;40(12):5666-5678.

101. Rothrock C, Cannon B, Hahm B, Lynch KW. A conserved signal-responsive sequence mediates activation-induced alternative splicing of CD45. *Mol Cell*. 2003;12(5):1317-1324.

102. Xu Z, Weiss A. Negative regulation of CD45 by differential homodimerization of the alternatively spliced isoforms. *Nat Immunol*. 2002;3(8):764-771.

103. Majeti R, Bilwes AM, Noel JP, Hunter T, Weiss A. Dimerization-induced inhibition of receptor protein tyrosine phosphatase function through an inhibitory wedge. *Science*. 1998;279(5347):88-91.

104. Ule J. High-throughput sequencing methods to study neuronal RNA-protein interactions. *Biochem Soc Trans*. 2009;37(Pt 6):1278-1280.

105. Oberdoerffer S, Moita LF, Neems D, Freitas RP, Hacohen N, Rao A. Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL. *Science*. 2008;321(5889):686-691.

106. Xue Y, Zhou Y, Wu T, et al. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol Cell*. 2009;36(6):996-1006.

107. Martinez NM, Pan Q, Cole BS, et al. Alternative splicing networks regulated by signaling in human T cells. *Rna-A Publication of the Rna Society*. 2012;18(5):1029-1040.

108. Rossbach O, Hung LH, Schreiner S, et al. Auto- and cross-regulation of the hnRNP L proteins by alternative splicing. *Molecular & Cellular Biology*. 2009;29(6):1442-1451.

109. Maillard I, Fang T, Pear WS. Regulation of lymphoid development, differentiation, and function by the notch pathway. *Annu Rev Immunol*. 2005;23:945-974.

110. Weiss A. Molecular and genetic insights into the role of protein tyrosine kinases in T cell receptor signaling. *Clinical Immunology & Immunopathology*. 1995;76(3 Pt 2):S158-62.

111. Liu P, Keller JR, Ortiz M, et al. Bcl11a is essential for normal lymphoid development. *Nat Immunol*. 2003;4(6):525-532.

112. Go WY, Liu X, Roti MA, Liu F, Ho SN. NFAT5/TonEBP mutant mice define osmotic stress as a critical feature of the lymphoid microenvironment. *Proc Natl Acad Sci U S A*. 2004;101(29):10673-10678.

113. Ikawa T, Kawamoto H, Goldrath AW, Murre C. E proteins and notch signaling cooperate to promote T cell lineage specification and commitment. *J Exp Med*. 2006;203(5):1329-1342.

114. Wan YY, Chi H, Xie M, Schneider MD, Flavell RA. The kinase TAK1 integrates antigen and cytokine receptor signaling for T cell development, survival and function. *Nat Immunol*. 2006;7(8):851-858.

115. Yamamoto-Furusho JK, Barnich N, Xavier R, Hisamatsu T, Podolsky DK. Centaurin beta1 down-regulates nucleotide-binding oligomerization domains 1- and 2-dependent NF-kappaB activation. *J Biol Chem*. 2006;281(47):36060-36070.

116. Ou CY, Kim JH, Yang CK, Stallcup MR. Requirement of cell cycle and apoptosis regulator 1 for target gene activation by wnt and beta-catenin and for anchorage-independent growth of human colon carcinoma cells. *J Biol Chem*. 2009;284(31):20629-20637.

117. Huelga SC, Vu AQ, Arnold JD, et al. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Reports*. 2012;1(2):167-178.

118. Lagier-Tourenne C, Polymenidou M, Hutt KR, et al. Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nat Neurosci*. 2012;15(11):1488-1497.

119. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology*. 2004;11(2-3):377-394.

120. Dikic I, Dikic I, Schlessinger J. Identification of a new Pyk2 isoform implicated in chemokine and antigen receptor signaling. *J Biol Chem*. 1998;273(23):14301-14308.

121. Davidson D, Viallet J, Veillette A. Unique catalytic properties dictate the enhanced function of p59fynT, the hemopoietic cell-specific isoform of the fyn tyrosine protein kinase, in T cells. *Molecular & Cellular Biology*. 1994;14(7):4554-4564.

