NORTHWESTERN UNIVERSITY

Investigations into the Role of RNA Binding Proteins and Alternative Splicing in the Epithelial-Mesenchymal Transition

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

For the degree

DOCTOR OF PHILOSOPHY

Field of Driskill Graduate Training Program in Life Sciences

By

Samuel Emerson Harvey

EVANSTON, ILLINOIS

June 2019

ABSTRACT

Investigations into the Role of RNA Binding Proteins and Alternative Splicing in the Epithelial-Mesenchymal Transition

Samuel Emerson Harvey

The critical importance of alternative mRNA splicing and the RNA binding proteins that orchestrate this essential layer of post-transcriptional gene regulation is increasingly recognized in gene regulatory programs. We and others have shown that alternative splicing plays a causal role during the Epithelial-Mesenchymal transition, a cell-developmental program that is hijacked by many cancers to acquire the ability to survive, invade, and metastasize to distant sites. While alternative splicing events may be the ultimate effectors of splicing in regulating cell function, RNA binding proteins play a critical role in regulating EMT and regulating cell-state plasticity in normal and oncogenic environments. This thesis describes four situations where RNA binding proteins and the global splicing regulons they control globally influence changes in cell-state, with a particular focus on the shift from the epithelial-cell state to the mesenchymal-cell state that occurs during EMT.

Chapter 1 describes coregulation of alternative splicing by hnRNPM and ESRP1, two RNA binding proteins that play opposite roles in regulating EMT, with hnRNPM necessary for EMT to occur while ESRP1 antagonizes EMT by promoting epithelial-associated alternative splicing. In this chapter, we describe a nuanced overlap of regulation of splicing by hnRNPM and ESRP1 where a small set of skipped exon splicing events controlled discordantly by these two factors is highly associated with EMT-

regulated biological processes and predicts breast cancer patient survival. These events are highly enriched in hnRNPM and ESRP1 shared binding sites in intronic sequence downstream of the 5' splice site, with one splicing event at APLP2 exon 7 showing competitive binding between hnRNPM and ESRP1 to drive APLP2 exon 7 skipping or inclusion in concordance with mesenchymal or epithelial splicing, respectively.

Chapter 2 examines cell-state specific regulation of alternative splicing by hnRNPM specifically by integrating epithelial and mesenchymal hnRNPM-depletion datasets from an *in vitro* model of EMT coupled with crosslinking-associated hnRNPM binding sites across the transcriptome in both cell states. hnRNPM regulates a partially overlapping set of genes and splicing events in these two states, with a shift towards EMT-promotive regulation at both the gene and splicing level in the mesenchymal state. This shift is manifest in increased enrichment of hnRNPM binding sites downstream of skipped exon events specifically in the mesenchymal state. Additionally, hnRNPM's necessity in the EMT phenotype is clarified through gene set enrichment analysis where hnRNPM depletion upregulates epithelial-associated genes such as CDH1 which are typically suppressed during EMT while playing little role in the gain of mesenchymal expressed genes. Ultimately, this chapter describes how hnRNPM, a ubiquitously expressed RNA binding protein, regulates distinct transcriptional processes in a cell-state specific manner while retaining the ability to suppress epithelial-associated gene regulation in both cell-states.

Chapter 3 describes an investigation of the role of RNA secondary structures, specifically RNA G-quadruplexes, during EMT. We previously identified that RNA G-

quadruplexes serve as an important binding site for hnRNPF, a splicing factor which this chapter shows promotes epithelial-associated alternative splicing and antagonizes the EMT phenotype. RNA G-quadruplexes are enriched near hnRNPF-dependent splicing events, and hnRNPF shows a high enrichment of G-quadruplex structures near binding sites compared to other hnRNP family members. In addition, this chapter describes the effect of emetine, a small molecule we previously identified as a G-quadruplex destabilizer, and its ability to promote an EMT associated transcriptional signature. The dichotomous roles of hnRNPF, an RNA binding protein that antagonizes EMT through recognition of G-quadruplexes, and emetine, which promotes EMT through G-quadruplex destabilization, highlights the critical role RNA secondary structure plays in regulating alternative splicing and cell phenotypes.

Chapter 4 discovers the role of AKAP8, a non-traditional RNA binding protein, as an antagonist of EMT-associated alternative splicing through differential RNA binding topology in epithelial and mesenchymal states. Integration of AKAP8 depletion RNA sequencing datasets in epithelial and mesenchymal cells with AKAP8 crosslinking binding sites shows that AKAP8 specifically antagonizes EMT associated alternative splicing in the epithelial state although AKAP8 expression does not change much during EMT. This antagonism is linked to increased binding of AKAP8 to proximal intronic sequences, which are more highly enriched in splicing-regulatory cis-elements, in the epithelial state compared to the mesenchymal state. Thus, this chapter identifies AKAP8 as another RNA binding protein with cell-state specific binding properties and splicing regulation that polarizes the EMT phenotype towards the epithelial state.

Together, these four chapters interrogate a fundamental question of interest in the field of post-transcriptional regulation; how and to what extent does RNA metabolism contribute to cell-phenotype change and cell-state plasticity? Using EMT as a model cell-state transition with clinically relevant importance, especially relating to the underlying mechanism of cancer metastasis, this thesis provides new insight into how RNA binding proteins, through direct and indirect interaction with downstream regulatory targets, play diverse and nuanced roles in bridging genotype to phenotype.

ACKNOWLEDGEMENTS

As I complete this dissertation, I am reminded of the African proverb, “It takes a village.” I am deeply grateful to the community of individuals, a village in a sense, that has guided me during my PhD. It is difficult to express how thankful I am for the guidance I received along the way, but here I hope to set down some words to acknowledge the generous individuals who lit the path on this journey of knowledge.

I must first acknowledge my mentor and thesis advisor, Chonghui Cheng. I remember our earliest interactions when I was searching for a thesis advisor, and we instantly connected with our passion for science and inquiry. Chonghui has been an extraordinary role model for me; she is an investigator with a tireless work-ethic and enthusiasm for her career, colleagues, and trainees coupled with uncompromising standards and integrity for the quality of the work she produces. This conduct of research has had a profound influence on me. Although I am only now nearing the end of my PhD training, I feel inspired to ask meaningful questions, pursue these questions with concise and falsifiable hypotheses, and communicate and defend my conclusions to the larger scientific community and public. All of these skills rest in large part on the mentorship of Chonghui. She has always challenged me to reach farther than I thought I was capable of, and for this reason I have grown immensely as a scientist, research, and person under her supervision. Thank you, Chonghui, for instilling in me these values and principles that will guide my future endeavors.

I would also like to thank the members of my committee: Marcus Peter, Elizabeth McNally, Eva Gottwein, and Xinshu (Grace) Xiao. I sincerely appreciate the thoughtful

discussions and helpful feedback I received from each of you. Thank you for challenging me to articulate clear hypotheses to consequential questions, formulate rational conclusions, and ultimately defend their validity. I would like to offer special thanks to Eva Gottwein and Grace Xiao in particular. Eva served as a generous host by offering me space in her laboratory when I transitioned to working in both Chicago and Houston when my primary advisor Chonghui moved her lab to Baylor College of Medicine. By opening her laboratory to me, she helped me navigate the challenges of working in two separate research locations. I am sincerely grateful for her support and mentorship. Also, I would like to thank Grace for her generous mentorship in the field of bioinformatics. Without her support I would have never achieved the level of proficiency I now hold in computational biology, and I am deeply honored that she has served as a collaborating senior author on several of the publications I contributed to during my PhD.

I could never have completed this journey without the support of the members of the Cheng lab, both past and present. So much of science is a team effort, and that was especially true in my case. I have been honored to work closely with many talented men and women, and every published work I have contributed to has involved the efforts of multiple co-authors. It has been a pleasure working with and knowing everyone who worked in the lab, and I am thankful for all of their support and collaboration. I am especially grateful to Huilin Huang and Jing Zhang, two talented post-docs I worked closely with on the RNA secondary structure work discussed in Chapter 3. I am also very thankful for my close collaboration with Xiaohui Hu and Rong Zheng, a post-doc

and graduate student in the lab, respectively, with whom the work in Chapters 2 and 4 was conducted.

During my PhD training I had the unique opportunity to work first at Northwestern University and then at Baylor College of Medicine. I am thankful for the support I received from the administration at both of these institutions, including the Northwestern MSTP office and DGP office as well as the administrators at the BCM Breast Center. I am very grateful for the support of the members of Eva Gottwein's lab for welcoming me into their lab during my intervals working at Northwestern in the latter half of my PhD. I am also indebted to the friends I made in the BCM Breast Center, with whom I shared illuminating conversations and experiences, both in and out of lab.

I would also like to recognize my undergraduate research mentors who instilled in me a love of science and molecular biology and motivated me to pursue a PhD. Special thanks goes out to Mark Forsyth for introducing me to the beauty of biomedical research in his microbiology laboratory at the College of William & Mary. I would also like to thank Michael Frodyma and Kevin Mann at Novozymes biologicals for showing me the perspective of industry research. Also, I would like to thank Adam Yuan and Qinqin Fu at the National University of Singapore for mentoring me during my study abroad and teaching me the art of structural biology.

To my wonderful friends in the Dungeons and Dragons group, with whom I met together in fellowship nearly every weekend, be it in person or over the internet during my time at BCM, please accept my gratitude. Without your support and compassion, I do not know where I would be today. Each of you have taught me the importance of

friendship in ways I could not have dreamed of when I first moved to Chicago, and I am so glad that we stuck together though times thick and thin over our long years of training. I also want to thank my old college roommate, Kobie Gordon, who has been a source of companionship and support over the years after undergraduate graduation. Cheers to friendship!

I would like to thank the members of my family. I thank my mother Ellen, my father Emerson, my brother John, and my sister Deborah for their unconditional love and support over the years of my life which has culminated in this achievement. I would be remiss to mention the endless love and encouragement from my mother-in-law Michiko, my father-in-law Khershed, and my brother-in-law Jahan for the grounding influence they have had and the tremendous growth I have experienced since they welcomed me into their family.

My final words are for my wife, Vivian. Without your love and support, especially during our years apart when the lab moved from Chicago to Houston, I am certain I would not have completed this work. Thank you from the bottom of my heart for the sacrifices you have made in support of my dream to become a physician scientist. Oh, how glad I am that our paths crossed and now we walk together!

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: Coregulation of alternative splicing by hnRNPM and ESRP1 during EMT

CHAPTER 2: hnRNPM regulates alternative splicing during EMT in a cell-state specific manner

LIST OF TABLES

CHAPTER 1: Coregulation of alternative splicing by hnRNPM and ESRP1 during EMT

*Adapted from the following published manuscript:

Harvey, SE, Xu, Y., Lin, X., Gao, X.D., Qiu, Y., Ahn, J., Xiao, X., Cheng, C. (2018). Coregulation of alternative splicing by hnRNPM and ESRP1 during EMT. RNA. PMID: 30042172

ABSTRACT

The epithelial-mesenchymal transition (EMT) is a fundamental developmental process that is abnormally activated in cancer metastasis. Dynamic changes in alternative splicing occur during EMT. ESRP1 and hnRNPM are splicing regulators that promote an epithelial splicing program and a mesenchymal splicing program, respectively. The functional relationships between these splicing factors in the genome-scale remain elusive. Comparing alternative splicing targets of hnRNPM and ESRP1 revealed that they coregulate a set of cassette exon events, with the majority showing discordant splicing regulation. Discordant splicing events regulated by hnRNPM show a positive correlation with splicing during EMT, however concordant events do not, indicating the role of hnRNPM in regulating alternative splicing during EMT is more complex than previously understood. Motif enrichment analysis near hnRNPM-ESRP1 coregulated exons identifies guanine-uridine rich motifs downstream of hnRNPM-repressed and ESRP1-enhanced exons, supporting a general model of competitive binding to these cis-elements to antagonize alternative splicing. The set of coregulated exons are enriched in genes associated with cell-migration and cytoskeletal reorganization, which are pathways associated with EMT. Splicing levels of coregulated exons are associated with breast

cancer patient survival and correlate with gene sets involved in EMT and breast cancer subtyping. This study identifies complex modes of interaction between hnRNPM and ESRP1 in regulation of splicing in disease-relevant contexts.

INTRODUCTION

Alternative RNA splicing is a fundamental mechanism of functional genome diversity that enables the nearly 21,000 protein-coding genes in the human genome to give rise to over 100,000 transcripts [1, 2]. Deep transcriptome sequencing has revealed that approximately 95% of all human multi-exon transcripts can undergo alternative splicing, positioning alternative splicing as a critical form of post-transcriptional gene regulation in a variety of cellular and biological processes [3-5]. Dysregulation of alternative splicing is increasingly implicated in a variety of human diseases, including cancer progression and survival [6, 7].

Alternative splicing has emerged as a central regulatory process during the epithelial-mesenchymal transition (EMT) [8-12]. EMT is a developmental program whereby epithelial cells transit to a mesenchymal phenotype, which occurs in natural processes such as organogenesis and wound healing [13, 14]. A mounting body of evidence suggests that EMT is aberrantly activated in cancer cells to mediate tumor recurrence and metastasis [15, 16]. Study of the molecular mechanism of EMT has been largely restricted to cellular signaling and transcriptional regulation. Recently, work from our group has demonstrated that alternative splicing of the gene CD44 causally contributes to EMT and breast cancer metastasis [10, 11, 17, 18]. Further evidence has also emerged to show the essential role of alternative splicing of other genes in controlling EMT [19, 20]. A variety of splicing regulatory proteins have also been implicated in EMT alternative splicing, however few have been shown to have essential functional roles during EMT [9, 12, 18, 21].

Investigating the mechanisms underlying the regulation of CD44 alternative splicing led us to identify antagonistic roles between two splicing factors, heterogeneous nuclear ribonucleoprotein M (hnRNPM) and epithelial splicing regulatory protein 1 (ESRP1). hnRNPM promotes CD44 variable exon skipping and favors a mesenchymal phenotype, whereas ESRP1 stimulates CD44 variable exon inclusion and promotes an epithelial cellular state [10, 11, 18, 22]. In addition, a recent study showed that hnRNPM promotes a set of alternative splicing events associated with cell-survival and resistance to inhibition of the PI3K-Akt pathway, two traits that are associated with EMT [23, 24]. Interestingly, hnRNPM is ubiquitously expressed but functions in a mesenchymal cell-state specific manner to regulate CD44 alternative splicing. This cell-state restricted activity of hnRNPM is guided in part by competition with ESRP1, which is expressed in epithelial tissues but not in mesenchymal tissues [18, 25, 26]. hnRNPM and ESRP1 share common guanine-uridine-rich (GU-rich) binding sites [27, 28], and the presence of ESRP1 suppresses the activity of hnRNPM by binding to the same GU-rich cis-elements near CD44 variable exons [18]. Given the critical roles hnRNPM and ESRP1 play in modulating EMT, we hypothesized that these two splicing factors compete to regulate not only CD44 alternative splicing, but also many other splicing events which may be associated with EMT. The balance between hnRNPM and ESRP1 splicing regulation may therefore control the phenotypic switch between an epithelial state and a mesenchymal state.

In this study, we analyzed splicing events coregulated by both hnRNPM and ESRP1. Our results show that hnRNPM and ESRP1 exhibit inverse activities in regulating most coregulated splicing events. Unexpectedly, they also display concordant activities

when regulating a subset of coregulated splicing events. Importantly, our results reveal that hnRNPM and ESRP1 regulate a set of cassette exons to promote and inhibit EMT, respectively. Cassette exons regulated discordantly by hnRNPM and ESRP1 are enriched in GU-rich motifs specifically in the downstream intron, corresponding with known hnRNPM and ESRP1 binding motifs and likely sites of competitive splicing regulation. This competitive mode of regulation is more widespread than previously appreciated. Coregulated cassette exons also stratify breast cancer patients by overall survival and correlate with cancer-relevant gene sets, highlighting the importance of hnRNPM and ESRP1 splicing regulation in cancer biology.

METHODS

Cell Lines

Maintenance of immortalized human mammary epithelial cells (HMLE) cells was conducted as previously described [10]. Human embryonic kidney 293FT, human breast carcinoma MDA-MB-231, and MDA-MB-231 metastatic derivative lines 4175 (LM2) and 1833 (BM1) were grown in DMEM supplemented with 10% FBS, L-glutamine, penicillin, and streptomycin. The ESRP1-HA overexpressing MDA-MB-231 cell line was described previously [11].

Plasmids, shRNAs, and ESRP1 overexpression

Two expression plasmids, pcDNA3-FLAG-hnRNPM and pcDNA3-FLAG-ESRP1 were subcloned from pECFP-hnRNPM [29] and pBRIT-ESRP1 [10], respectively. The CD44 variable exon 5 minigene was described previously [10]. The MARK3 exon 17 minigene was constructed through PCR amplification of MARK3 exon 17 and approximately 400 nucleotides of flanking intron followed by cloning into the BamH1 site of the CD44v5 minigene. Primers for MARK3 minigene construction are listed in Table 1. The control, hnRNPM, and ESRP1 shRNAs were described previously [10, 18]. Control shRNA: 5'-CCCGAATTAGCTGGACACTCAA-3'. hnRNPM shRNA: 5'-GGAATGGAAGGCATAGGATTT-3'. The ESRP1 shRNA was obtained from Open Biosystems (clone V2LHS_155255) with sequence 5'–CGCATAAGATCTTGGAATAATA-3'.

Transfection, semi-quantitative RT-PCR, and qRT-PCR (qPCR)

Briefly, 2.25 x 10^5 HEK293FT cells were plated in 24-well plates twenty-four hours prior to transfection. Co-transfection of hnRNPM and/or ESRP1 with splicing minigenes was performed using Lipofectamine 2000 (Invitrogen) per the manufacturer's instructions. For HEK293FT transfections, 100 ng of minigene was used. RNA was extracted from cells using the E.Z.N.A. Total RNA Kit (Omega Bio-Tek). RNA concentration was measured using a Nanodrop 2000 (Thermo Fisher Scientific). cDNA was generated via reverse transcription using the GoScript Reverse Transcription System (Promega) with 1 μL GoScript RT and 250 ng of RNA in a total volume of 20 μL followed by incubation at 25˚C for 5 min, 42˚C for 30 min, and 70˚C for 15 min. Semi-quantitative RT-PCR assaying for splicing products was performed using Hot StarTaq DNA polymerase (Qiagen), and PCR cycles were run for 30 or fewer cycles.

Primers for semi-quantitative analysis were designed in constitutive exons flanking each variable exon [30]. Semi-quantitative PCR generates both exon inclusion and skipping products which were separated through agarose gel electrophoresis. PCR product intensity was measured using ImageJ image analysis software. qPCR was performed using GoTaq qPCR Master Mix (Promega) per the manufacturer's instructions on a CFX Connect Real-Time PCR system (BioRad) using a two-step protocol and supplied software. For every qPCR sample per biological replicate, two technical replicates were performed and Ct counts were averaged. Normalization and quantification of qPCR data was done using the $2^{-\Delta Ct}$ method relative to TATA-binding protein (TBP) expression [31]. Primer specificity was verified with melt-curve analysis during qPCR.

Primers for semi-quantitative and qPCR analysis were designed to generate products spanning introns (Table 1).

Table 1: Primers for cloning, validation, and minigene construction

Primers for Minigene Creation	**Forward Primer Sequence (5'-3')**	**Reverse Primer Sequence (5'-3')**
MARK3 exon 17 minigene	CTTCCTGTTGGCAGTTCCATTC	GAACTCGAGTAAGCCCTGATCAC
Primers for ESRP1 knockdown qRT-PCR	**Forward Primer Sequence (5'-3')**	**Reverse Primer Sequence (5'-3')**
ESRP1	CAGAGGCACAAACATCACAT	AGAAACTGGGCTACCTCATTGG
TBP	GGAGAGTTCTGGGATTGTAC	CTTATCCTCATGATTACCGCAG
Primers for minigene semi-quantitative RT-PCR	**Forward Primer Sequence (5'-3')**	**Reverse Primer Sequence (5'-3')**
Minigene semi-quantitative primers	GAGGGATCCGGTTCCTGCCCC	CCAGCGGATAGAATGGCGCCG
Primers for endogenous semi-quantitative RT-PCR		
Gene/Isoform Name	**Forward Primer Sequence (5'-3')**	**Reverse Primer Sequence (5'-3')**
APLP2	ATGAGGAGAATCCTACTGAAC	TCATCTGCAGAGGTCTCGAA
ABI1	TCAGCCTCATGTTAATGGA	CTCATCTTCATAATCCACTGGT
ARHGAP17	GAAACGACTCTGACTCGGGG	GCAGATACAGGGTCCTTCGG
SPAG9	GTGGGGACAGGAAATGGTGT	AAGCTGTGCATGTGCCATTG
ARHGEF12	TGAACCGAGAGTCACCAACAG	ACCGTCAGCCCAAATCCATT
ARHGAP10	CGTGACTACAGCTGTCCCTG	GCTGTGTTCTGCTTCACACG
MYO1B	CACGAAGGGAACTGAGACGG	TCCAGCATTGGCTCTGAAGAA
PRRC2B	CTCAGATGGAGATGAAAGGC	TCCGGTTCGCTTCATCTG

Quantitative immunoblotting

Whole cell lysates or RNA pull-down samples were separated by 10% SDS-PAGE, transferred to a PVDF membrane (BioRad), and probed with the appropriate antibody.

Primary antibodies used in western blots included HA-HRP (Roche Applied Science), hnRNPM (OriGene). GAPDH (GE) and β-actin (Sigma-Aldrich) were used as loading controls. After incubation with HRP-tagged secondary antibodies, if appropriate, blots were visualized via chemiluminescence (Thermo-Fisher).

RNA Pull-down Assays

5'-biotinylated nucleotides were used for RNA pull-down experiments. The APLP2 exon 7 associated probes include GU1: 5'-biotin-UCUGUGUGGUGUCCCUGCCCACUC GGGUGUUUGCU, which was mutated to GU1 mut: 5'-biotin-UCUACGUAAUCUCCCUGCCCACUCGCCUGCAUGCU and GU2: 5'-biotin-CGUGUGUCUGGUGGUGCUUGGUGGUGAUGGUGC, which was mutated to GU2 mut: 5'-biotin- CGUACGUCUCCUGCAGCUAGCUAAUGAUACUCC. Biotinylated RNA oligos (10 μL at 40 μM) were immobilized on 50 μL of streptavidin beads (50% slurry; Thermo Fisher) in a total volume of 400 μL 1X binding buffer (20 mM Tris, 200 mM NaCl, 6 mM EDTA, 5 mM sodium fluoride, 5 mM β-glycerophosphate, 2 mg/mL aprotinin, pH 7.5) for 2 hours at 4˚C in a rotating shaker. After immobilization, beads were washed 3 times in 1X binding buffer, then 200 μg MDA-MB-231 or MDA-MB-231 ESRP1-HA cell lysates were suspended with the beads in 400 μL of 1X binding buffer for incubation at 4˚C overnight. Beads were then washed 3 times in 1X binding buffer, resuspended in 60 μL of 2X Laemmli sample buffer (Bio-Rad), and boiled for 5 min. Ten μL of sample was analyzed under denaturing conditions on 10% SDS-PAGE and detected via immunoblotting.

RNA Sequencing Analysis

Total RNA was extracted from LM2 or BM1 cells stably expressing control or hnRNPM shRNAs using Trizol, and poly-A-selected RNA-seq libraries were generated using TruSeq stranded mRNA library preparation kits (Illumina) and subjected to 100-base-pair PE stranded RNA-seq on an Illumina HiSeq 4000. RNA-seq reads were aligned to the human genome (GRCh37, primary assembly) and transcriptome (Gencode version 24 backmap 37 comprehensive gene annotation) using STAR version 2.5.3a [32] with the alignEndsType parameter set as EndToEnd. Differential alternative splicing of cassette exons was quantified using rMATS version 3.2.5 [33] and differential splicing events were collected using the following cutoffs: FDR < 0.05, |ΔPSI| ≥ 0.1, and average junction reads per cassette event ≥ 10. Control cassette exons not differentially spliced were defined by the following filters: FDR > 0.5, minimum PSI for control or hnRNPM knockdown samples < 0.85, maximum PSI > 0.15, and average junction reads per cassette event ≥ 10. RNA sequencing data for H358 shESRP1/2 was processed in the same way after retrieving raw reads from GEO record GSE75492. After rMATS analysis, cassette exons events were collapsed if they shared identical cassette exon coordinates. Other ESRP1-regulated cassette exons were obtained from the supplemental materials of published studies [12, 27]. EMT-regulated cassette exons were obtained from the supplemental materials of two previous studies [8, 9]. Sequencing datasets for hnRNPM are deposited in the Gene Expression Omnibus at GSE112516.

Motif Enrichment and RNA Motif Maps

Kmer enrichments were calculated using 250 bp of the sequence flanking the cassette exons. To avoid enrichment of canonical splice site motifs, 9 nucleotides downstream of 5' splice sites and 25 nucleotides upstream of 3' splice sites were removed. To assess enrichment of hexamers, a previously published motif analysis method was used [34]. For the GU-rich motif RNA map analysis, the top 12 6-mer motifs from an ESRP1-SELEX-Seq analysis were obtained [27]. The motifs were UGGUGG, GGUGGG, GUGGUG, GUGGGG, GUGUGG, GGUGUG, UGUGGG, GGUGGU, GUGGGU, UGGGGU, GGGGGU, UGGGGG. The 250 bp of sequence flanking upstream and downstream hnRNPM and ESRP1 regulated cassette exons as well as control exons obtained from the rMATS alternative splicing analysis were also obtained. The GU-rich motif score was computed in a custom python script by counting the number of nucleotides covered by any one of the GU-rich motifs in a sliding window of 50 bp shifted 1-nucleotide at a time across the 250 bp interval in all of the regulated and control cassette exons. The GU-rich motif score was set equal to the percent of nucleotides covered by the motifs in each of the sliding windows and plotted for regulated exons, stratified by inclusion or skipping, and control exons.

TCGA BRCA Survival Analysis, Genomic Alterations, GSEA, and Gene Ontology

Processed TCGA BRCA level 3 RNA Seq V2 data for exon junctions and gene expression were downloaded from the Genomic Data Commons Legacy Archive [35]. Cassette exon PSI values in each patient sample were calculated using the following

equation from the exon junction files: PSI = (Inclusion junction reads / 2) / ((Inclusion junction reads / 2) + (Skipping junction reads)). Patients were clustered into two groups using K-means clustering by PSI value. Kaplan-Meier survival analysis was conducted between these two groups using overall survival. p-values were computed using log-rank tests.

For GSEA analysis, correlation values between the given cassette exon PSI values and all genes in the TCGA BRCA RNA-seq V2 datasets were computed. Genes were then ranked by correlation and GSEA was performed using the Broad Institute javaGSEA desktop application [36, 37].

For analysis of genomic alterations in ESRP1 and hnRNPM, the TCGA BRCA Provisional dataset containing tumor samples with genome sequencing and copy number variation analysis (n = 963) was queried using cBioPortal (www.cbioportal.org). The genomic alterations figure was derived from the cBioPortal Oncoprint analysis page.

Gene ontology analysis was conducted using DAVID v6.8 with the gene list composed of all genes containing an exon in the 213 hnRNPM-ESRP1 coregulated exons [38, 39]. GOTERM enrichment was restricted to GOTERM BP-DIRECT, GOTERM MF-DIRECT, and GOTERM CC-DIRECT. The background gene set was composed of all genes in hnRNPM and ESRP1 RNA sequencing datasets with FPKM > 3 in at least one sample.

Statistical analyses

All data were presented as mean ± standard deviation, unless specifically indicated. Correlation was assessed using pearson correlation. Statistical significance tests included two tailed, unpaired Student's t-tests, Fisher's exact tests, hypergeometric tests, and log-rank tests. p-value < 0.05 was considered statistically significant. $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***) where indicated.

RESULTS

Splicing factors ESRP1 and hnRNPM coregulate a set of cassette exons

We previously showed that ESRP1 antagonizes the splicing activity of hnRNPM on CD44 alternative splicing and that hnRNPM stimulates EMT-associated splicing in a mesenchymal cell-type specific manner [18]. In an effort to better understand how ESRP1 and hnRNPM functionally interact with each other globally, we compared alternative splicing events from transcriptome-profiling datasets in response to hnRNPM or ESRP1 perturbation. We performed differential alternative splicing analysis when hnRNPM was perturbed by shRNA in two well-established mesenchymal cell lines: MDA-MB-231-derived lung-metastatic LM2 and bone-metastatic BM1 [40, 41]. We chose mesenchymal cell lines to investigate hnRNPM-mediated alternative splicing to avoid masking of hnRNPM function by ESRP1, which is not expressed in mesenchymal tissues [25, 26]. We focused on cassette exons as they are the most common form of alternative splicing [5]. Cassette exon splicing levels are reported using the Percent Spliced In (PSI) metric, which is a measure of the relative abundance of the exon inclusion isoform. We obtained a set of 1635 hnRNPM-regulated alternative cassette exons by taking the union of all significantly regulated exons (FDR < 0.05, $|\Delta PSI| \geq 0.1$, average junction reads per cassette event ≥ 10) after hnRNPM knockdown in LM2 and BM1 cells. Using the union of cassette exons identified from previously published alternative splicing profiling datasets for ESRP1, including ESRP1 overexpression in MDA-MB-231 cells and ESRP1/2 knockdown in prostate epithelial PNT2 and lung non-small cell carcinoma H358 cells [9, 12, 27], we derived a corresponding set of 1300 ESRP1-regulated cassettes. Although a

variety of cell lines were considered in our analysis, out of the 1635 cassette exons significantly regulated by hnRNPM, 1548 (95%) were detected in a background set of all exons expressed in the ESRP1 datasets. Similarly, out of the 1300 cassette exons regulated by ESRP1, 1124 (86%) were detected in a background set of exons from the hnRNPM datasets. These results show that the combined datasets, although encompassing different cell types, contained the vast majority of assayed cassettes and were suitable for comparative analysis.

To determine cassette exons regulated by both hnRNPM and ESRP1, we intersected the hnRNPM-regulated cassettes with ESRP1-regulated cassettes and obtained a statistically significant overlap of 213 coregulated cassette exons (Fig. 1A, , p = 7.9e-115, hypergeometric test). Cassette exons where hnRNPM and ESRP1 both promote exon inclusion or exon skipping were defined as concordant exons. Cassette exons where hnRNPM and ESRP1 promote opposite splice isoforms were defined as discordant exons. Nearly two-thirds of the coregulated exons (134/213, 63%) were regulated discordantly while the remaining exons (79/213, 37%) were regulated concordantly (Fig. 1A-B). The fact that the majority of coregulated events show discordant regulation mirrors the antagonistic role that ESRP1 and hnRNPM play in favoring cell-state specific splicing programs [18]. By contrast, the concordant splicing regulation by ESRP1 and hnRNPM suggests that they cooperate to control splicing within a subset of genes. Thus, the coregulation of splicing between ESRP1 and hnRNPM is not purely antagonistic and may be more complex than previously understood.

As hnRNPM and ESRP1 play important roles in regulating alternative splicing during EMT, we overlapped the coregulated exons with a set of EMT regulated alternative splicing events derived from previous studies [8, 9]. Over 30% (67/213) of the coregulated exons overlap with EMT, representing a statistically significant overlap (Fig. 1C, p = 2.63e-80, hypergeometric test). We then analyzed the regulatory roles of hnRNPM and ESRP1 on these EMT-associated splicing events. For both discordant and concordant exons coregulated by ESRP1 and hnRNPM, ESRP1-mediated splicing inversely correlated with the EMT-associated splicing, in line with the role of ESRP1 as an epithelial specific splicing regulator (Fig. 2A-C) [12, 25, 26]. Interestingly however, hnRNPM showed bi-directional correlation.

For the hnRNPM-ESRP1 discordant exons, we observed a positive correlation between hnRNPM-mediated splicing and EMT-associated splicing, indicating that hnRNPM promotes splicing that occurs during EMT (Fig. 2D). For the hnRNPM-ESRP1 concordant exons, however, hnRNPM's activity inversely correlated with EMT-associated splicing (Fig. 2E). This bi-directional difference in EMT-splicing regulation was statistically significant (Fig. 2F). These results suggest that although the majority of hnRNPM-regulated events are consistent with its role in driving a mesenchymal splicing program in opposition to ESRP1, hnRNPM may also be involved in a small subset of splicing events regulated in favor of an epithelial splicing pattern when functioning in concert with ESRP1.

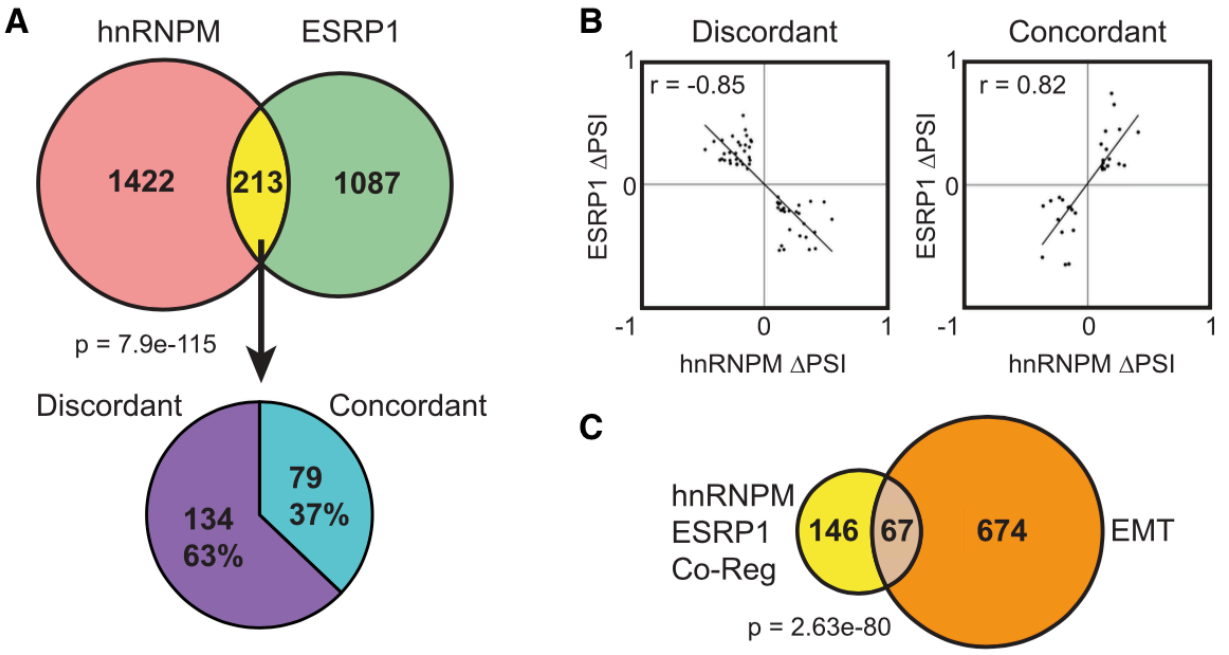

FIGURE 1. hnRNPM and ESRP1 coregulate a set of cassette exons in discordant and concordant manners. (A) RNA-sequencing analysis of hnRNPM knockdown and ESRP1 perturbation data sets identified splicing factor–dependent cassette exon events. In total, 213 cassette exon events showed overlapping regulation by hnRNPM and ESRP1 (p-value by hypergeometric test). The majority, 63% (134/213) cassette exons, were regulated discordantly by hnRNPM and ESRP1. The minority, 37% (79/213) cassette exons, were regulated concordantly. (B) Discordant exons show a negative correlation of hnRNPM versus ESRP1 ΔPSI splicing changes, while concordant exons show a positive correlation. Positive ΔPSI indicates promotion of exon skipping, while negative ΔPSI indicates promotion of exon inclusion. (C) 67 of the hnRNPM-ESRP1 coregulated cassette exons overlap with cassette exons regulated during EMT (P-value by hypergeometric test).

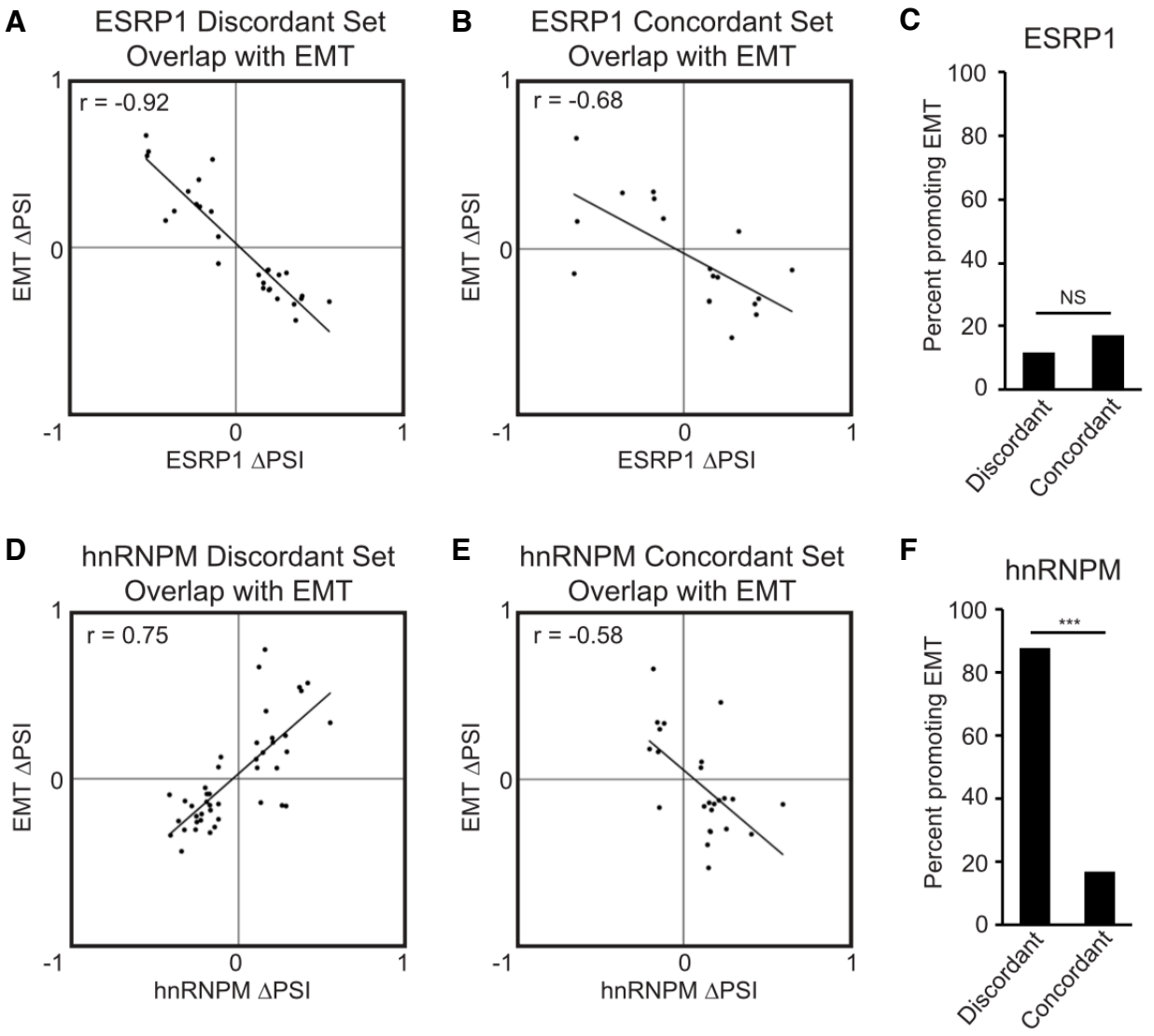

Figure 2: hnRNPM and ESRP1 show divergent splicing regulatory relationships with EMT. Both ESRP1 discordantly regulated exons (A) and concordantly regulated exons (B) show a negative correlation with EMT splicing. Positive and negative ΔPSI indicate promotion of exon skipping and inclusion, respectively. (C) 12% of ESRP1-regulated discordant exons promote EMT compared to 17% of concordant exons. (D) hnRNPM discordantly regulated exons show a positive correlation with EMT. (E) hnRNPM concordant exons show a negative correlation with EMT. (F) 88% of hnRNPM discordant exons promote EMT compared to 17% of hnRNPM concordant exons ($P < 0.001$ by Fisher's exact test).

Validation of ESRP1 and hnRNPM coregulated cassette exons

We chose to validate hnRNPM splicing activity in LM2 cells, which is a mesenchymal cell line that does not express ESRP1 (Fig. 3A), and we chose human mammary epithelial cells (HMLE) to validate ESRP1 activity where ESRP1 is robustly expressed (Fig. 3A). hnRNPM is expressed similarly in LM2 and HMLE cells (Fig. 3A). We experimentally validated four of the concordant and four of the discordant coregulated splicing events using RT-PCR upon shRNA-mediated hnRNPM or ESRP1 knockdown to confirm splicing regulation observed in the RNA sequencing studies (Fig. 3B-C). The validation set confirms that hnRNPM and ESRP1 coregulate splicing events in both concordant and discordant manners (Fig. 3C). Moreover, we examined the specificity of the hnRNPM and ESRP1 splicing regulatory relationships by using two splicing minigenes, with one containing a discordantly regulated exon at CD44 variable exon 5 and the other harboring a concordant exon at *MARK3* exon 17. Co-transfection experiments of the CD44v5 minigene with hnRNPM or ESRP1 in 293FT cells showed that hnRNPM promotes v5 exon skipping, whereas ESRP1 inhibits it (Fig. 4A), supporting discordant regulation of hnRNPM and ESRP1. By contrast, co-transfection of the concordant *MARK3* exon 17 minigene with hnRNPM or ESRP1 both resulted in dose-dependent increases in exon skipping, mirroring the concordant regulation of splicing observed in the RNA-sequencing data (Fig. 4B). These results show that hnRNPM and ESRP1 function in both discordant and concordant fashions that are dependent on splicing substrates.

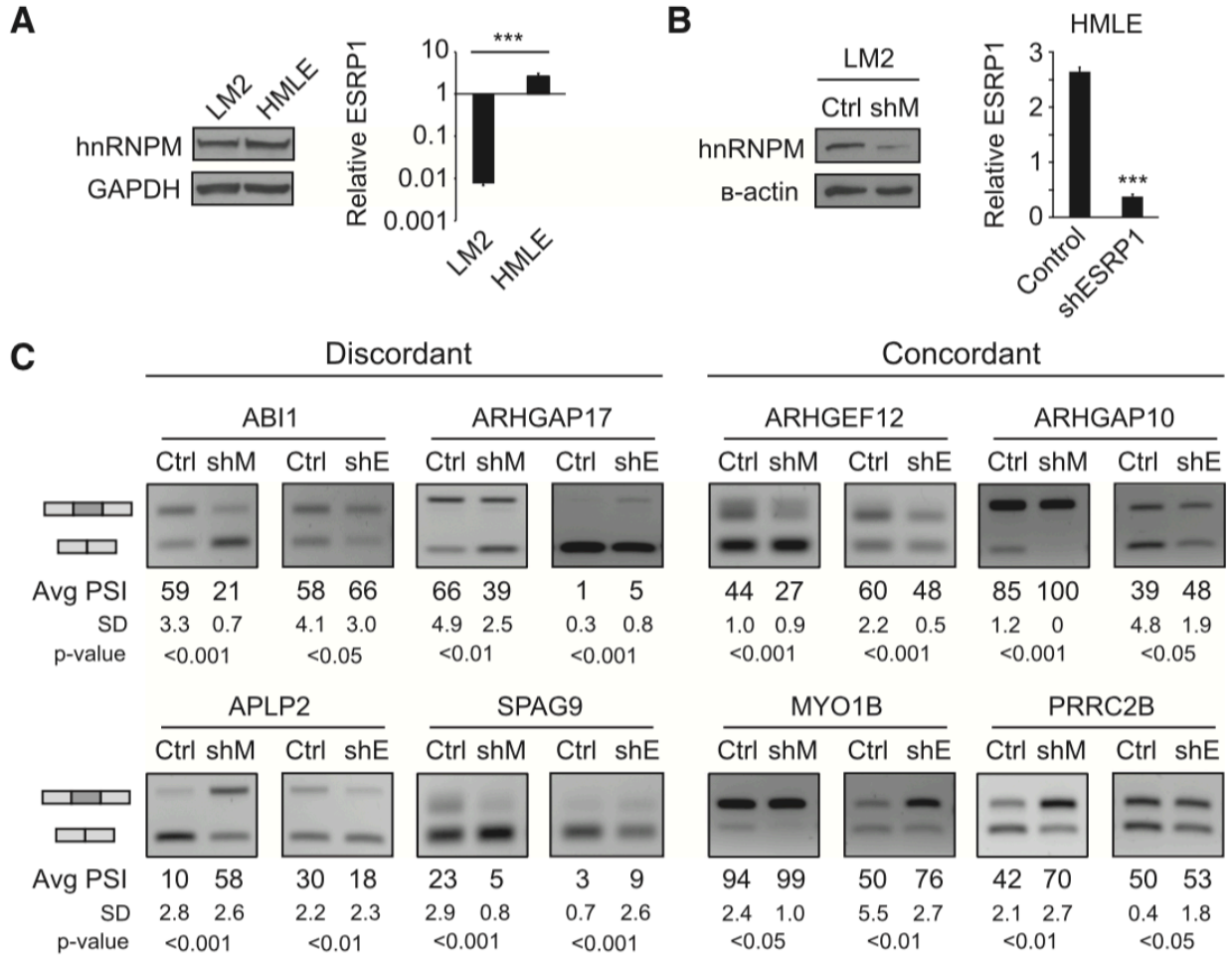

FIGURE 3. Validation of hnRNPM and ESRP1 coregulation of cassette exons. (A) Western blot showing expression of hnRNPM in LM2 and HMLE cells. qPCR showing expression of ESRP1 relative to TBP in LM2 and HMLE cells (Error bars = S.E.M, n = 3, $P < 0.001$). (B) Western blot showing hnRNPM knockdown in LM2 cells and qPCR showing ESRP1 knockdown relative to TBP in HMLE cells (error bars = S.E.M, n = 3, $P < 0.001$). (C) Cassette exon splicing events regulated by hnRNPM and ESRP1. hnRNPM validation was conducted in LM2 cells, while ESRP1 validation was conducted in HMLE cells. Avg PSI represents an average of three experiments. SD represents standard deviation of PSI values (P-value calculated by Student's t-test).