122. Van Parijs L, Refaeli Y, Lord JD, Nelson BH, Abbas AK, Baltimore D. Uncoupling IL-2 signals that regulate T cell proliferation, survival, and fas-mediated activation-induced cell death. *Immunity*. 1999;11(3):281-288.

123. Yu Y, Wang J, Khaled W, et al. Bcl11a is essential for lymphoid development and negatively regulates p53. *J Exp Med*. 2012;209(13):2467-2483.

124. Appleby MW, Gross JA, Cooke MP, Levin SD, Qian X, Perlmutter RM. Defective T cell receptor signaling in mice lacking the thymic isoform of p59fyn. *Cell*. 1992;70(5):751-763.

125. Beinke S, Phee H, Clingan JM, Schlessinger J, Matloubian M, Weiss A. Proline-rich tyrosine kinase-2 is critical for CD8 T-cell short-lived effector fate. *Proc Natl Acad Sci U S A*. 2010;107(37):16234-16239.

126. Miyazaki T, Takaoka A, Nogueira L, et al. Pyk2 is a downstream mediator of the IL-2 receptor-coupled jak signaling pathway. *Genes Dev*. 1998;12(6):770-775.

127. Slattery C, Ryan MP, McMorrow T. E2A proteins: Regulators of cell phenotype in normal physiology and disease. *Int J Biochem Cell Biol*. 2008;40(8):1431-1436.

128. Schlissel M, Voronova A, Baltimore D. Helix-loop-helix transcription factor E47 activates germ-line immunoglobulin heavy-chain gene transcription and rearrangement in a pre-T-cell line. *Genes Dev*. 1991;5(8):1367-1376.

129. Lynch KW, Weiss A. A CD45 polymorphism associated with multiple sclerosis disrupts an exonic splicing silencer. *J Biol Chem*. 2001;276(26):24341-24347.

130. Jacobsen M, Schweer D, Ziegler A, et al. A point mutation in PTPRC is associated with the development of multiple sclerosis. *Nat Genet*. 2000;26(4):495-499.

131. Liu G, Razanau A, Hai Y, et al. A conserved serine of heterogeneous nuclear ribonucleoprotein L (hnRNP L) mediates depolarization-regulated alternative splicing of potassium channels. *J Biol Chem*. 2012;287(27):22709-22716.

132. Goehe RW, Shultz JC, Murudkar C, et al. hnRNP L regulates the tumorigenic capacity of lung cancer xenografts in mice via caspase-9 pre-mRNA processing. *J Clin Invest*. 2010;120(11):3923-3939.

133. Shankarling G, Cole BS, Mallory MJ, Lynch KW. Transcriptome-wide RNA interaction profiling reveals physical and functional targets of hnRNP L in human T cells. *Mol Cell Biol*. 2014;34(1):71-83.

134. Rossbach O, Hung LH, Schreiner S, et al. Auto- and cross-regulation of the hnRNP L proteins by alternative splicing. *Molecular & Cellular Biology*. 2009;29(6):1442-1451.

135. Xue Y, Zhou Y, Wu T, et al. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol Cell*. 2009;36(6):996-1006.

136. Li H, Qiu J, Fu XD. RASL-seq for massively parallel and quantitative analysis of gene expression. *Curr Protoc Mol Biol*. 2012;Chapter 4:Unit 4.13.1-9.

137. Huang Y, Li W, Yao X, et al. Mediator complex regulates alternative mRNA processing via the MED23 subunit. *Mol Cell*. 2012;45(4):459-469.

138. Kim JE, White FM. Quantitative analysis of phosphotyrosine signaling networks triggered by CD3 and CD28 costimulation in jurkat cells. *J Immunol*. 2006;176(5):2833-2843.

139. Amit M, Donyo M, Hollander D, et al. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep*. 2012;1(5):543-556.

140. Gelfman S, Cohen N, Yearim A, Ast G. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. *Genome Res*. 2013;23(5):789-799.

141. Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB. RNA bind-n-seq: Quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell*. 2014;54(5):887-900.