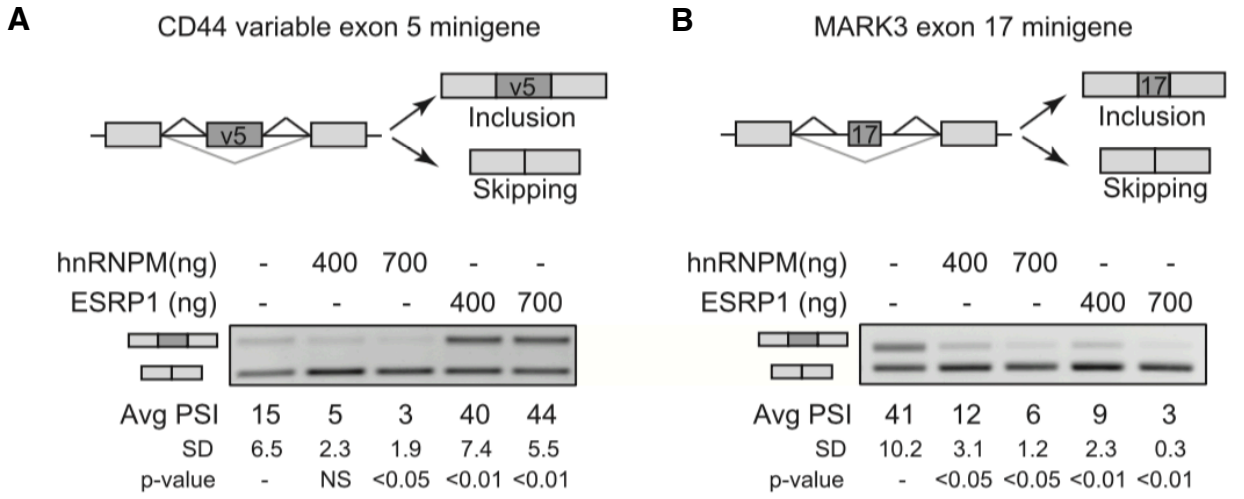

Figure 4: hnRNPM and ESRP1 show concordant and discordant regulation of splicing minigenes. (A) Splicing minigene analysis showing cotransfection of CD44 variable exon 5 minigene with hnRNPM in 293FT cells promotes exon skipping, while ESRP1 promotes exon inclusion. (B) Cotransfection of MARK3 exon 17 minigene and hnRNPM or ESRP1 in 293FT cells shows that both promote exon skipping. Avg PSI represents an average of three experiments. SD rep- resents standard deviation of PSI values (P-value calculated by Student's t-test).

ESRP1 and hnRNPM discordantly regulated exons are enriched in shared GU-rich binding sites

In order to better understand the functional relationship between hnRNPM and ESRP1 in coregulating alternative splicing, we performed motif enrichment analysis on the introns near all hnRNPM and ESRP1 coregulated splicing events (Fig. 5A). Both hnRNPM and ESRP1 are known to bind GU-rich cis-elements primarily in introns [9, 26-28]. We observed selective enrichment of multiple GU-rich hexamers downstream of hnRNPM-repressed and ESRP1-enhanced events, with both sets of events showing enrichment of two identical motifs, GUGUGU and GUGGUG, among the top three enriched motifs (Fig 5A-C). The observation that GU-motif enrichment was observed downstream of exons regulated oppositely by hnRNPM and ESRP1 suggests that GU-motifs are enriched in discordantly regulated exons. These data support a model where hnRNPM and ESRP1 compete for shared binding sites directly downstream of cassette exons to regulate alternative splicing antagonistically.

We also noted significant enrichment of a UGCAUG motif downstream of hnRNPM-enhanced and ESRP1-repressed events (Fig. 5A). This sequence corresponds to the well-known binding motif of the RBFOX family of RNA binding proteins, of which only RBFOX2 is expressed in the cell lines used in this study [42, 43]. In total, 67/213 (31%) of hnRNPM-ESRP1 coregulated splicing events contain a splice-site proximal UGCAUG motif within 250 nucleotides of a splice site. The motif was more enriched downstream of the cassette exons investigated, with 45/213 (21%) containing a motif downstream while 30/213 (14%) contained a motif upstream of the cassette exon.

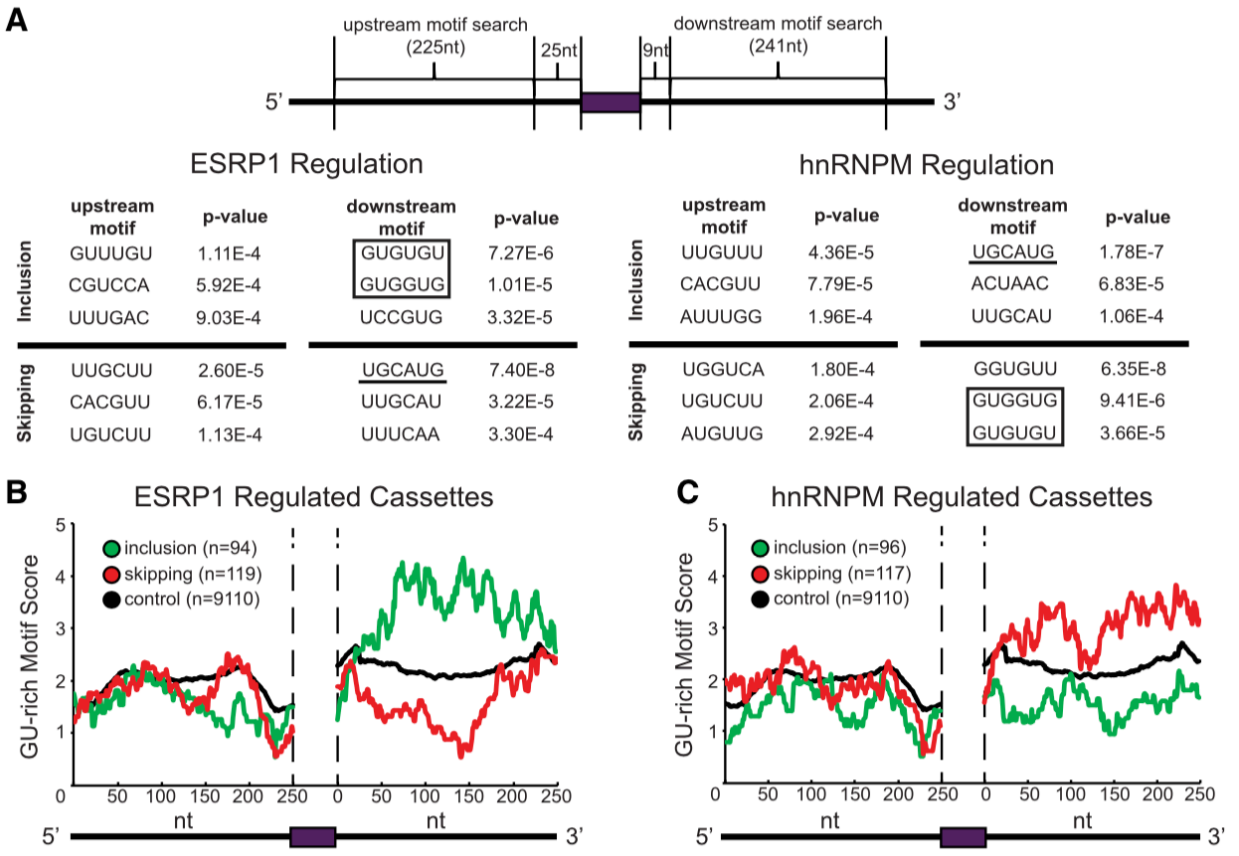

Figure 5. hnRNPM and ESRP1 show common motif enrichment downstream from discordantly regulated exons. (A) K-mer enrichment analysis showing the top three enriched 6-mers in introns flanking hnRNPM and ESRP1 coregulated cassette exons. Two identical GU-rich motifs (GUGUGU and GUGGUG, black box) were enriched downstream from cassette exons in ESRP1-enhanced and hnRNPM-repressed events. The RBFOX motif (UGCAUG, underlined) was enriched downstream from ESRP1-repressed and hnRNPM-enhanced events. (B,C) RNA motif map analysis of GU-rich motifs in introns flanking hnRNPM-ESRP1 coregulated exons with respect to ESRP1 regulation (B) and hnRNPM regulation (C) reveals enrichment of GU-rich motifs in downstream introns of ESRP1-enhanced and hnRNPM-repressed cassette exon splicing events. Inclusion events (green). Skipping events (red). Control events (black).

To experimentally examine the binding relationships of hnRNPM and ESRP1, we analyzed their ability to bind to the GU-rich motifs downstream of the discordantly regulated APLP2 cassette exon 7, which was validated in Fig. 3C. APLP2 contains multiple occurrences of GU-rich motifs identified from the motif analysis within 250 nucleotides downstream of APLP2 cassette exon 7 (Fig. 6A). RNA pull-down assays were conducted in MDA-MB-231 cells, which do not express ESRP1, and MDA-MB-231 cells ectopically expressing HA-tagged ESRP1 (Fig. 6B). Both cell lines express hnRNPM equally (Fig. 7E). As predicted from the RNA-seq data and validation experiments, an increase in APLP2 exon 7 PSI was observed upon ESRP1 overexpression (Fig. 3F). In order to examine binding of hnRNPM and ESRP1 to the GU-rich motifs downstream of APLP2 exon 7, we designed two 5' biotinylated RNA probes: GU1 and GU2 (Fig. 3G, top panel). hnRNPM binds GU2 much more strongly than GU1, while ESRP1 bound GU1 and GU2 relatively equally (Fig. 6D, bottom panel). GU2 contains more GU-nucleotide stretches than GU1, and this may have contributed to the binding differences of hnRNPM on these probes. As both ESRP1 and hnRNPM do not have clearly defined binding sites, instead binding to degenerate GU-rich sequences, there could be variations in their affinity to different GU repeats. These binding activities are specific because binding of both proteins was abolished upon disruption of GU-rich motifs (Fig. 6E). To determine whether hnRNPM binding is decreased in the presence of ESRP1, we compared relative hnRNPM binding on the GU2 probe in parental MDA-MB-231 cells to that in MDA-MB-231 cells overexpressing ESRP1-HA. By comparing the amount of hnRNPM that was associated with the GU2 probe relative to input, we found a 70% reduction in hnRNPM

binding in the ESRP1-HA-expressing MDA-MB-231 cells (Fig. 6E), suggesting that ESRP1 is capable of competing with hnRNPM for the same binding sites on the APLP2 pre-mRNA.

ESRP1 and hnRNPM coregulated exons are enriched in EMT processes and correlate with breast cancer signatures and patient survival

In order to better understand the relevance of the complex regulation of splicing by hnRNPM and ESRP1 to disease phenotypes, we performed gene ontology analysis using DAVID on the set of 213 coregulated exons [38, 39]. We observed significant GOTERMs associated with cell polarity, cell adhesion, and cytoskeletal dynamics (Fig. 7A), all processes that are critical for EMT. To assess the contribution of hnRNPM and ESRP1 coregulation of splicing in breast cancer patient samples, we mined the publicly available The Cancer Genome Atlas (TCGA) RNA-sequencing data for invasive breast carcinoma (BRCA) and calculated the PSI values for hnRNPM and ESRP1 coregulated splicing events. After stratifying the patients based on PSI levels for each exon via 2-means clustering, we observed alternative splicing events that are positively or negatively correlated with patient overall survival (FDR < 0.05, log rank test). We found that increased *SPAG9* exon 24 inclusion levels predict poorer patient survival while increased *ZMYND8* exon 22 inclusion is associated with a better prognosis in breast cancer (Fig. 7B). *SPAG9* exon 24 PSI increases during EMT, with exon inclusion stimulated by hnRNPM and inhibited by ESRP1. Conversely, *ZMYND8* exon 22 exon skipping is promoted by hnRNPM and antagonized by ESRP1.

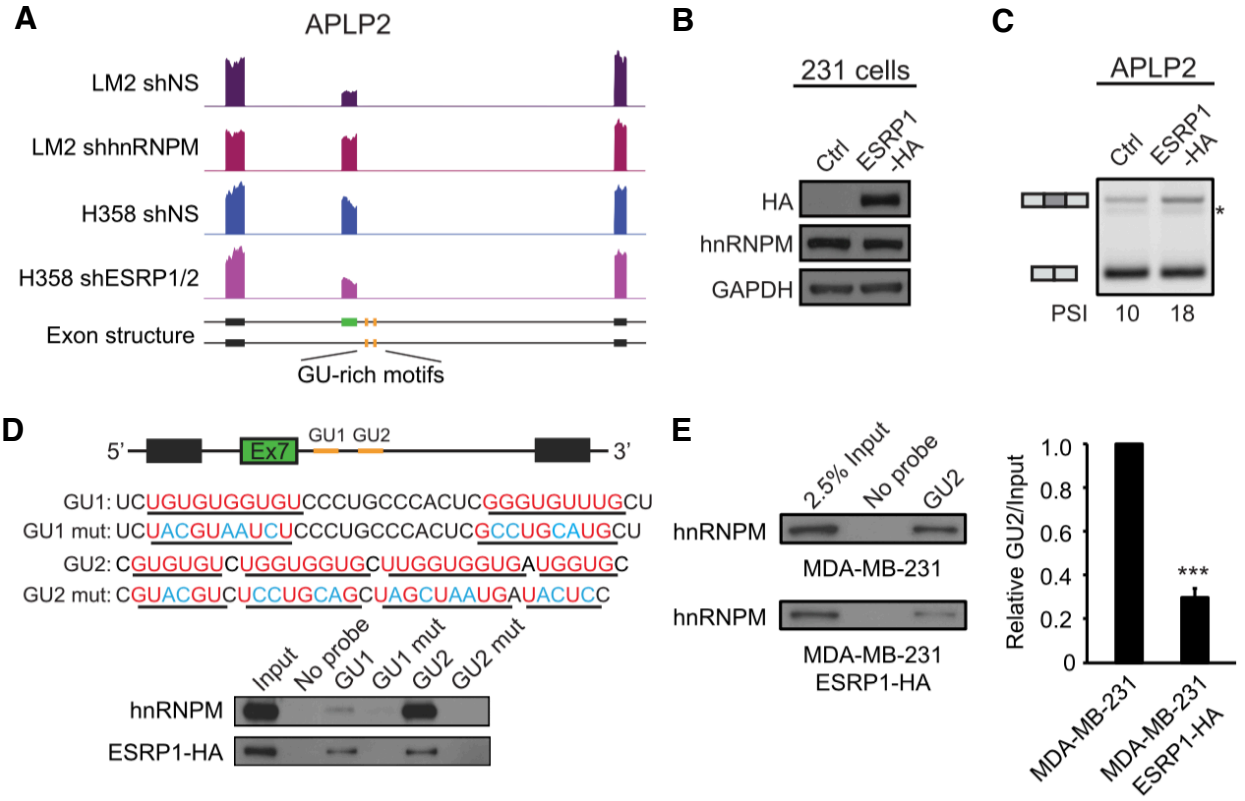

Figure 6: hnRNPM and ESRP1 compete for binding sites near APLP2 cassette exon 7. (A) Genome browser plot of RNA sequencing data sets showing hnRNPM knockdown promotes APLP2 exon 7 inclusion and ESRP1 depletion promotes skipping. Black bars indicate constitutive exons. Green bar indicates variable exon 7. Yellow bars indicate location of two clusters of GU-rich motifs within 250 nucleotides (nt) downstream from APLP2 exon 7. (B) Immunoblot of HA-tagged ESRP1 overexpression in MDA-MB-231 cells and endogenous hnRNPM. (C) ESRP1-HA overexpression in MDA-MB-231 cells results in increased APLP2 exon 7 inclusion. (∗) Indicates nonspecific band. (D) (Upper panel) The two GU-rich regions identified within 250 nt downstream from APLP2 exon 7 were used to design RNA probes GU1 and GU2 containing stretches of GU nucleotides underlined and in red with mutant probes GU1-mut and GU2-mut with mutated sequences colored in blue. (Lower panel) RNA pull- down analysis using RNA probes blotting for endogenous hnRNPM and overexpressed ESRP1-HA in MDA-MB-231 cells shows that hnRNPM and ESRP1 bind common GU-rich sequences. (E) RNA pull-down experiments using a static amount of the biotinylated GU2 RNA probe and cell lysate assaying for

hnRNPM in the MDA-MB-231 cell line, which does not express ESRP1, and the same line with ESRP1-HA overexpression. A total of 2.5% input was provided as a loading control for both samples. Overexpression of ESRP1 leads to less hnRNPM binding, suggesting that ESRP1 competes for the same GU2 binding site. (Error bars = S.E.M, n = 3, [$***$] P-value < 0.001 by Student's t-test).

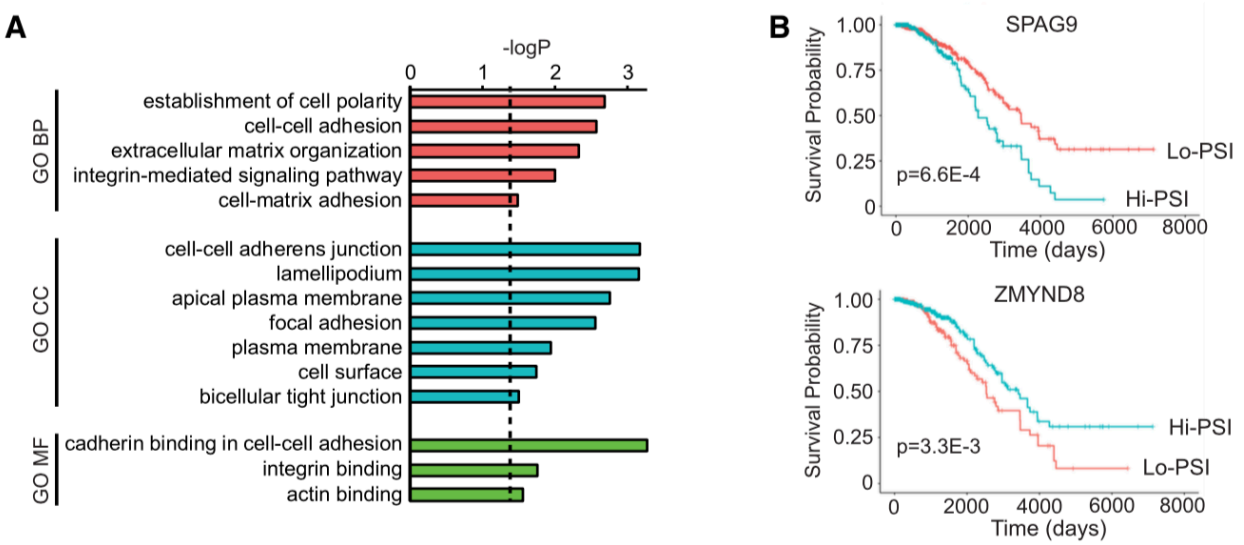

Figure 7. hnRNPM-ESRP1 coregulated exons are associated with EMT associated gene ontology terms and predict breast cancer survival. (A) Gene ontology analysis of genes that contain hnRNPM-ESRP1 coregulated exons identified significant terms associated with cell polarity, adhesion, migration, and the cytoskeleton. Direct GOTERMS BP, MF, and CC were queried using DAVID. (B) PSI levels of SPAG9 exon 24 and ZMYND8 exon 22 stratify breast cancer patients by overall survival. p-value calculated by log-rank test.

Intrigued by the role these exons played in predicting breast-cancer patient survival, we performed Gene Set Enrichment Analysis (GSEA) on gene signatures that are correlated with *SPAG9* exon 24 and *ZMYND8* exon 22 inclusion. Within the TCGA dataset, a published EMT gene signature was positively correlated with *SPAG9* exon 24 inclusion, while this signature was negatively correlated with *ZMYND8* exon 22 inclusion [44] (Fig. 8A-B). Genes downregulated by mir-200, a potent repressor of EMT [45, 46], also showed a positive correlation with *SPAG9* exon 24 inclusion but a negative correlation with *ZMYND8* exon 22 inclusion. Moreover, genes downregulated in basal subtype and claudin-low subtype breast cancers, which are generally non-hormone dependent and resistant to conventional therapies [47, 48], were negatively correlated with *SPAG9* exon 24 inclusion. Conversely, these genes were positively correlated with *ZMYND8* exon 22 inclusion.

As hnRNPM and ESRP1 showed primarily opposing functions in splicing regulation, we investigated the occurrence of genomic alterations in the genes encoding these two proteins in 963 TCGA breast tumors where genome sequencing and copy number analyses were available (Fig. 8C). Alterations in ESRP1 occurred in 16% of patients compared to 1.7% with hnRNPM alterations, with the majority of alterations representing gene amplification through copy number variation. Amplification or mutations in these genes were mostly mutually exclusive, with only two patients showing amplification of both genes.

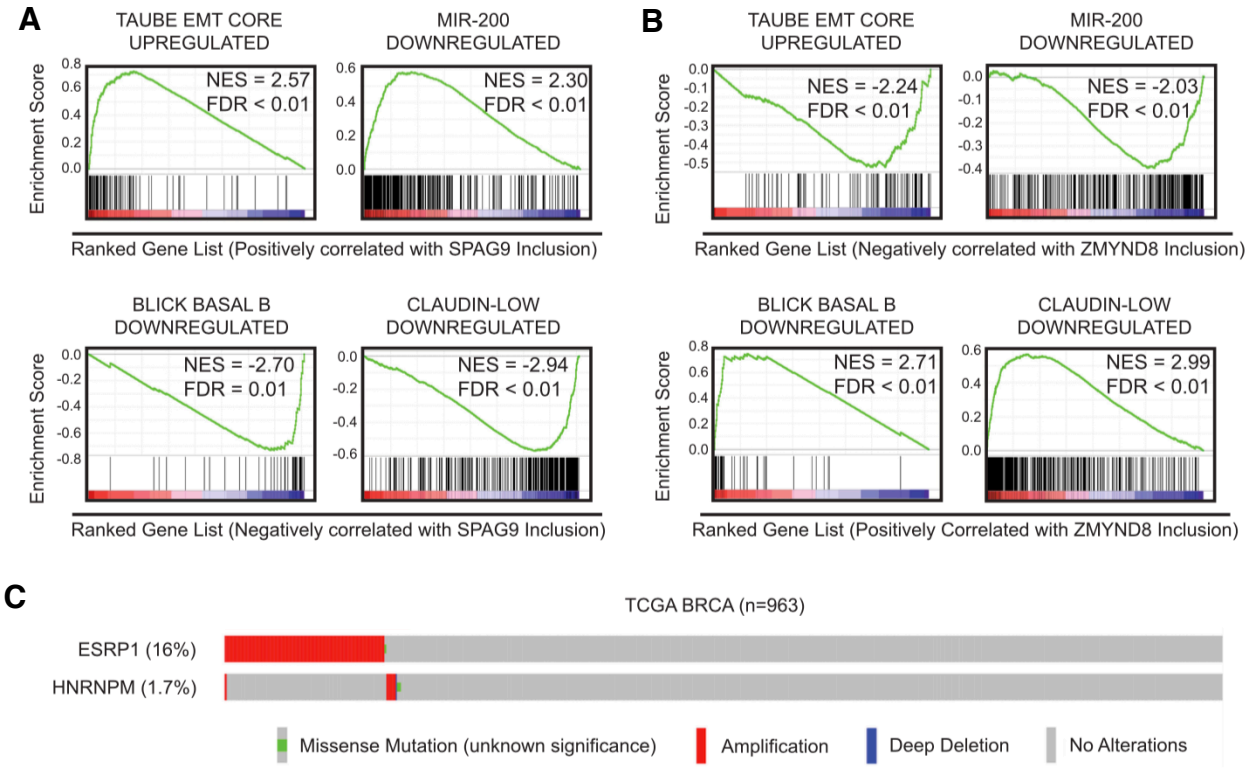

Figure 8: hnRNPM and ESRP1 coregulated splicing events correlate with EMT and invasive breast cancer gene signatures and show mutually exclusive genomic alterations. (A) Genes positively correlated with SPAG9 exon 24 inclusion are upregulated during EMT and downregulated by mir-200, while genes negatively correlated with SPAG9 exon 24 inclusion are downregulated in basal and claudin-low breast cancer subtypes. (B) Genes upregulated during EMT and downregulated by mir-200 are negatively correlated with ZMYND8 exon 22 inclusion, while genes downregulated in basal and claudin-low breast cancers are positively correlated with ZMYND8 exon 22 inclusion. (C) Visualization of TCGA BRCA tumors with different genomic alterations in ESRP1 or hnRNPM. Alterations in ESRP1 and hnRNPM are mostly mutually exclusive.

DISCUSSION

In summary, this study reveals widespread coregulation of alternative splicing by hnRNPM and ESRP1, both identified as key regulators of EMT splicing programs [12, 18]. hnRNPM-regulated cassette exons significantly overlap with ESRP1-regulated cassette exons, with the majority of coregulated events showing discordant splicing regulation, suggesting that hnRNPM and ESRP1 largely serve to functionally antagonize one another. We also observed a subset of splicing events regulated concordantly by hnRNPM and ESRP1 that inversely correlate with EMT splicing. Although these events represent a minority of coregulated events, these results suggest that hnRNPM is partly correlated with antagonistic regulation of splicing during EMT. These results are surprising as we observed that hnRNPM is required for cells to undergo EMT [18]. Interestingly, this scenario is reminiscent of that observed for RBM47, an RNA binding protein that inhibits EMT [49]. RBM47 showed primarily discordant regulation of splicing events compared to EMT, but also concordant regulation of a subset of splicing events that promote EMT [9]. These findings highlight the importance of understanding the combinatorial regulation of splicing by different factors with respect to a complex biological process such as EMT. Whether the hnRNPM-regulated splicing events that oppose EMT play a functional role during EMT or are important for regulating cellular processes that are highly active in epithelial cell states will be an interesting area for future study.

Some of the alternative splicing events coregulated by hnRNPM and ESRP1 have been investigated in detail to understand their functional contributions to EMT. CD44 contains multiple variable exons that undergo extensive alternative splicing during EMT.

Exon skipping of all variable exons of CD44 to generate the CD44s isoform is required for EMT and has been shown to promote Akt-signaling, mediate invadopodia activity, and attenuate degradation of EGFR to promote sustained RTK signaling, all of which have implications during EMT [10, 17, 50]. ESRP1 and hnRNPM directly regulate alternative splicing of CD44 in a discordant manner, with ESRP1 driving exon inclusion and hnRNPM promoting exon skipping. Another hnRNPM-ESRP1 coregulated splicing event that plays a role during EMT is alternative splicing of EXOC7 exon 7 and exon 8. Skipping of EXOC7 exon 7 and 8 produces a mesenchymal isoform capable of promoting actin polymerization and increased cell invasion compared to the epithelial isoform where part of exon 8 is included [20]. ESRP1 was shown to regulate EXOC7 splicing to promote production of the epithelial isoform, although the role of hnRNPM is not known. In addition, hnRNPM and ESRP1 coregulate exon skipping of TCF7L2 exon 4. The exon-4 skipped TCF7L2 isoform is capable of greater activation of Wnt/β-catenin target gene promoters compared to inclusion isoforms [51], and this isoform is upregulated in the mesenchymal state. It is worth noting that in our study the CD44 and TCF7L2 splicing events were discordantly regulated by hnRNPM and ESRP1, with ESRP1 promoting production of the epithelial isoform and hnRNPM promoting the mesenchymal splicing pattern. While the precise molecular consequences of most of the hnRNPM-ESRP1 regulated splicing events during EMT are unknown, our observation that hnRNPM-ESRP1 coregulated splicing events are enriched in genes associated with cytoskeleton remodeling and cell adhesion present interesting avenues for further study.

Analysis of motifs near hnRNPM and ESRP1 coregulated cassettes identified enrichment of GU-rich motifs downstream of hnRNPM-repressed and ESRP1-enhanced events. Previous studies of hnRNPM-mediated alternative splicing identified enrichment of GU-rich sequences and hnRNPM binding sites by cross-linking and immunoprecipitation (CLIP) analyses specifically near hnRNPM-repressed but not hnRNPM-enhanced events, which is consistent with our study [28]. Similarly, GU-rich motifs have been previously identified downstream of ESRP1-enhanced and upstream of ESRP1-repressed exons, indicating position-specific function of ESRP1 in regulating alternative splicing [9, 12, 26, 27]. The enrichment of GU-rich motifs in hnRNPM-ESRP1 discordant exons supports a model of competitive binding between the two proteins, which is supported by our binding analysis downstream of variable exon 7 of APLP2 as well as our previous study of CD44 alternative splicing [18].

We also identified striking enrichment of UGCAUG, the motif recognized by the RBFOX family of RNA binding proteins, near hnRNPM-ESRP1 coregulated cassettes. RBFOX motifs have been previously shown to enrich near exons regulated by hnRNPM [52] and ESRP1 [9, 12, 27]. RBFOX proteins were recently shown to function in a complex with hnRNPM, along with other splicing factors, where the presence of both factors was necessary for their full splicing activity [52]. We found a small but significant overlap (approximately 7%, $p = 4.52E-7$, Fisher's exact test) between the hnRNPM-ESRP1 coregulated exons and a recently published set of cassette exons regulated by RBFOX2 in HeLa cells [42]. These preliminary results suggest that RBFOX2 is involved in regulating the set of hnRNPM-ESRP1 coregulated exons. Since RBFOX2 has been

shown to promote alternative splicing during EMT [8, 21, 53], we speculate that RBFOX2 may function in concert with hnRNPM and against ESRP1 to regulate EMT-associated alternative splicing. A detailed analysis of RBFOX2's regulatory relationship with hnRNPM and ESRP1 is an interesting area for future study.

Through analysis of hnRNPM and ESRP1 coregulated cassette exons using breast cancer TCGA data, we identified splicing events that stratify breast cancer patients by overall survival and observed correlations with EMT and breast cancer subtype gene sets. These data indicate that hnRNPM and ESRP1 splicing function is associated with breast cancer progression and prognosis. We also show that in breast tumors few genomic alterations exist in hnRNPM while copy number variations in ESRP1 are present in a subset of tumors, and these alterations appear to be mutually exclusive. We previously showed that hnRNPM expression positively correlates with breast tumor grade [18], and others have shown that hnRNPM expression promotes tumor aggressiveness and predicts poor prognosis [54]. The relationship between ESRP1 expression and cancer progression is more complicated, with previous studies reporting a positive correlation between ESRP1 expression and longer patient survival in colorectal, pancreatic, and breast cancers [55, 56]. Conversely, other studies have shown directly opposite correlations with ESRP1 and survival in breast cancers, ovarian cancers, and melanomas [57-59]. These conflicting results implicate both increased and decreased ESRP1 expression in tumor progression and suggest that ESRP1 function in cancer is highly tissue and context dependent.

In-depth analysis of hnRNPM and ESRP1 splicing regulation informs our understanding of splicing factor binding and functional dynamics in the context of disease-relevant splicing programs and indicates the importance of understanding the competitive and cooperative mechanisms of splicing regulation that allow precise modulation of alternative splicing. Taken together, we show that coregulation of alternative splicing by hnRNPM and ESRP1 is wide-spread and primarily antagonistic, although a subset of events is regulated concordantly. Furthermore, we demonstrate that in controlling hnRNPM-ESRP1 discordantly regulated events, hnRNPM promotes alternative splicing in the same direction as EMT. hnRNPM and ESRP1 splicing antagonism is explained, at least in part, by competition for GU-rich elements downstream of coregulated exons. Lastly, hnRNPM-ESRP1 coregulated splicing events correlate with EMT and breast cancer-associated gene sets and predict breast cancer patient survival. Taken together, this study highlights the complex regulation of alternative splicing by ESRP1 and hnRNPM as well as the relevance of this regulatory interaction in EMT and cancer.

CHAPTER 2: hnRNPM regulates alternative splicing during EMT in a cell-state specific manner

*This work was completed in collaboration with Rong Zheng

ABSTRACT

The epithelial-mesenchymal transition (EMT) is a fundamental developmental program that is abnormally activated in cancer cells, resulting in therapeutic resistance and metastasis. Dramatic changes in alternative splicing occur during EMT, however the master regulators of this critical regulatory layer are poorly understood. We previously identified the ubiquitously expressed RNA binding protein hnRNPM as a causal driver of EMT and breast cancer metastasis whereby hnRNPM drives alternative exon skipping of the cancer stem cell marker CD44 in a mesenchymal-cell-state restricted manner. However, the scope of hnRNPM-mediated regulation of the EMT post-transcriptional landscape and the mechanisms by which hnRNPM functions in a cell-state specific fashion remain uncertain. By integrating breast epithelial and mesenchymal hnRNPM-depletion RNA-sequencing datasets coupled with individual nucleotide crosslinking and immunoprecipitation of hnRNPM, we identified cell-state specific hnRNPM transcriptional and alternative splicing programs correlated to hnRNPM binding patterns. hnRNPM seems to play a role in suppressing epithelial specific genes in both cell states, however the hnRNPM epithelial cell-state transcriptional program enriches for gene sets critical for the cell cycle while the mesenchymal signature involves pathways involved in inflammation. We found that hnRNPM shows a shift to promoting EMT-regulated splicing during EMT.

INTRODUCTION

The role of RNA binding proteins in determining the flow or information from genotype to phenotype within the vast network of post-transcriptional gene regulation is an area of intense study [60]. The mechanisms whereby RNA binding proteins are able to control cell plasticity and cell phenotypes is incompletely understood. In fact, approximately 80% of RNA binding proteins do not show tissue specific gene expression [61], however we and others have identified RNA binding proteins with limited changes in gene expression that play key roles in modulation of cell-state, with a specific focus on epithelial to mesenchymal cell transitions [8, 18, 26, 27, 62-65]. We previously identified the ubiquitously expression splicing factor hnRNPM as a cell-state restricted RNA binding protein critical for EMT and breast cancer metastasis [18]. hnRNPM drives CD44 exon skipping which is an essential alternative splicing switch during EMT [10], however due to competition with other RNA binding proteins such as ESRP1, hnRNPM's role in promoting EMT is most manifest in the mesenchymal cell state.

Competition among splicing factors and the importance of splicing factor complexes in regulation of alternative splicing has been increasingly demonstrated, including in the relationship between hnRNPM and RBFOX2 as components of the LASR complex of splicing regulators [52]. In addition, differences in transcriptome binding and specific pre-mRNA targets for splicing factors has also been observed. Of note is the differential binding of NOVA to targets in motorneurons vs. whole spinal cord in mouse despite similar expression in both cell types [66]. In this study, we were motivated to examine the scope of hnRNPM cell-state specific regulation of splicing during EMT

through integration of cell-state specific hnRNPM-depletion RNA sequencing datasets coupled with crosslinking and immunoprecipitation methods. By combining these methods, we sought to identify genes and alternative splicing events regulated by hnRNPM cell-state dependent and independent manners. Specifically, we sought to understand how hnRNPM might function in a mesenchymal-specific manner in regulating EMT alternative splicing. By intersecting hnRNPM binding sites with hnRNPM cell-state specific splicing events, we have uncovered shifts in hnRNPM function towards the mesenchymal phenotype correlated with increased hnRNPM binding activity near hnRNPM-regulated cassette exons in the mesenchymal state.

METHODS

Cell Lines and EMT induction

HMLE/Twist-ER cells were grown in Mammary Epithelial Cell Growth Medium (Lonza, USA). To induce EMT in HMLE/Twist-ER cells, final concentration of 20nM tamoxifen (TAM) was added to its culture medium, by splitting cells every other day, until a complete mesenchymal morphology was obtained.

Plasmids and shRNAs

hnRNPM shRNA (shM2) was cloned into the pLKO.1 vector corresponding to the targeting sequence GGAATGGAAGGCATAGGATTT. Another hnRNPM shRNA (shM4) was purchased from Open Biosystems targeting the sequence GCATAGGATTTGGAATAAA. Nonspecific shRNA used as a control was cloned into the pLKO.1 vector targeting the following sequence GCCCGAATTAGCTGGACACTCAA.

High-throughput RT-PCR analysis

Primers for 37 hnRNPM-regulated splicing events were designed to validate the hnRNPM cell-state specific differential alternative splicing analysis. Primers are listed in Table 2. RT-PCR was conducted using RNA was extracted from cells using the E.Z.N.A. Total RNA Kit (Omega Bio-Tek). RNA concentration was measured using a Nanodrop 2000 (Thermo Fisher Scientific). cDNA was generated via reverse transcription using the GoScript Reverse Transcription System (Promega) with 1 µL GoScript RT and 250 ng of RNA in a total volume of 20 µL followed by incubation at 25°C for 5 min, 42°C for 30 min,

and 70°C for 15 min. Semi-quantitative RT-PCR assaying for splicing products was performed using Hot StarTaq DNA polymerase (Qiagen), and PCR cycles were run for 30 or fewer cycles. 10 μL of RT-PCR reactions were analyzed and quantified on a QIAxcel Advanced System.

Table 2: Primers used for hnRNPM cell-state RNA sequencing validation

SPLICING EVENT	FORWARD PRIMER	REVERSE PRIMER
SE:chr11:129993506-129993674:129992199-129992408:129996594-129996725:+:APLP2	ATGAGGAGAATCCTACTGAAC	TCATCTGCAGAGGTCTCGAA
SE:chr11:72403797-72403830:72399499-72399582:72404369-72404515:-:ARAP1	GGCTTCCACGATCGCTACTT	TGGGTGGCCTGAGTTTCTTC
SE:chr11:85701292-85701442:85694908-85695016:85707868-85707972:-:PICALM	CAGCAGCCAACTTTTCACCC	TTAAGGCCAGCTGAAGGGTG
SE:chr12:111891480-111891649:111890017-111890644:111893832-111894060:-:ATXN2	GTTTCCCAGCAGCACAACAG	AGCAGTAGAAGGGAGGAGGG
SE:chr12:132479411-132479501:132476717-132476775:132489621-132489729:+:EP400	GCTAGCATATCTTTGACTGATGACG	CTGAGAACTCGGCAGGAAGG
SE:chr12:22611417-22611519:22609904-22610095:22612425-22612476:-:C2CD5	ACACCTGGAGAGTGCAAGTT	TTGCACAAGGTGGAGCAGAA
SE:chr12:27829996-27830029:27829360-27829532:27832421-27832572:+:PPFIBP1	CCATTTGGGACCCTTCCTCC	GAATCTGGAGATGGAGGCGG
SE:chr14:102664091-102664184:102661274-102661457:102674939-102676471:+:WDR20	CACCTAACAGCCACAGCAGA	AGCTCTCTCCCTGCTTCAGA
SE:chr14:105707601-105707751:105695156-105695250:105718843-105718916:-:BRF1	TCTTGGCAAGAGAGCTCTGC	AGCGTGGACTCACACACTTT
SE:chr14:20849711-20849845:20849471-20849560:20850071-20850222:-:TEP1	AGAGCTGGGAAGAAGCAGTG	AACTTCGGAAGGTGCCTGAG
SE:chr14:76203908-76203950:76201537-76201632:76211437-76211551:+:TTLL5	TGAGTCTTCCAGGCGAAACC	CAACTTCCCTGGCCCTTTCT
SE:chr16:88037900-88038017:88017665-88017865:88039695-88039871:+:BANP	CCCAGACGTGCAACAAAGTG	AGTTCAGGGTGATGAGCGTG
SE:chr17:30536368-30536464:30535125-30535328:30538134-30538257:+:RHOT1	CTCCACCACAAGCCTTCACT	GCAGGTGCACATACAGCGTA
SE:chr17:61803997-61804055:61791365-61791468:61805717-61805773:-:STRADA	ACTGAAGTAGGCCTACACAGT	AGTGAGCAGCTCGTAACACC
SE:chr19:42862938-42863106:42862295-42862459:42863249-42863394:+:MEGF8	TACTCCCTGCACTGTCCTGA	CCACCATGGTGTCCCCTAAC
SE:chr1:17961042-17961057:17958815-17958961:17961329-17961511:+:ARHGEF10L	TGTGTGAGACGTTGACGGAG	CTTGTCATAGGTGCCACCGT
SE:chr1:53372190-53372283:53370705-53370762:53373539-53373626:-:ECHDC2	TGAAGGAGCGGGAACAGATG	CTCGCGTGGTCTCAATCAGT
SE:chr22:19462590-19462623:19459209-19459331:19466605-19466695:-:UFD1L	CGGGGTTTCTTCGTTGCATT	ATGCCCTCATCAGCCACAAA
SE:chr22:45723722-45723944:45719055-45719308:45726483-45726612:+:FAM118A	GGTTGTCGCCCATGATCTGA	AGGACGCCGTACTTCATGTG
SE:chr2:27997290-27997397:27995540-27995559:28002299-28002564:+:MRPL33	GCCAAGAGCAAGTCAAA	AAGTCATTTTCAATCCACCG
SE:chr3:15046063-15046120:15045382-15045493:15055095-15055296:+:NR2C2	TCTCCACCGACTCTGCTGTA	GACGAGGTTGTCTGAGGTGG
SE:chr3:9475728-9475865:9475528-9475634:9475960-9476169:+SETD5	TATGGGACCACTCAGAGGCA	GGCCATTCAGGTCAGAACGA
SE:chr4:148984298-148984451:148968042-148968202:148985566-148985658:+:ARHGAP10	CGTGACTACAGCTGTCCCTG	GCTGTGTTCTGCTTCACACG
SE:chr4:175219116-175219213:175204827-175205094:175220222-175220361:+:CEP44	GCTCCCAGCTTTGAAGGTCT	ACACAGTCCACCTCTTCAGG

SE:chr4:1940177-1940259:1936870-1936989:1952798-1952930:+:WHSC1	GGACAAGCACAGTCTTCGGA	AGTCACTCCTCGCTCAGACT
SE:chr4:83782783-83782861:83778841-83778917:83784470-83784545:-SEC31A	ACTTTGAGGATGATTCTCGTGGA	TCCAGATGAGGGTAGAAATTCAGA
SE:chr5:102523014-102523077:102522020-102522140:102526542-102526683:+:PPIP5K2	AAGCGCACCTAACCTACAGG	GTCTTCCGTGGAACGTCAGA
SE:chr5:175786483-175786570:175782573-175782752:175786813-175786921:-:KIAA1191	CAGCTGGACCATGTTCCTCA	CAGGGTCCTCGAGGGTATCA
SE:chr5:37157414-37157522:37153841-37154095:37157771-37157970:-:C5orf42	GCAGTTCCATCGTCTGCAGA	TAGTTGGTGAGCTGCAGAGC
SE:chr5:64951462-64951480:64948327-64948372:64954194-64954328:+:TRAPPC13	CTCAAACCATTGGATGTGAAAACC	GCTCCAGTGAAACCTTCTCCA
SE:chr5:74695134-74695212:74685416-74685512:74696029-74696122:-:COL4A3BP	ATTGGCCTACATCCTTGCCC	GCATCTCCGCCTACATCCTG
SE:chr7:6438292-6438349:6431554-6431672:6439756-6439819:+:RAC1	AAACCGGTGAATCTGGGCTT	TGATGCAGGACTCACAAGGG
SE:chr7:91506191-91506292:91502635-91504078:91509368-91509427:-:MTERF1	GGGATGCAGAGCCTTTCCTT	CTCCAGGCTGTCGTTTCCTT
SE:chr8:130916744-130916831:130915557-130915596:130951853-130952000:-:FAM49B	CAGGTGTTGAGGGGCTCC	CAGTGCTGGTGATTCTGTCCT
SE:chr8:142154246-142154358:142148120-142148236:142160932-142161052:+:DENND3	GCGGGAATAGGACCTATGGC	TTCTCAGGCCTGAAGCACAG
SE:chr8:73942570-73942630:73939174-73939287:73944276-73944368:+:TERF1	GCAGCGGCAAAAGTAGTAGA	AGGTCTTGTTGCTGGGTTCC
SE:chrX:13781863-13781974:13780462-13780563:13785245-13785403:+:OFD1	TCAGATGTGGACAAGCTAGCT	TCTTTCACTTGCTGTTCCCGT

Antibodies for Western Blot

Antibodies used for western blotting were as follows: hnRNPM (Origene), E-cadherin and plakoglobin (Cell Signaling Technology), Fibronectin and N-cadherin (BD), vimentin (Thermo), GAPDH (Millipore).