142. Ray D, Kazan H, Chan ET, et al. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol*. 2009;27(7):667-670.

143. Carrozza MJ, Li B, Florens L, et al. Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell*. 2005;123(4):581-592.

144. Li B, Gogol M, Carey M, Lee D, Seidel C, Workman JL. Combined action of PHD and chromo domains directs the Rpd3S HDAC to transcribed chromatin. *Science*. 2007;316(5827):1050-1054.

145. Yuan W, Xie J, Long C, et al. Heterogeneous nuclear ribonucleoprotein L is a subunit of human KMT3a/Set2 complex required for H3 lys-36 trimethylation activity in vivo. *J Biol Chem*. 2009;284(23):15701-15707.

146. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet*. 2009;41(3):376-381.

147. Shankarling G, Cole BS, Mallory MJ, Lynch KW. Transcriptome-wide RNA interaction profiling reveals physical and functional targets of hnRNP L in human T cells. *Molecular & Cellular Biology*. 2014;34(1):71-83.

148. Hong S, Noh H, Chen H, et al. Signaling by p38 MAPK stimulates nuclear localization of the microprocessor component p68 for processing of selected primary microRNAs. *Science Signaling [Electronic Resource]*. 2013;6(266):ra16.

149. Samaan S, Tranchevent LC, Dardenne E, et al. The Ddx5 and Ddx17 RNA helicases are cornerstones in the complex regulatory array of steroid hormone-signaling pathways. *Nucleic Acids Res*. 2014;42(4):2197-2207.

150. Davis BN, Hilyard AC, Lagna G, Hata A. SMAD proteins control DROSHA-mediated microRNA maturation. *Nature*. 2008;454(7200):56-61.

151. Gregory RI, Yan KP, Amuthan G, et al. The microprocessor complex mediates the genesis of microRNAs. *Nature*. 2004;432(7014):235-240.

152. Mori M, Triboulet R, Mohseni M, et al. Hippo signaling regulates microprocessor and links cell-density-dependent miRNA biogenesis to cancer. *Cell*. 2014;156(5):893-906.

153. Suzuki HI, Yamagata K, Sugimoto K, Iwamoto T, Kato S, Miyazono K. Modulation of microRNA processing by p53. *Nature*. 2009;460(7254):529-533.

154. Sabin LR, Zhou R, Gruber JJ, et al. Ars2 regulates both miRNA- and siRNA-dependent silencing and suppresses RNA virus infection in drosophila. *Cell*. 2009;138(2):340-351.

155. Emery VC, Bishop DH. Characterization of punta toro S mRNA species and identification of an inverted complementary sequence in the intergenic region of punta toro phlebovirus ambisense S RNA that is involved in mRNA transcription termination. *Virology*. 1987;156(1):1-11.

156. Sabin LR, Zheng Q, Thekkat P, et al. Dicer-2 processes diverse viral RNA species. *PLoS ONE [Electronic Resource]*. 2013;8(2):e55458.

157. Bortz E, Westera L, Maamary J, et al. Host- and strain-specific regulation of influenza virus polymerase activity by interacting cellular proteins. *mBio*. 2011;2(4).

158. Mallory MJ, Jackson J, Weber B, Chi A, Heyd F, Lynch KW. Signal- and development-dependent alternative splicing of LEF1 in T cells is controlled by CELF2. *Molecular & Cellular Biology*. 2011;31(11):2184-2195.

159. Arce L, Yokoyama NN, Waterman ML. Diversity of LEF/TCF action in development and disease. *Oncogene*. 2006;25(57):7492-7504.

160. Carlsson P, Waterman ML, Jones KA. The hLEF/TCF-1 alpha HMG protein contains a context-dependent transcriptional activation domain that induces the TCR alpha enhancer in T cells. *Genes Dev*. 1993;7(12A):2418-2430.

161. Vlasova-St Louis I, Bohjanen PR. Coordinate regulation of mRNA decay networks by GU-rich elements and CELF1. *Curr Opin Genet Dev*. 2011;21(4):444-451.

162. Wang ET, Ward AJ, Cherone J, et al. Antagonistic regulation of mRNA expression and splicing by CELF and MBNL proteins. *Genome Res*. 2015.

163. Marquis J, Paillard L, Audic Y, et al. CUG-BP1/CELF1 requires UGU-rich sequences for high-affinity binding. *Biochem J*. 2006;400(2):291-301.