Deep RNA sequencing and data analysis

Two biological replicates for control and hnRNPM knockdown HMLE/Twist-ER cells with and without TAM treatment were collected with 1ml TRIzol for a 10cm dish. RNAs were extracted followed by the TRIzol Reagent kit from Invitrogen. The purified RNAs were submitted to the Genomic Facility at University of Chicago for RNA quality validation, RNA-seq library generation and paired-end sequencing on HiSeq 4000. RNA-seq reads were aligned to the human genome (GRCh37, primary assembly) and

transcriptome (Gencode version 24 backmap 37 comprehensive gene annotation) using STAR version 2.6.1a [32] with the following non-standard parameters --outFilterMultimapNmax 1 --outSAMstrandField intronMotif --outFilterType BySJout --alignSJoverhangMin 8 --alignSJDBoverhangMin 3 --alignEndsType EndToEnd. Only uniquely aligned reads were retained for downstream analysis.

Differential alternative splicing was quantified using rMATS version 4.0.2 [33] using the following non-default parameters --readLength 100 --cstat 0.01 --libType fr-secondstrand. To identify significant differential splicing events, we set up the following cutoffs: FDR < 0.05, ΔPSI ≥ 0.1, and average junction reads per event per replicate ≥ 20. Differential gene expression analysis was performed by counting reads over genes from the same annotation as alighment using featureCounts version 1.5.0 with the following non-default parameters -s 2 -p -C -B. Differential gene expression analysis was conducted using DESeq2 performed on genes with at least 10 counts present in the library with the lowest sequencing depth [67]. Significantly regulated genes were defined as genes with an | log2FC | > 2 and FDR < 0.05. Regulon overlap analysis between differential alternative splicing events and differentially expressed genes was assessed with Pearson correlation and hypergeometric testing.

Gene set enrichment analysis (GSEA) was conducted using the GSEA pre-rank method where differentially expressed genes were ranked by log2FC before conducting GSEA analysis using gene set level permutation 10000 times [36, 37].

iCLIP assay and analysis

Two biological replicates of HMLE/Twist-ER untreated (Epitheilal) and tamoxifen treated (Mesenchymal) cells were UV crosslinked and iCLIP was performed identically to a previously published protocol using an anti-hnRNPM antibody (Origene) [68].

Data analysis was conducted using the Clip Tool Kit (CTK) v1.0.3 [69]. After processing and mapping iCLIP reads in CTK, replicates were pooled and single nucleotide resolution binding sites for hnRNPM were determined using the following strategy: 1) Statistically significant iCLIP peaks were called and delimited by one half the peak height (adjusted $p < 0.05$). 2) Significant crosslinking induced truncations sites (CITS) were called (adjusted $p < 0.05$. 3) using the CTK CITS.pl script. 3) CITS were further filtered by retaining only those that overlapped with a significant peak. 4) CITS were extended upstream and downstream by 10 nucleotides to identify local binding sites for further analysis.

De novo motif analysis was conducted using HOMER v4.10 findMotifsGenome.pl script with the following non-default parameters -p 4 -rna -S 10 -len 4,5,6 -size 100 -chopify. De novo motifs were computed compared to a background of shuffled human introns. Metagenes and other analyses were computed using custom R and python scripts.

For the GU-rich hnRNPM binding motif RNA map analysis, 14 GU-rich 5-mer motifs used in a previous study to identify hnRNPM binding motifs were used to screen for motif enrichment [70]. The motifs were = ['TGTGT', 'GTGTG', 'TTGTG', 'GTGTT', 'TGTTG', 'TGTGG', 'GTTGT', 'GGTGT', 'TGGTT', 'TTGGT', 'TGGTG', 'GTGGT', 'GTTGG',

'GGTTG']. The 500 bp of sequence flanking upstream and downstream hnRNPM and ESRP1 regulated cassette exons as well as control exons obtained from the rMATS alternative splicing analysis were also obtained. The motif score was computed in a custom python script by counting the number of nucleotides covered by any one of the GU-rich motifs in a sliding window of 50 bp shifted 1-nucleotide at a time across the 250 bp interval in all of the regulated and control cassette exons. The motif score was set equal to the percent of nucleotides covered by the motifs in each of the sliding windows and plotted for regulated exons, stratified by inclusion or skipping, and control exons.

For the CLIP binding metagenes, hnRNPM binding sites were overlapped with four regions of 500 bp flanking the upstream and downstream 5' and 3' splice sites flanking all hnRNPM-regulated cassette exons in each cell state. Each region was divided into 100 bins of 5 nucleotides each, and the percent occurrence of at least one binding site in each bin was plotted for hnRNPM-inclusion, skipping, and control cassette exons.

Statistical analyses

All data were presented as mean ± standard deviation, unless specifically indicated. Correlation was assessed using pearson correlation. Statistical significance tests included Fisher's exact tests and hypergeometric tests. p-value < 0.05 was considered statistically significant. $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***) where indicated.

RESULTS

hnRNPM regulates partially overlapping sets of genes and splicing events in epithelial and mesenchymal cell states

In order to understand the scope of cell-state specific regulation of EMT by the splicing factor hnRNPM, we utilized the well-established *in vitro* EMT model system HMLE-Twist ER where addition of tamoxifen transits human mammary epithelial cells (HMLE) to a mesenchymal state in 14-28 days (Fig. 1A). The epithelial and mesenchymal cells showed clear loss of epithelial markers such as E-cadherin and gain of mesenchymal markers such as N-cadherin (Fig. 1B). In the epithelial and mesenchymal cell states, hnRNPM was stably depleted using shRNA (Fig. 1C) and polyA selected RNA-seq was completed. hnRNPM expression does not change much during EMT.

The effect of hnRNPM depletion on the transcriptome and post-transcriptional regulation was completed through a bioinformatics pipeline to compute differentially expressed genes (DEGs) and differential alternative splicing (DAS) events dependent on hnRNPM expression. For comparison, EMT-regulated DEGs and DAS events were also determined (Fig. 2A-B). More upregulated DEGs and more skipping DAS events were identified after hnRNPM depletion, while the distribution was more even during EMT, suggesting that hnRNPM suppresses gene expression and suppresses exon inclusion, of which the role of hnRNPM as a driver of exon skipping is consistent with previous findings [65, 71]. Interestingly, hnRNPM-regulated splicing events did not change gene expression much during EMT (Fig. 2C).

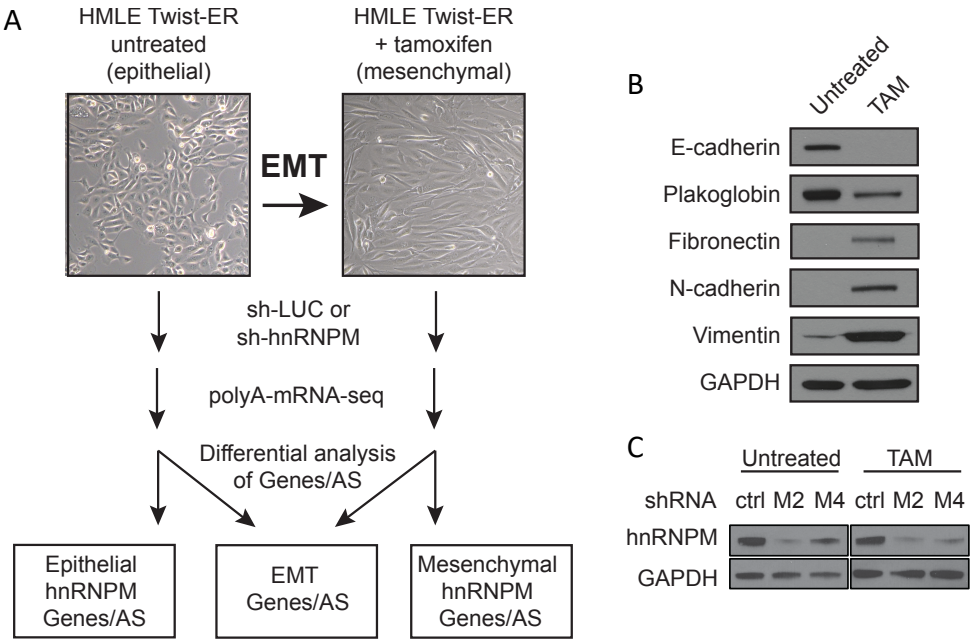

Figure 1: hnRNPM regulates alternative splicing and gene expression in a cell-state specific manner. (A) HMLE Twist-ER cells were treated with tamoxifen to induce EMT *in vitro*. hnRNPM was knocked down in HMLE Twist-ER cells untreated (epithelial cells) and tamoxifen treated (mesenchymal cells) and RNA sequencing were followed. Differential gene expression and differential alternative splicing (AS) analysis was conducted to identify cell-state specific hnRNPM regulation. (B) Western blot showing loss of epithelial markers and gain of mesenchymal markers upon tamoxifen (TAM) induction. (C) Western blot showing hnRNPM knockdown after knockdown with two independent shRNAs. shM2 was the shRNA used for the RNA seq analysis.

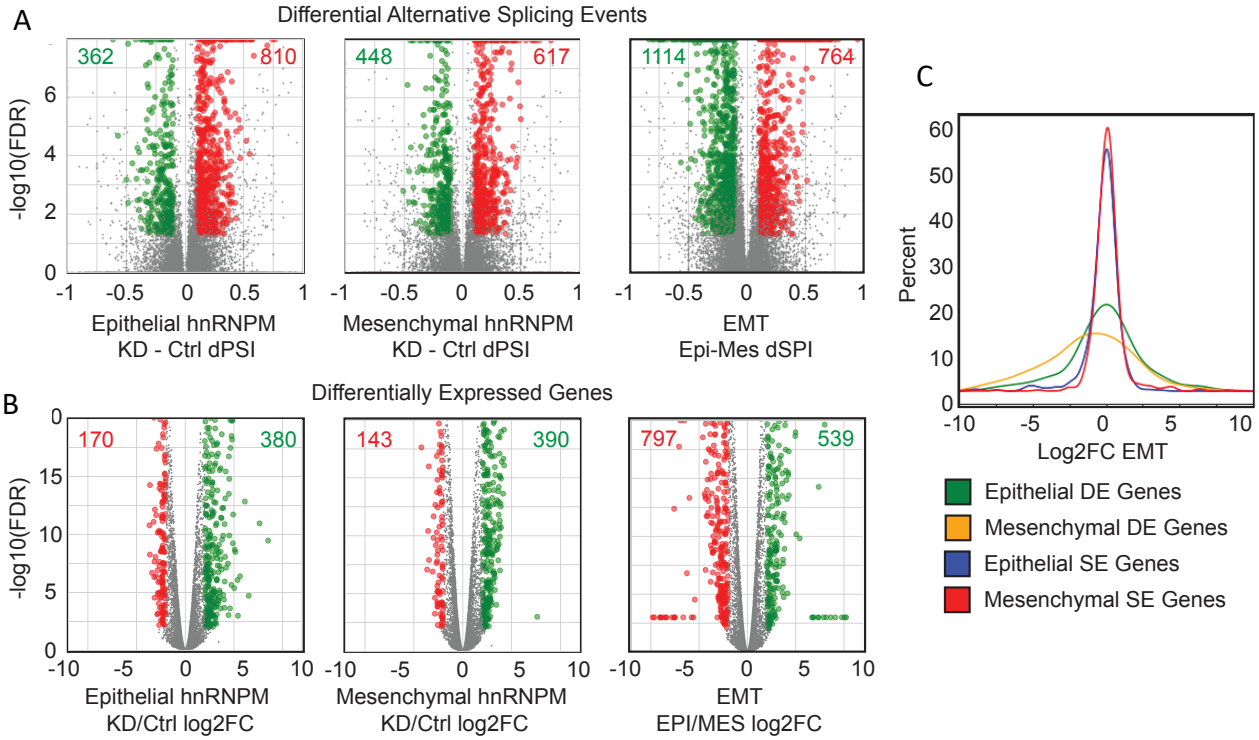

Figure 2: Differential alternative splicing and differential gene expression quantification. (A) Volcano plots of differential alternative splicing events (FDR < 0.05, [dPSI] >= 0.1, Junction Counts >= 20) and (B) differentially expressed genes (DEGs). (FDR <0.01, |log2FC|>2) (B) after hnRNPM KD in epithelial (left), mesenchymal (middle) cells as well as during EMT (right). (C) Kernel Density Estimate plot showing fraction of genes undergoing changes in gene expression during EMT for DEGs regulated upon hnRNPM knockdown in epithelial (green) and mesenchymal (yellow) states as well as differential alternative skipped exon events upon hnRNPM KD in epithelial (blue) and mesenchymal (red) cell states.

The distribution of splicing events regulated by hnRNPM is heavily skewed towards skipped exons, the most common form of alternative splicing (Fig. 3) [5], and this distribution is largely the same between epithelial and mesenchymal cells.

Validation of the alternative splicing results from the RNA-seq analysis was conducted by validating nearly 30 skipped exon events for the hnRNPM epithelial and mesenchymal RNA seq using two independent shRNAs targeting hnRNPM (Fig. 4). Pearson correlation coefficients between the RNA-seq and RT-PCR were above 0.8, with the shM2 RNA, the same used for the RNA-seq, showing a higher correlation than the other shM4 RNA.

hnRNPM cell-state specific DEGs and DAS skipped exons shift towards the EMT regulatory direction in the mesenchymal state

As we previously showed that hnRNPM is required for CD44 exon skipping and thus EMT in a mesenchymal cell-state specific manner [18], we overlapped the DEGs and DAS SE altered upon hnRNPM knockdown with those that change during EMT (Fig. 5). We observed statistically significant overlaps between both epithelial and mesenchymal hnRNPM regulons with EMT, and in both DEGs and DAS SE we observed a shift from EMT-concordant to EMT-discordant changes upon hnRNPM depletion in the epithelial state, supporting our hypothesis that hnRNPM is important in maintaining the epithelial phenotype.

Intrigued by this observation, we performed Gene Set Enrichment Analysis (GSEA) on gene ranked by the magnitude of their changes upon hnRNPM knockdown and

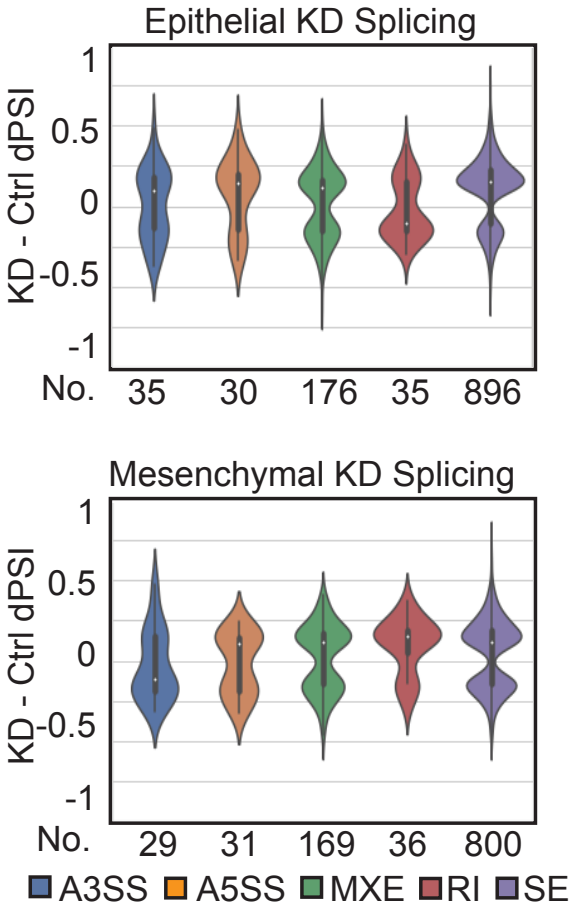

Figure 3: Distribution of hnRNPM cell-state differential splicing events. Violin plots showing distribution and direction of splicing of Alternative 3' Splice Sites (A3SS), Alternative 5' Splice Sites (A5SS), Mutually Exclusive Exons (MXE), Retained Introns (RI), and Skipped Exons (SE) upon hnRNPM knockdown in epithelial (top) and mesenchymal (bottom) cell states.

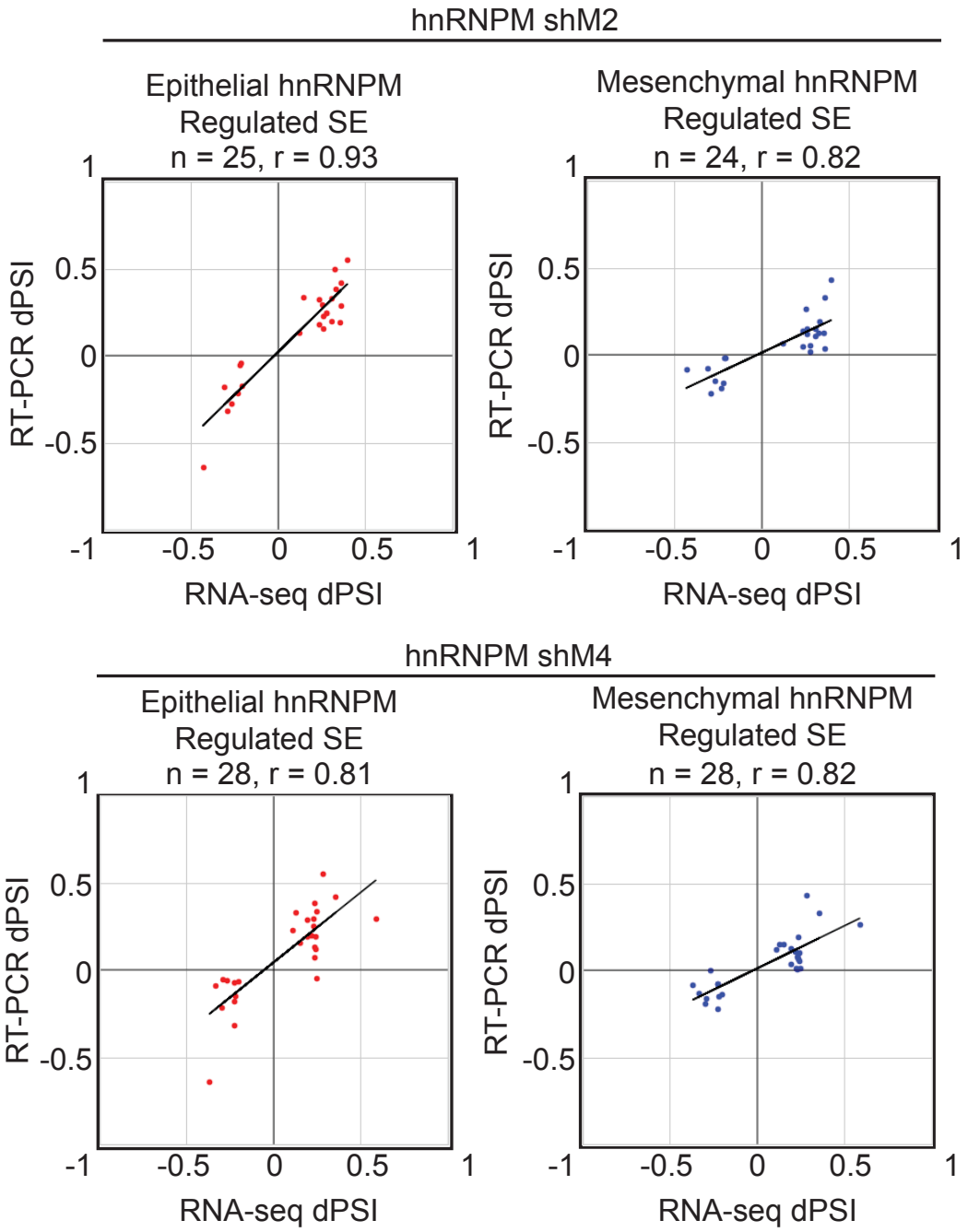

Figure 4. Validation of hnRNPM cell-state splicing events. Scatterplots showing correlation of hnRNPM-regulated alternative skipped exons comparing RNA-seq determined deltaPSI (PSI) (x-axis) and high-throughput RT-PCR delta PSI (y-axis). Two independent shRNAs targeting hnRNPM (shM2 and shM4) were used for the RT-PCR result.

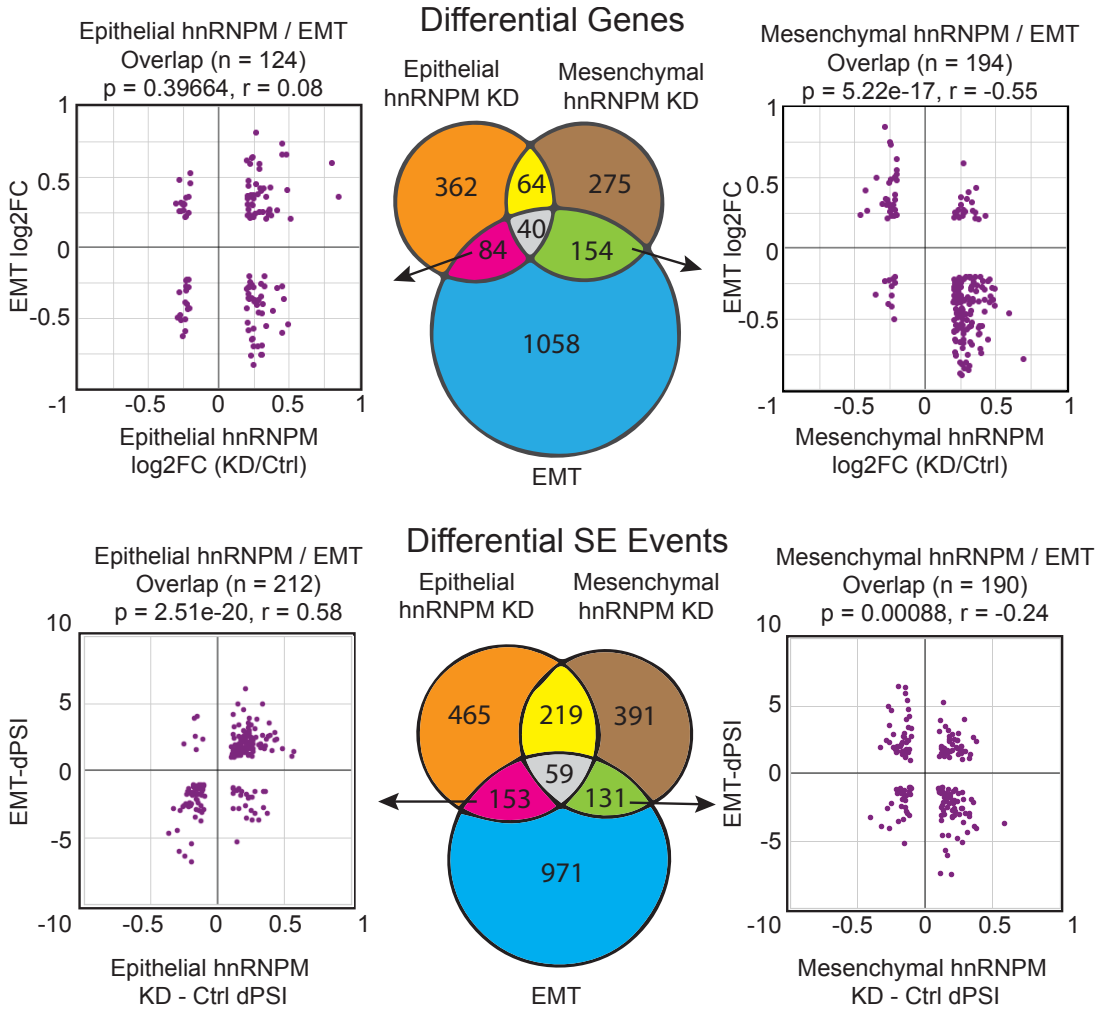

Figure 5: hnRNPM shows cell-state changes in gene regulation and splicing.

(Upper Panel) (Center) Venn diagrams showing significantly DEGs upon epithelial hnRNPM KD, mesenchymal hnRNPM KD, and during EMT. DEGs shared between epithelial hnRNPM cell-state and EMT (left panel) show no significant relationship. DEGs shared between mesenchymal hnRNPM cell-state and EMT (right panel) show a larger negative correlation, suggesting depletion of hnRNPM antagonizes EMT-driven gene expression in mesenchymal cells. (Lower Panel) (Center) Venn diagrams showing significantly differentially regulated SE upon epithelial hnRNPM KD, mesenchymal hnRNPM KD, and during EMT. SE events shared between epithelial hnRNPM cell-state and EMT (left panel) are significant and show positive correlation, while SE events shared between mesenchymal hnRNPM cell-state and EMT (right panel) shift to a negative correlation, showing loss of hnRNPM antagonizes EMT splicing in the mesenchymal cells.

compared enrichment of general cell program hallmarks as well as a curated set of EMT and breast cancer specific genes sets (Figure 6, left panel). We observed some overlap in epithelial and mesenchymal hnRNPM dependent gene signatures, with hnRNPM depletion causing upregulation of interferon and inflammatory responses, which were opposed during EMT. However, we also observed some uniquely enriched gene sets, with cell-cycle associated genes decreasing upon hnRNPM knockdown only in the epithelial state while genes involved in apoptosis response, fatty acid metabolism, and UV response increased upon hnRNPM knockdown only in the mesenchymal state. These results suggest that while there are some overlapping hnRNPM functions in both cell states, hnRNPM has the ability to impinge on other cellular programs in a cell-state specific manner.

As we were specifically interested in cell-state specific differences in hnRNPM's role during EMT, we focused on examining enrichment of previously published gene sets upregulated/downregulated in EMT phenotypes as well as breast cancer subtype associated signatures (Figure 6, right panel). Interestingly, we observed that hnRNPM depletion causes upregulation of genes typically downregulated during EMT, such as E-cadherin, supporting the notion that hnRNPM is required for EMT. However, genes that are upregulated during EMT were largely unaffected by hnRNPM. These results suggest that hnRNPM plays a more important role in facilitating loss of epithelial character during EMT and suppressing these genes in the mesenchymal state as opposed to promoting acquisition of mesenchymal gain-of-function genes.

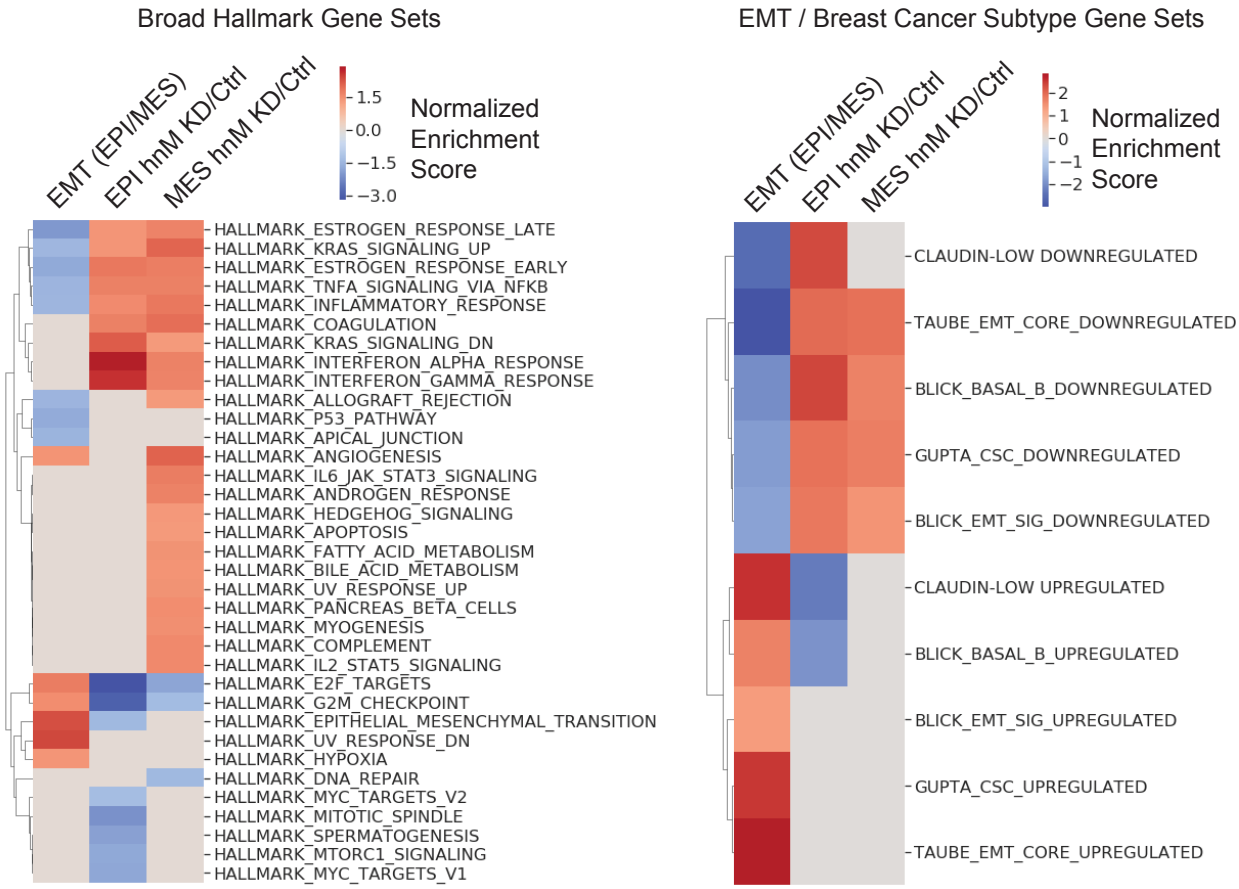

Figure 6: hnRNPM depletion regulates distinct gene sets in different cell states. (Left panel) hnRNPM depletion regulates some similar and some distinct gene sets, with general antagonism to those regulated during EMT. (Right panel) hnRNPM depletion upregulates epithelial associated gene sets that are generally downregulated during EMT, however hnRNPM depletion has no significant effect on gene sets upregulated during EMT. Color bars described normalized enrichment scores calculated during GSEA.

hnRNPM shows similar binding patterns in both epithelial and mesenchymal states

We performed hnRNPM iCLIP in both epithelial and mesenchymal cell states (HMLE Twist-ER Untreated and Tamoxifen treated cells, respectively) to see if global differences in hnRNPM binding across the transcriptome might explain the cell-state specific differences observed in hnRNPM function (Fig. 7A). hnRNPM binding sites in both cell states were highly correlated among replicates. We performed a stringent filtering procedure in the CLIP analysis leveraging the individual nucleotide resolution of the method to identify 54377 and 27764 high-confidence hnRNPM binding sites in the epithelial and mesenchymal state, respectively. Using these binding sites, we examined the proportion of binding sites belonging to different gene regions. hnRNPM primarily binds distal introns in nearly 90% of binding sites, with no major shift in binding distribution occurring between cell states (Fig, 7B-C). De novo motif analysis using the stringent binding sites observed GU-rich motifs that were consistent between the two cell states (Fig 8). Identification of GU-rich motifs validates the stringency of our motif analysis as these motifs are highly similar to those identified for hnRNPM through previous CLIP analysis [71].

hnRNPM shows increased motif enrichment downstream of mesenchymal hnRNPM-regulated skipped exons

By integrating the hnRNPM iCLIP binding sites with the cell-state specific hnRNPM-regulated skipped exons, we mapped hnRNPM motifs and binding sites across

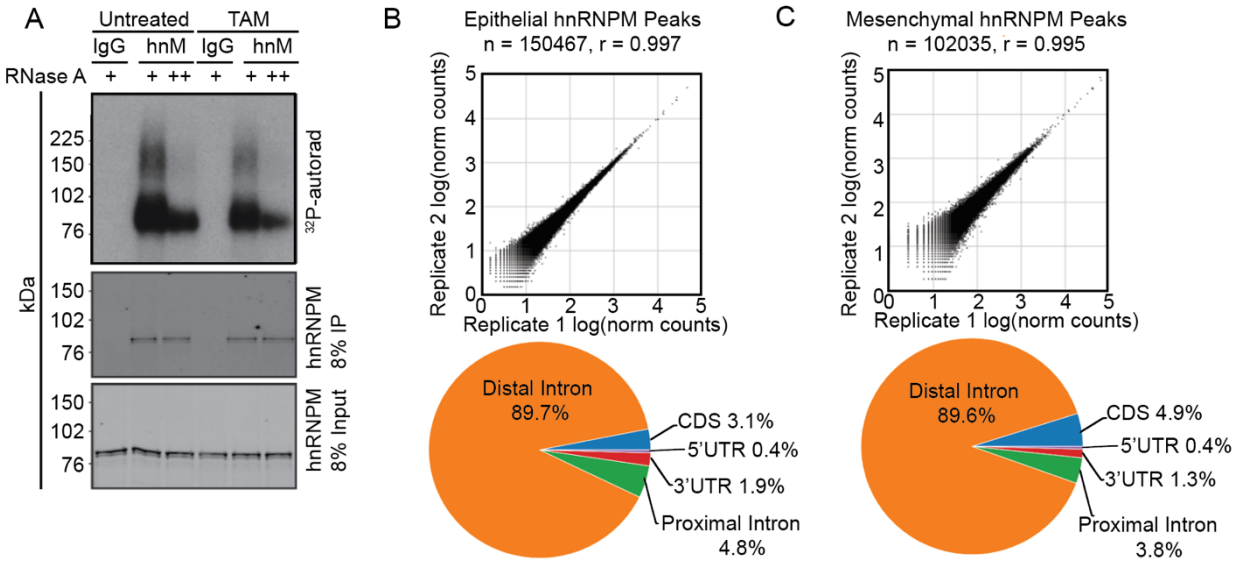

Figure 7: hnRNPM cell-state specific iCLIP reveals similar binding profiles in both cell states. (A) hnRNPM iCLIP in HMLE Twist-ER untreated (epithelial) and tamoxifen-treated (mesenchymal) cells. (Upper panel) Autoradiogram after crosslinking and IP with hnRNPM shows hnRNPM-RNA complexes abolished with excess RNAse A treatment. Blots for hnRNPM in IP (middle panel) and input (bottom panel) showed expected hnRNPM band. *Figure contributed by Dr. Ryan Flynn, Stanford University. (B,C) (Top panels) Significant binding peaks ($p < 0.001$) identified in epithelial (B) and mesenchymal (C) hnRNPM iCLIP show strong correlation with primary genomic distributions (bottom panels) for significant binding sites in distal introns > 500 bp from exonic splice sites, followed by proximal introns < 500 bp from exonic splice sites and then locations on gene bodies.

hnRNPM Epithelial Top CLIP Motifs

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-1209	-2.786e+03	49.13%	33.59%
2		1e-1178	-2.713e+03	80.57%	66.34%
3		1e-1040	-2.396e+03	62.83%	48.07%

hnRNPM Mesenchymal Top CLIP Motifs

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-652	-1.502e+03	60.03%	43.66%
2		1e-583	-1.343e+03	89.67%	77.82%
3		1e-484	-1.116e+03	82.17%	69.81%

Figure 8: hnRNPM de novo motif analysis reveals consistent GU-rich motifs in both cell states. De novo motif analysis using significant hnRNPM binding clusters identifies similar GU-rich motifs in both epithelial and mesenchymal cell states.

the proximal introns within 500 bp of hnRNPM-regulated spliced sites (Figs 9 and 10). hnRNPM motifs were GU-rich pentamers derived from a previous study of hnRNPM [70]. hnRNPM shows strongest enrichment of motifs near hnRNPM-skipping exons in the immediately flanking variable exons, with no discernible enrichment upstream or downstream of the variable exon, consistent with previous reports [71]. However, we noted greater enrichment in the mesenchymal state downstream of the variable exon compared to the epithelial state. This is consistent with our previously published observation that hnRNPM binds more strongly downstream of CD44 variable exon 8 during EMT, and that this mechanism is essential for EMT describes part of hnRNPM's mesenchymal restricted function during EMT [18]. Mapping the hnRNPM binding sites across the same hnRNPM-regulated cassette exons observed similar flaking enrichment of binding sites in proximal introns around the variable exon (Fig. 10), however the iCLIP experimental design used in this study was unable to determine changes in binding intensity that might exist downstream of the variable exon in the mesenchymal state.

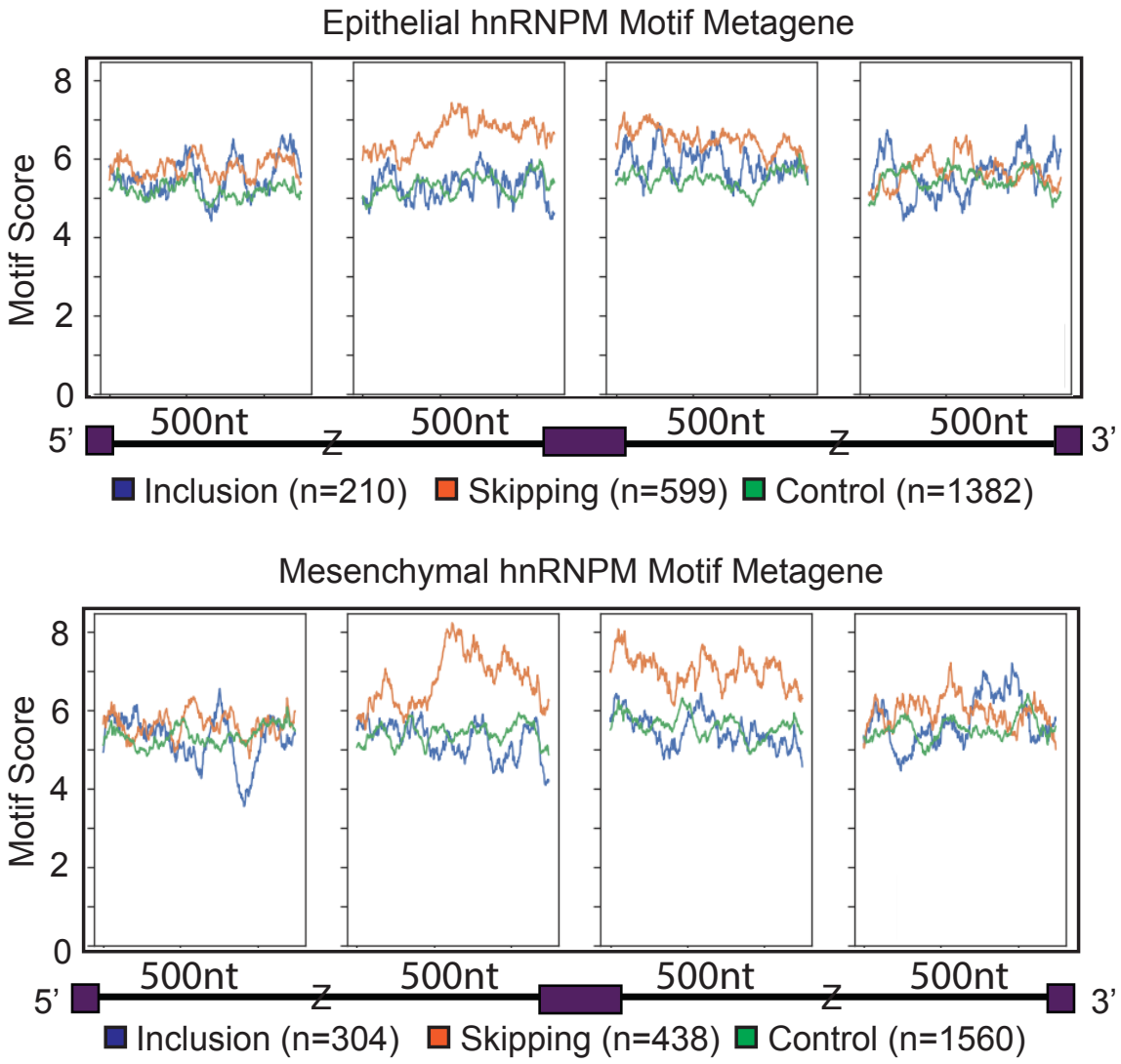

Figure 9: hnRNPM binding motifs are enriched proximal to hnRNPM-mediated skipping exons. RNA motif maps of GU-rich motifs within proximal introns (500 bp within splice sites) as well as 50 bp of flanking exonic sequence show higher enrichment near hnRNPM-driven epithelial (top panel) and mesenchymal (bottom panel) skipping exons compared to inclusion or control cassette exons. This enrichment is most pronounced flanking the central variable exon, with greater enrichment seen downstream of the variable exon in the mesenchymal compared to the epithelial state.

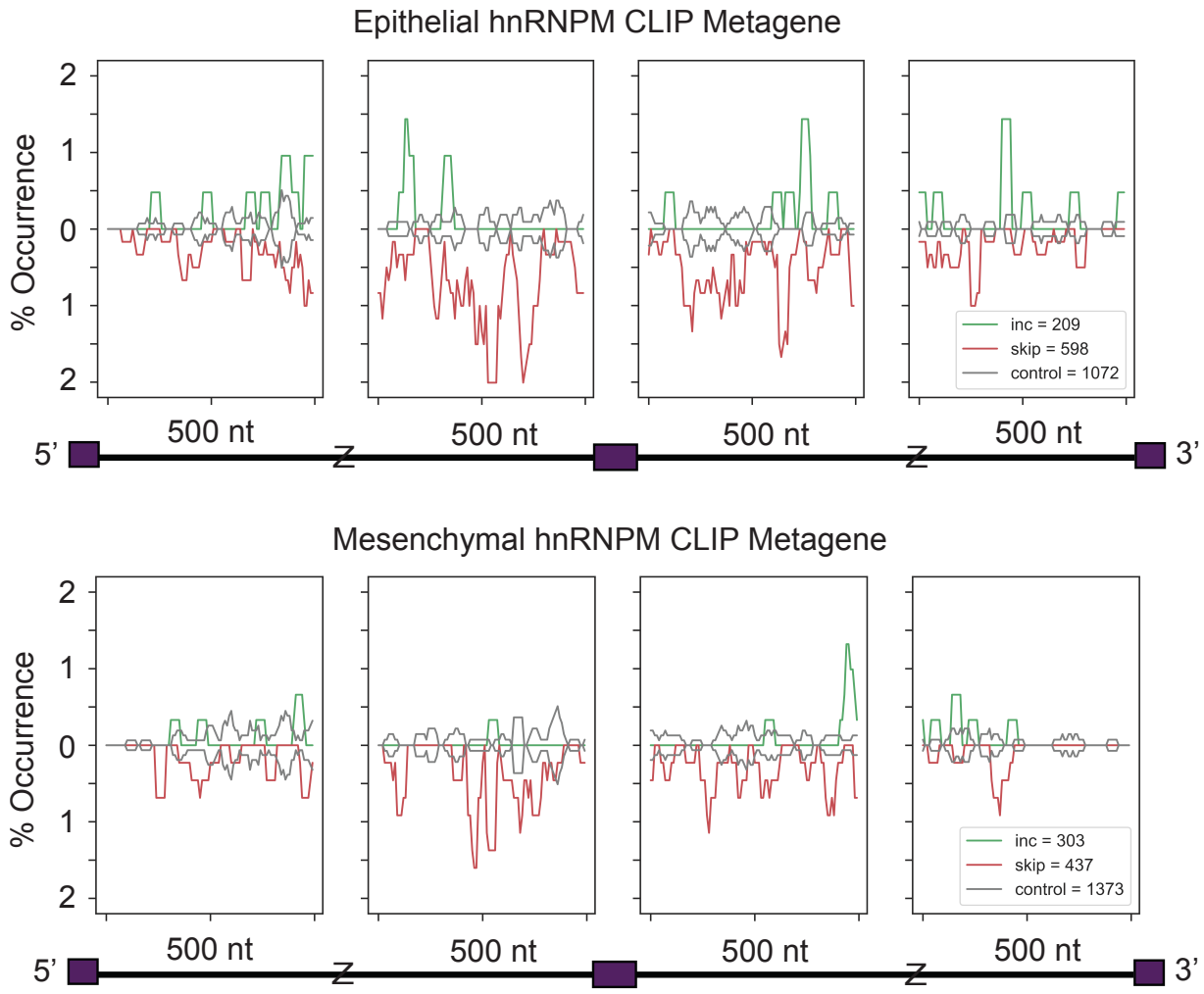

Figure 10: hnRNPM iCLIP sites are enriched proximal to skipping exons. Metagene plotting the percent occurrence of (upper panel) epithelial hnRNPM binding sites near epithelial regulated hnRNPM cassette exons and (lower panel) mesenchymal hnRNPM binding sites near epithelial regulated hnRNPM cassette exons. hnRNPM binding is strongest near skipping events immediately proximal to the variable exon in both cell states.

DISCUSSION

In this study we sought to understand how hnRNPM, a non-tissue-specific RNA binding protein, serves as a necessary regulator for EMT. hnRNPM, as a member of the heterogeneous nuclear ribonucleoprotein family, has been traditionally assumed to function as a member of the splicing repressive complex, a group of proteins that generally suppress splice site recognition [72]. Indeed, as our and others' studies have identified, the majority of hnRNPM regulated splicing events show hnRNPM driving exon skipping [28]. However, recent investigations into members of the hnRNP family and other RNA binding proteins using transcriptome-wide methods have shown that the functional role of splicing factors in cell phenotypes is far from binary; for example, a study investigating the RNA binding protein RBM47, a close homolog of hnRNPR, implicated gain of RBM47 as suppressive of breast cancer metastasis and another study showed that depletion of RBM47 accelerated EMT [9, 49]. However, a closer investigation of RBM47 in splicing regulation during EMT showed that nearly 40% of the splicing events regulated by RBM47 are regulated concordantly during EMT. This raises the interesting question as to how RNA binding proteins seem to play phenotypically necessary roles in cell phenotypes through gain or loss of function studies while maintaining regulatory input into many RNA processing events, including alternative splicing, that are more nuanced and less binary. Is it possible that only a key set of splicing events are regulated by the RNA protein that contribute to the observed phenotype, or is the effect of the RNA binding protein more distributed? In our study, hnRNPM shows a shift in regulation of both genes and splicing events towards the EMT direction of regulation when comparing the regulons

in the epithelial and mesenchymal state as demonstrated by the shift of correlation from negative to positive in Figure 5. This shift is not total but is consistent in both levels of gene regulation.

One approach to answer this question has been to integrate other orthogonal sources of information, such as localizing RNA binding sites or motifs, to better understand which regulatory targets of a particular RNA binding protein may be directly regulated whereas others may be indirect effects. Several recent studies have combined RNA-sequencing based alternative splicing analyses upon splicing factor perturbation with individual nucleotide crosslinking and immunoprecipitation (CLIP) techniques to obtain a core set of splicing events directly regulated by the RNA binding protein of interest [42, 63, 71]. While CLIP technologies are still limited in sensitivity and specificity, integrating this information has proven invaluable in improving the reliability of inferring bona-fide RNA binding protein regulatory targets across the transcriptome. Generally, no more than 50% of the splicing events regulated by an RNA binding protein contain an exon-proximal binding site, suggesting that either current CLIP methods are not sensitive enough, or more likely, many of the splicing events altered upon RNA binding protein perturbation represent indirect effects. In our particular study, only around 15 – 25% of splicing regulatory targets of hnRNPM contain a splice site proximal binding site.

One of the largest manifestations of this effort to integrate RNA binding protein RNA sequencing and crosslinking information across more than one cell line is the ENCODE project, and recent efforts to harmonize the RNA binding protein depletion datasets with the crosslinking datasets for each RNA binding protein are just now being

completed and reviewed [73, 74]. Small scale efforts focusing on single proteins in multiple cell lines are also rare, with one notable example showing that NOVA binding targets differ between whole spinal cord and motor neurons in mice, with motor neuron targets enriching for splicing events in cytoskeletal targets [66]. In our study, we observed similar binding topology of hnRNPM binding both upstream and downstream of hnRNPM-regulated cassette exons in both cell state. However, we noticed an increase in hnRNPM RNA binding motifs downstream of cassette exons in the mesenchymal cell state (Figure 9), an observation that is consistent with our previously published result that hnRNPM acquires the ability to bind downstream of CD44 exon 8 to promote variable exon skipping exclusively in the mesenchymal state [18]. The mechanisms of this observation remain unknown, although our previous observations that splicing factor competition between hnRNPM and the epithelial specific RNA binding protein ESRP1 and the observation that hnRNPM functions more actively in dynamic splicing factor complexes are interesting areas for future study [18, 65, 70].

CHAPTER 3: hnRNPF and the small molecule emetine regulate alternative splicing during EMT through association with RNA G-quadruplexes

*This work was completed in collaboration with Huilin Huang and Jing Zhang, adapted from the following published manuscripts:

Zhang, J., **Harvey, S.E.**, & Cheng, C. (2019). A high-throughput screen identifies small molecule modulators of alternative splicing by targeting RNA G-quadruplexes. Nucleic acids research. PMID: 30698802

Huang H.*, Zhang J.*, **Harvey, S.E.***, Hu X., Cheng C. (2017). RNA G-quadruplex secondary structure promotes alternative splicing via the RNA-binding protein hnRNPF. Genes Dev. PMID: 29269483 (*These authors contributed equally to this work)

ABSTRACT

It is generally thought that splicing factors regulate alternative splicing through binding to RNA consensus sequences. In addition to these linear motifs, RNA secondary structure is emerging as an important layer in splicing regulation. Analysis of RNA-binding protein footprints revealed that G quadruplexes are enriched in heterogeneous nuclear ribonucleo-protein F (hnRNPF)-binding sites and near hnRNPF-regulated alternatively spliced exons in the human transcriptome. Mining breast cancer TCGA (The Cancer Genome Atlas) data sets, we demonstrate that hnRNPF negatively correlates with an EMT gene signature and positively correlates with patient survival. These data suggest a critical role for RNA G-quadruplexes in regulating alternative splicing. Small molecules have the potential to modulate RNA G-quadruplex integrity. We previously characterized emetine as a small molecule that disrupts RNA G-quadruplexes. Transcriptome analysis reveals that emetine globally regulates alternative splicing, including splicing of variable exons that contain splice site-proximal G-quadruplexes. Our data suggest the use of emetine for investigating mechanisms of G-quadruplex-associated alternative splicing.

INTRODUCTION

Nearly all human genes are estimated to undergo alternative splicing, providing the complexity and diverse functions of human proteomes [4, 5, 75-78]. This conserved post-transcriptional process must be precisely regulated, and accumulating evidence has indicated that aberrant alternative splicing causes human diseases, including cancer [6, 18, 79]. The code for regulating alternative splicing resides in the pre-mRNA. The majority of the code identified thus far is comprised of short nucleotide motifs located near exonic splice sites [72, 80]. These consensus sequences recruit RNA binding proteins to splice sites to promote or inhibit spliceosome assembly. In contrast to these linear consensus motifs, there is a void in our understanding of how RNA secondary structures contribute to the splicing code.

One type of RNA secondary structures is the RNA G-quadruplex. RNA G-quadruplexes are assembled through the interactions between four guanines that are organized in a cyclic hoogsteen hydrogen-bonding arrangement [81-84]. G-quadruplexes functioning as cis-elements to regulate splicing have been reported in a growing number of genes, such as fragile X mental retardation 1 (FMR1) and the tumor suppressor TP53 [85, 86]. Recent profiling of RNA G-quadruplexes has revealed widespread and evolutionarily conserved G-quadruplex structures in the human transcriptome [87]. Given the prevalence of RNA G-quadruplex structures in the transcriptome, it is essential to define the functions of these G-quadruplexes in normal and disease states as well as the factors that recognize them in order to execute splicing regulation. Several RNA binding proteins have been shown to interact with RNA G-quadruplexes that impact biological

consequences [88-94]. However, the relationship between splicing factor recognition of G-quadruplexes and regulation of alternative splicing remains largely unclear.

In exploring the effect of alternative splicing in cellular functions, we and others showed that alternative splicing of the cell adhesion molecule CD44 is dynamically regulated during epithelial-mesenchymal transition (EMT), a developmental program that is abnormally activated during tumor metastasis [10, 16, 18, 19, 95]. Switching of splice isoforms from CD44 variants (CD44v) to CD44 standard (CD44s) is necessary for cells to transit from an epithelial cellular state to a mesenchymal phenotype. A cis-acting RNA element designed “I-8” is located in the intron immediately downstream of the CD44 variable exon 8 (v8) and was identified to modulate the splicing of CD44 [18].

In this study, we investigated the role RNA G-quadruplexes may play in the binding and splicing regulation mediated by a group of RNA binding proteins of the heterogeneous ribonucleoprotein family. We show that hnRNPF both binds and associates with G-quadruplex containing alternative splicing across the transcriptome more than other splicing factors of the same family. Depletion of hnRNPF promotes an EMT-associated gene signature. These results connect hnRNPF and RNA G-quadruplexes to EMT, a critical process that drives tumor metastasis.

Although much effort has been directed toward understanding the importance of RNA G-quadruplexes in biology, few small molecules capable of targeting these structures to modulate biological functions have been discovered or developed. Some progress has been made in the discovery of small molecules designed to selectively

target splicing factors and alternative splicing events [96-98]. The identification of small molecules targeting RNA secondary structures that are associated with disease-relevant alternative splicing, such as G-quadruplexes, has been very limited. We previously found that the analogous small molecules emetine and cephaeline regulate alternative splicing by interfering G-quadruplex structures [99]. We also showed that G-quadruplex integrity, specifically a quadruplex present in the intron of CD44, is important for maintaining CD44 variable exon inclusion and sustaining the epithelial cell state [100]. Here we show that treatment with emetine promotes loss of an epithelial gene signature and emetine-regulated alternative splicing is associated with G-quadruplexes across the transcriptome. Therefore, emetine has the potential to regulate cellular processes and alternative splicing by targeting G-quadruplexes.

METHODS

Cell culture and treatment

The maintenance of immortalized human mammary epithelial cells HMLE, HMLE-Twist, and MCF10A were described previously [10, 18]. Immortalized human mammary epithelial (HMLE) cells and HMLE derivative cell lines were grown as described in a 1:1 mix of MEGM (Lonza) and DMEM/F12 (supplemented with 5 µg/ml insulin, 10 ng/ml EGF, 0.5 µg/ml hydrocortisone, and 1X Pen/Strep). To passage, cells were incubated with 0.15 percent trypsin at 37°C for 5-10 minutes. To inactivate trypsin, CS-DMEM was then added to the cells. Following centrifugation at 1,000 rpm for 5 minutes, trypsin-containing medium was aspirated, and cells were resuspended in MEGM and replated. HMLE cell lines were passaged every two to three days at a density of 1:2 to 1:4. It was frequently beneficial to change media (one-half to full volume) on HMLE cells on days that cells were not passaged. MCF10A cells were grown in DMEM/F12 supplemented with 5% horse serum, 10 µg/mL insulin, 20 ng/mL EGF, 100 ng/mL cholera toxin (Sigma), 0.5 µg/mL hydrocortisone, and 1X Pen/Strep. To passage, cells were trypsinized with 0.05% trypsin for about 20 minutes at 37°C. Once cells were detached, they were rinsed off the plate with DMEM supplemented with 5% calf serum to inactivate trypsin. Cells were then centrifuged, trypsin was removed, and cells were resuspended in MCF10A growth media and replated.

Plasmids and shRNAs

The plasmids expressing human hnRNPF shRNAs (shF) or non-specific control shRNA (shNS) were generated by insertion of target sequences (shF: GAGTGATGTTATCTAAGTTTA; shNS: GCCCGAATTAGCTGGACACTCAA) flanked by human miR-30 sequences into the XhoI and EcoRI sites of the TRC lentiviral shRNA vector (Open Biosystems).

Western blot and immunofluorescence

Cells were lysed in RIPA buffer and analyzed by western blotting as previously published [101]. Antibodies used for western blotting were anti-hnRNPF (Santa Cruz) and GAPDH (GE) used as a loading control.

Predicted G-quadruplexes analysis from CLIP-seq data

High throughput CLIP-seq data and exon array data of hnRNPs were retrieved from the Gene Expression Omnibus (GEO) and the annotation of RNA binding motifs were provided by Dr. Gene Yeo. Pairwise correlation of hnRNP RNA binding motifs was conducted after merging overlapping hnRNP CLIP-sites. hnRNP-binding sites +/- 25 nucleotide flanking sequences were analyzed by QGRS Mapper (http://bioinformatics.ramapo.edu/QGRS/index.php) with parameters: Max length = 30nt, Min G-group = 2, loop size = 0-36, to identify potential G-quadruplexes (PGQs). PQGs in randomly shuffled sequences were analyzed as a background control. A Fisher's exact test was used to evaluate the statistical significance of enrichment of PGQs for each set

of hnRNP-binding sites compared to the shuffled background control. Central enrichment of PGQs in hnRNP CLIP-sites was analyzed by calculating the distance from the 5' end of the PGQ to the center of each corresponding CLIP-site. PGQs/CLIP or PGQs/kb were calculated by dividing the numbers of PGQs by the total number or length of hnRNP-binding sites, while the enrichment fold change was calculated by normalizing PGQ/motif of each hnRNP to that of the corresponding shuffled background sequences.

hnRNPF Depletion RNA sequencing, alternative splicing, G-quadruplex enrichment analysis, and Gene Set Enrichment analysis

RNA was extracted from HMLE cells stably expressing shNS or shF from two biological replicates using Trizol and poly-A selected RNA-seq libraries were generated using TruSeq Stranded mRNA Library Prep Kits (Illumina) and subjected to 100 bp PE stranded RNA sequencing on an Illumina Hiseq 4000. RNA-seq reads were aligned to the human genome (GRCh37, primary assembly) and transcriptome (Gencode v24 backmap 37 comprehensive gene annotation) using STAR v2.5.3a [32] using the following parameters: STAR --runThreadN 16 --alignEndsType EndToEnd --quantMode GeneCounts --outSAMtype BAM SortedByCoordinate. Differential gene expression was quantified using Cuffdiff v2.2.1 [102] using the following non-default parameters: --max-bundle-frags 1000000000 --library-type fr-firststrand: Differential alternative splicing was quantified using rMATS v3.2.5 [33] using the following non-default parameters: -t paired -len 100 -analysis U -libType fr-firststrand. Significant differentially spliced cassette exons were quantified using only unique junction reads and the following cutoffs: FDR < 0.05, delta

PSI $\geq$ 0.2, average junction reads per cassette event per replicate $\geq$ 20. Control cassette exons were identified by the following filters: FDR > 0.5, minimum PSI for shNS or shF < 0.85, maximum PSI > 0.15, average junction reads per cassette event per replicate $\geq$ 20. These filters were selected to identify cassette exons with evidence of alternative splicing but were not differentially spliced upon hnRNPF depletion. G-quadruplex enrichment analysis was conducted by curating a strand-specific list of all sequences in the GRCh37 version of the human genome matching the regular expression GxNy, where N is any nucleotides other than G, x $\geq$ 3 and 1 $\leq$ y $\leq$ 7 to identify all sequences with G-tracts of 3 or more separated by 1-7 of any other base. Bedtools v2.26.0 and custom bash and python scripts were used to intersect hnRNPF-dependent alternative splicing events and the 250 or 150 nucleotides flanking each splice site with these G-quadruplexes. Twelve highly expressed cassette exons near PGQs were selected for semiquantitative PCR validation per the following criteria: all exon junctions present in Gencode gene annotation, fpkm for associated gene $\geq$ 10, junction reads $\geq$ 70 per cassette exon event per sample, and junction reads $\geq$ 10 per inclusion or skipping isoform. Processed TCGA BRCA Level 3 data for gene and exon expression was downloaded from the Genomic Data Commons Legacy Archive. Gene Set Enrichment Analysis was performed using the Broad Institute javaGSEA desktop application. GSEA using hnRNPF depletion RNA seq was conducted with all genes with a minimum expression of FPKM $\geq$ 1 in control or hnRNPF knockdown datasets. GSEA using TCGA BRCA gene and exon expression data included all genes in all BRCA samples for which exon expression data were available ranked by pearson correlation coefficient with hnRNPF expression.

Emetine Treatment RNA sequencing analysis and G-quadruplex prediction

RNA was extracted using Trizol from HMLE-Twist and MCF10A cells that were treated with DMSO or emetine (10 μM) for 24 hours. Two biological replicates were performed per experimental condition. Poly-A selected RNA-seq libraries were generated using TruSeq Stranded mRNA Library Prep Kits (Illumina) and subjected to 100 bp PE stranded RNA sequencing on an Illumina Hiseq 4000. RNA-seq reads were aligned to the human genome (GRCh37, primary assembly) and transcriptome (Gencode v24 backmap 37 comprehensive gene annotation) using STAR v2.5.3a [32] with alignEndsType set to "EndToEnd". Differential alternative splicing was quantified using rMATS v3.2.5 [33] using the same transcriptome. Significant differentially spliced cassette exons after emetine treatment were called using only unique junction reads and the following cutoffs: FDR < 0.05, delta PSI > 0.1, average junction reads per cassette event > 20. Predicted G-quadruplexes were curated from a strand-specific list of all sequences in the GRCh37 version of the human genome matching the regular expression GxNy, where N equals any series of nucleotides other than G, x > 2 or 3 and 1 < y < 7 to identify all sequences with G-tracts of 2-3 or more separated by 1-7 of any other base. Bedtools v2.25.0 and custom bash and python scripts were used to intersect emetine-dependent cassette exons, the flanking variable exons, and the 250 nucleotides flanking each splice site with these G-quadruplexes. Emetine RNA sequencing data has been deposited under accession number GSE113505.

Statistical analyses

All data were presented as mean ± standard deviation, unless specifically indicated. Correlation was assessed by pearson correlation. Statistical significance tests included Fisher's exact tests, hypergeometric tests, and log-rank tests. p-value < 0.05 was considered statistically significant. $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***) where indicated.

RESULTS

The splicing factor hnRNPF is a potential G-quadruplex binding protein

Having observed that G-quadruplexes promote alternative RNA splicing, we aimed to identify RNA binding proteins that mediate this regulation. We predicted the ability of RNA binding proteins to bind to G-quadruplexes by analyzing previously published CLIP-seq data [28] for several heterogeneous nuclear ribonucleoproteins (hnRNPs), an abundant family of proteins that regulate alternative splicing. This study was the largest comparative analysis of CLIP-seq data available at the time, including six hnRNP CLIP-seq datasets. We merged overlapping CLIP sites between each dataset and computed pairwise correlations with hierarchical clustering (Fig. 1). The hnRNPH dataset was poorly correlated with and distinct from the other hnRNP datasets. This might be caused by a batch effect because the hnRNPH dataset was not generated by the study authors and was analyzed from a previous publication [103]. We excluded the hnRNPH dataset from further analysis.

We used Quadruplex forming G-rich Sequences (QGRS) Mapper [104] to score the strength of predicted G-quadruplex (PGQ) structures within the CLIP-seq hnRNP-binding regions. We examined the regions that comprised the defined hnRNP-binding sites and 25-nt of flanking sequence both upstream and downstream of the binding sites. Bioinformatics analysis showed that PGQs were most significantly enriched in the binding regions of hnRNPF relative to random control sequences, which exhibited the most significant enrichment out of the five hnRNPs analyzed (Fig. 2A). There were 3304 PGQ-containing binding sites within 13616 hnRNPF-binding regions, comprising 24% of total

hnRNPF binding sites (p = 1.18E-67, Fisher's exact test), compared to the second highest hnRNPA1 with 450, representing about 22% of 2043 hnRNPA1 binding sites (p = 1.20E-13, Fisher's exact test). As a more stringent analysis, we utilized the G-score provided by QGRS mapper to obtain higher confidence G-quadruplexes. The G-score is a metric to predict how likely a PGQ is to form a stable quadruplex structure and prioritizes shorter loop sequences between each G-tract as well as loop sequences that are relatively equal in length [104]. To determine a relevant G-score threshold unlikely to occur by chance, we quantified the number of G-quadruplexes per G-score occurring in 10000 randomly shuffled sequences. G-quadruplexes with G-scores 19 or greater had a p-value less than 0.05. Thus we chose G-scores ≥ 19 as an appropriate threshold to identify significant G-quadruplexes in the hnRNP-binding sites, labeled PGQ-hi (Fig. 2A). This analysis revealed that hnRNPF had the most significant enrichment of PGQ-hi, compared to shuffled control, among all hnRNPs considered (p = 1.53E-81, Fisher's exact test) (Fig. 2B). These data show that PGQs, especially those with a high probability of forming RNA G-quadruplex structures, are significantly enriched within hnRNPF-binding regions.

hnRNPF and hnRNPA1 had the first and second highest percentage of CLIP sites containing PGQs, respectively. By quantifying the position of the start positions of the 5' end of each PGQ relative to the center of the CLIP site, we observed a stronger central

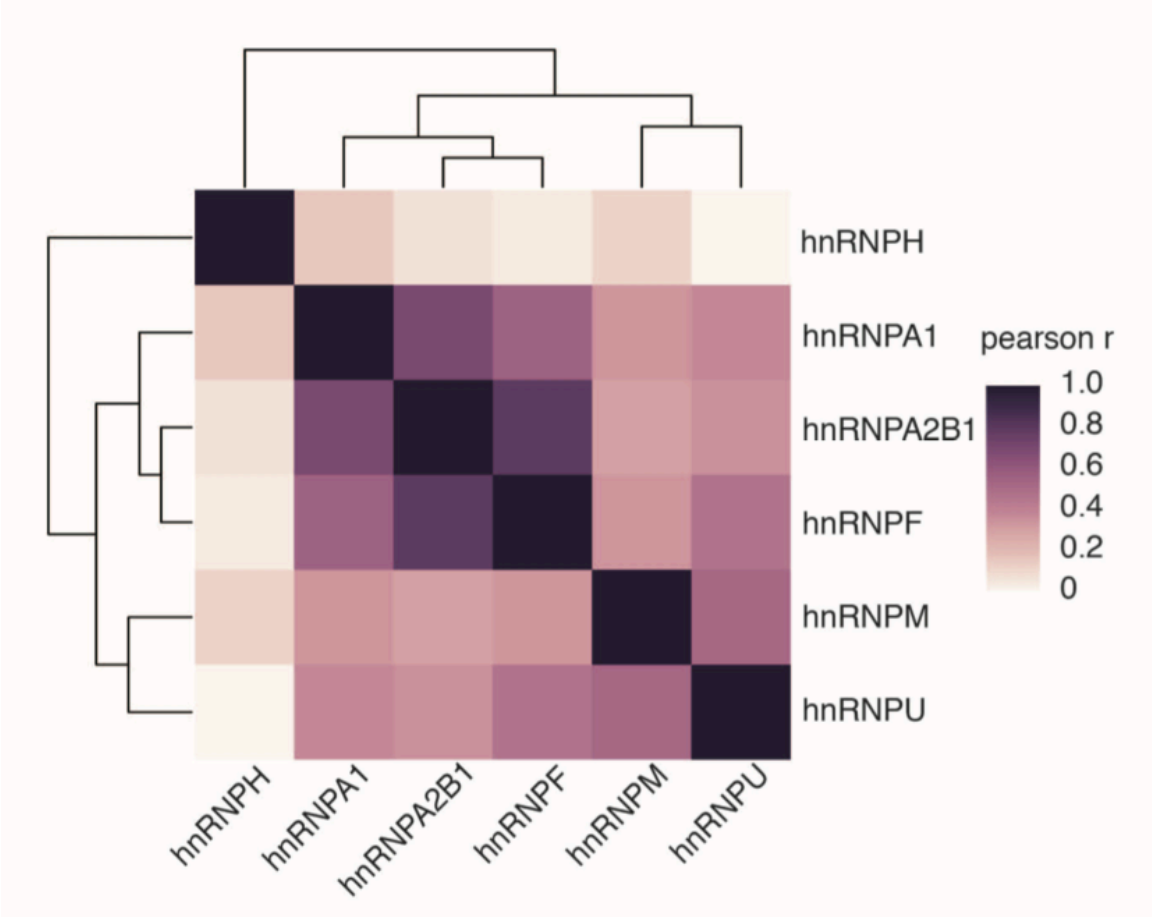

Figure 1. hnRNPH dataset is not well correlated with other hnRNPs. Pairwise correlation and hierarchical clustering of hnRNP CLIP-seq datasets from Huelga et al., 2012. hnRNPH shows weakest correlation with other hnRNPs.

enrichment of PGQs in hnRNPF bound CLIP sites compared to those of hnRNPA1 (Fig. 2C). The percentage of PGQs in hnRNPF CLIP-sites is highest near the center of hnRNPF CLIP-sites and decreases further from the center, showing that G-quadruplexes are quite proximal to the hnRNPF binding site. By contrast, PGQs in hnRNPA1 CLIP-sites are more dispersed and shifted upstream of the center of the CLIP-sites. The increased central enrichment of PGQs within hnRNPF binding sites suggests that hnRNPF binding is directly associated with the presence of a PGQ.

QGRS mapper by default identifies one G-quadruplex per hnRNP binding site. Because PGQs that contain more than four G-tracts have the potential to adopt distinct conformations by utilizing different G-tracts and thus form multiple overlapping G-quadruplexes, we computed all possible G-quadruplexes by requiring no more than four G-tracts within each PGQ. Among all tested hnRNPs, hnRNPF-binding sites showed the most remarkable enrichment of overlapping PGQs (Fig. 2D), supporting its greater potential to form G-quadruplexes. Taken together, the enrichment of PGQs, especially those with high G-scores and overlapping G-quadruplexes, within close vicinity of hnRNPF-binding sites suggests that hnRNPF preferentially binds to G-quadruplex-containing RNA regions compared to other hnRNPs. Examples of hnRNPF CLIP sites [28] containing PGQs that are located in introns proximal to hnRNPF-regulated alternatively spliced exons were shown in Fig. 3.

G-quadruplexes are enriched near alternative exons regulated by hnRNPF

To examine whether hnRNPF regulates G-quadruplex-mediated alternative splicing globally across the transcriptome, we performed RNA-sequencing analysis in

control and hnRNPF-silenced HMLE cells (Fig. 4A,B). hnRNPF knockdown caused significant differential splicing of 190 cassette exons (FDR < 0.05, delta PSI ≥ 0.2). Among them, hnRNPF promoted inclusion of 136 (71.6%) exons and skipping of 54 (28.4%) exons (Fig. 4C), consistent with a previous exon array analysis showing that hnRNPF predominately promotes exon inclusion [28].

To test whether hnRNPF mediates alternative splicing of exons containing proximal G-quadruplexes, we quantified enrichment of G3N7 PGQ near hnRNPF-dependent cassette exons. We found that within 250 nucleotides upstream or downstream of the alternative exon, 43 out of 190 (22.6%) of hnRNPF-dependent exons contain a PGQ, compared to 303/3471 (8.7%) of control cassette exons not differentially spliced upon hnRNPF depletion, representing a statistically significant enrichment (p = 2.36E-8, Fisher's exact test) (Fig. 4D middle bars). Of these 43 hnRNPF-regulated exons containing PGQs, hnRNPF promoted inclusion of 37/43 (86.0%), consistent with the major role of hnRNPF in promoting exon inclusion. The PGQ enrichment near the cassette exon was more significant than the enrichment near the upstream or downstream exons flanking the cassette exon (Fig. 4D). Two cassette exons with PGQs in introns proximal to the cassette exon that were validated with semi-quantitative PCR are shown in in Fig. 4E. These observations suggest that primarily G-quadruplex sequences located proximal to the 3' and 5' splice sites of the cassette exon are targets for hnRNPF binding and splicing regulation. 40/43 (93%) of the cassette exons contained PGQs exclusively in flanking intronic sequence, with the remaining 3 events containing PGQs overlapping or entirely within the cassette exon itself, consistent with previously

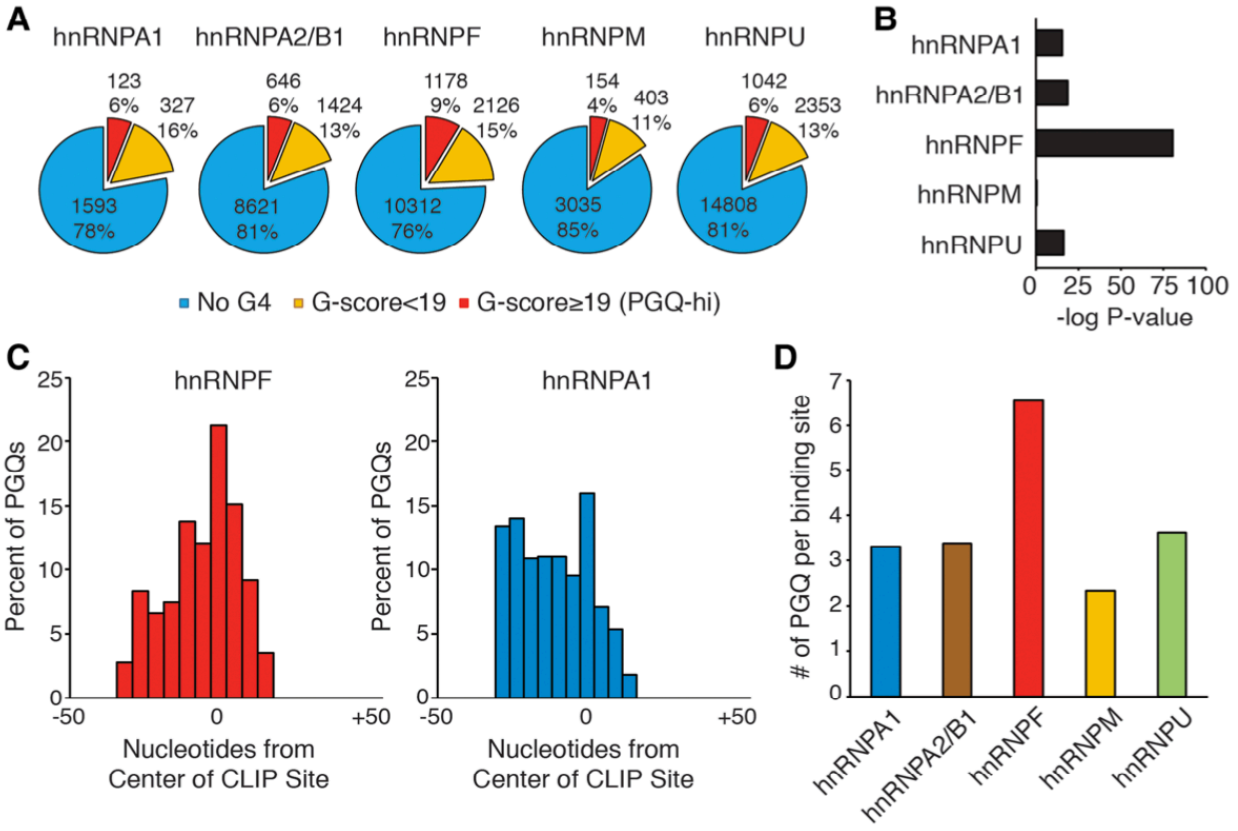

Figure 2. Transcriptome-wide enrichment of predicted G-quadruplexes in hnRNPF binding regions. (A) Percentage of QGRS predicted G-quadruplexes (PGQs) in hnRNP CLIP binding sites and 25-nt upstream and downstream flanking sequences. The fraction of CLIP sites containing PGQs with G-scores ≥ 19 (PGQ-hi) are colored in red while sites containing PGQ with G-score < 19 or sites lacking PGQs are colored in yellow and blue, respectively. (B) –logP values for Fisher exact tests comparing enrichment of PGQ-hi in hnRNP CLIP binding sites compared to shuffled control sequences. (C) Histograms displaying the percentage and location of the start of PGQs relative to centers of hnRNPF (left panel) and hnRNPA1 (right panel) CLIP-sites. PGQs occurring within +/- 25 nt of the CLIP site are shown, but histograms are plotted +/- 50 nt of the center of the CLIP site. (D) Average number of overlapping PGQs/binding site within the hnRNP-CLIP binding sites +/- 25 nt.

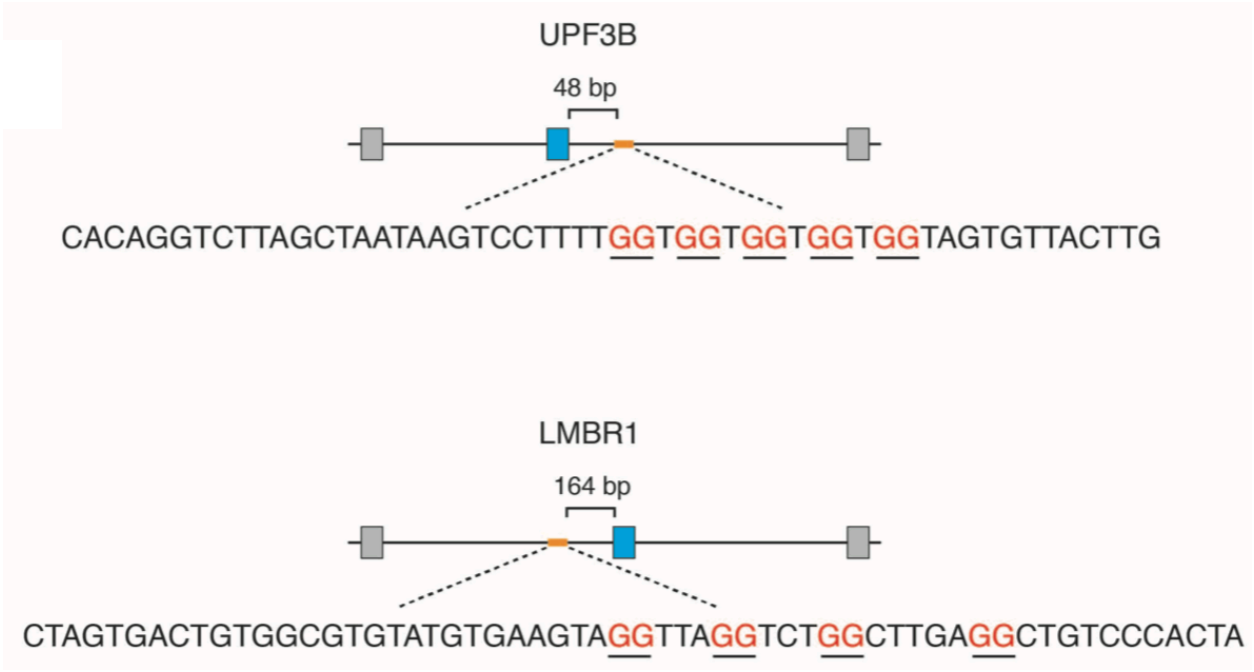

Figure 3: Examples of hnRNPM CLIP sites near hnRNPF regulated exons. (Bottom) Examples are shown for hnRNPF CLIP-seq binding sites that contain predicted G-quadruplexes from mining hnRNPF CLIP data and hnRNPF-regulated alternative exons identified in from a published manuscript [71]. Constitutive exons are colored gray and the cassette exons are colored blue. hnRNPF binding regions are depicted in yellow. Zoomed sequences display the full hnRNPF CLIP binding sequence with the G-quadruplex guanines underlined and colored red.

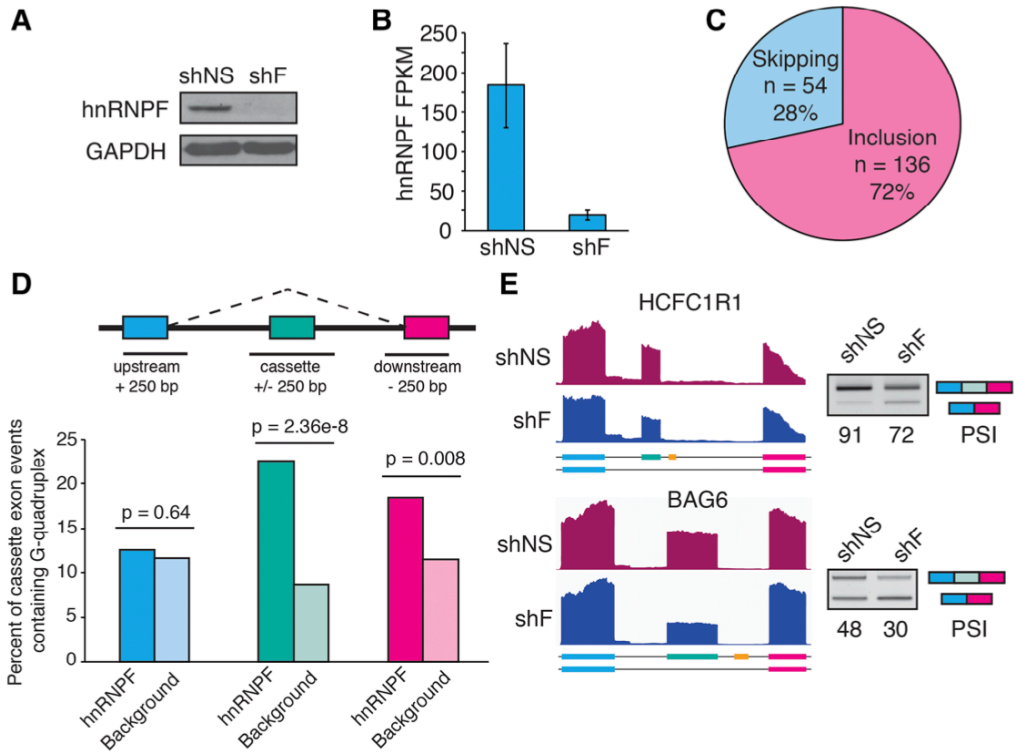

Figure 4. Predicted G-quadruplexes are enriched near hnRNPF-regulated cassette exons. (A) Western blot images showing hnRNPF expression in HMLE cells expressing control (shNS) and hnRNPF-targeting (shF) shRNAs *Completed by Dr. Jing Zhang (B) hnRNPF expression in control and hnRNPF knockdown HMLE cells quantified by RNA-seq. Error bars represent 95% confidence intervals. (C) Proportion of hnRNPF-regulated cassette exon splicing events where hnRNPF promotes exon inclusion (pink) vs. exon skipping (blue). (D) Diagram of cassette exon alternative splicing events and regions scanned for predicted G-quadruplexes (PGQs) within exons and 250 nucleotides proximal to splice sites flanking the exon (top panel). Percentage of hnRNPF-regulated cassette exons (n = 190) containing PGQs upstream, proximal, or downstream of the cassette exon compared to the same regions in non-hnRNPF-regulated cassette exons (n = 3471) (bottom panel). p-value calculated by Fisher's exact test. (E) Tracks of hnRNPF-regulated exons nearby PGQs shown in yellow bars (left panels). RT-PCR images showing increased exon skipping upon knockdown of hnRNPF in HMLE cells (right panels). *RT-PCRs completed by Dr. Jing Zhang.

published results identifying that the majority of hnRNPF binding sites are located in introns [28].

hnRNPF depletion promotes an EMT gene signature

Our next goal was to determine the functional role of hnRNPF. We performed Gene Set Enrichment Analysis (GSEA) of genes differentially expressed upon hnRNPF knockdown. We found that hnRNPF depletion resulted in upregulation of an EMT-associated gene set (Fig. 5A), implying that hnRNPF inhibits EMT. This observation made us curious as to the correlation between hnRNPF expression and CD44 variable exon skipping, which leads to increased production of CD44s, a splicing switch essential for EMT [18]. Evidence supporting the notion that hnRNPF promotes CD44 variable exon inclusion came from analysis of The Cancer Genome Atlas (TCGA) Breast Invasive Carcinoma (BRCA) RNA-seq dataset. We found that hnRNPF expression significantly positively correlates with the levels of CD44 variable exons while negatively correlating with all constitutive CD44 exons (Fig. 5B).

hnRNPF expression correlates with breast cancer patient survival

Abnormal activation of EMT promotes cancer metastasis [16, 95]. To explore the relevance of hnRNPF's function in cancer, we mined the TCGA breast cancer database and performed GSEA on all genes ranked by correlation with hnRNPF gene expression. We found that the hnRNPF gene-set positively correlated with gene signatures that were downregulated during EMT and in basal-like or claudin-low breast cancer subtypes (Fig.

5C). Conversely, hnRNPF gene sets negatively correlated with gene signatures that were upregulated in these phenotypes. These results derived from a large dataset of human breast cancers strongly suggest that hnRNPF antagonizes EMT and invasive breast cancer.

The association of hnRNPF with breast cancer phenotypes associated with patient survival prompted us to examine the relationship between hnRNPF expression in breast cancer and overall patient survival. We performed two-means clustering of the TCGA BRCA samples by hnRNPF expression. Our results showed that patients with hnRNPF highly-expressing tumors exhibited a significant survival advantage compared to patient with hnRNPF lowly-expressing tumors ($p = 0.011$, Log Rank Test) (Fig. 5D). These results strongly support our findings that hnRNPF antagonizes EMT and cancer progression.

G-quadruplexes are enriched near alternative exons regulated by emetine

Our previous publication observed that emetine regulates alternative splicing in a G-quadruplex-dependent manner [99], thus we aimed to identify the effect of emetine on alternative splicing globally across the transcriptome. We performed RNA sequencing (RNAseq) using two human cell lines MCF10A and HMLE-Twist with or without emetine treatment. Emetine treatment caused significant differential splicing of thousands of splicing events in both cell lines, of which nearly 60% represented skipped exon (SE) events (Fig. 6A). We focused on SE events as they were the most frequently regulated by emetine and the most common form of alternative splicing [5]. To identify a stringent set of SE regulated by emetine, we overlapped the SE events in MCF10A and HMLE-

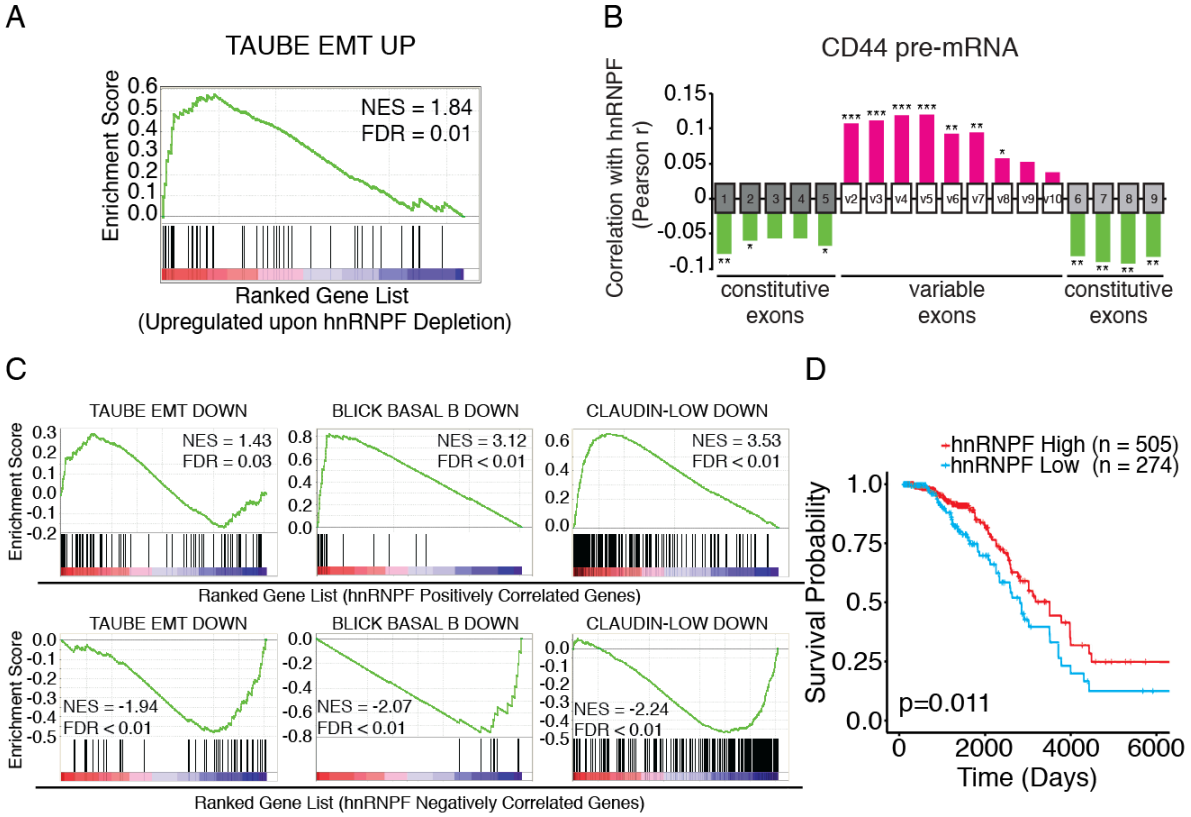

Figure 5. hnRNPF inhibits EMT and positively correlates with breast cancer patient survival. (A) GSEA analysis showing that hnRNPF depletion caused an enrichment in an EMT signature. NES = Normalized Enrichment Score. (B) Correlation of hnRNPF gene expression with CD44 exon expression in TCGA BRCA patient RNA-seq datasets showing strong positive correlation with CD44 variable exons (pink) and weaker negative correlation with CD44 constitutive exons (green). * = $p < 0.05$, ** = $p < 0.01$, * = $p < 0.001$, pearson correlation. (C) GSEA analysis of genes positively (top panel) and negatively (bottom panel) correlated with hnRNPF in TCGA BRCA patient RNA-seq samples revealed hnRNPF positively correlates with genes downregulated during EMT, in basal-subtype breast cancers, and in claudin-low breast cancer while hnRNPF negatively correlates with genes upregulated in these same gene sets. (D) Kaplan-Meier survival curves showing a significant overall survival difference between hnRNPF-High and hnRNPF-Low expressing tumors in TCGA BRCA patients. p-value computed by log-rank test.

Twist and identified 1970 events shared between both cell lines and they were strongly positively correlated (pearson r = 0.89, Fig. 6B). To test whether emetine mediates alternative splicing of cassette exons containing proximal G-quadruplexes, we identified all SE containing potential G-quadruplexes (PGQ) inside the cassette exon, flanking exons, or within 250 nucleotides of a splice site. We found that approximately 60 percent of emetine regulated SE contained a PGQ with 2-nt guanine tracts and approximately 10 percent contained a more stringent PGQ with 3-nt guanine tracts (Fig. 6C). As examples, two of the emetine-regulated cassette exons identified as containing an exon-proximal G-quadruplex, AKR1A1 exon 7 and RFC5 exon 2, were previously identified to contain G-quadruplexes capable of regulating alternative splicing [105]. Experimental validation of these two SEs showed that emetine treatment promoted AKR1A1 exon 7 skipping and RFC5 exon 2 inclusion in both HMLE-Twist and MCF10A cells (Fig. 7A,B). Taken together, these results illustrate that emetine regulates alternative splicing events that are enriched for G-quadruplexes globally across the transcriptome.

We also examined whether cells treated with emetine showed an EMT signature. We performed GSEA analysis using differentially expressed genes from RNAseq data of the emetine treated epithelial MCF10A cells. Noticeably, GSEA analysis showed significant negative enrichment of an EMT-downregulated gene set in response to emetine treatment (Fig. 8), suggesting that emetine favors the loss of an epithelial phenotype [44].

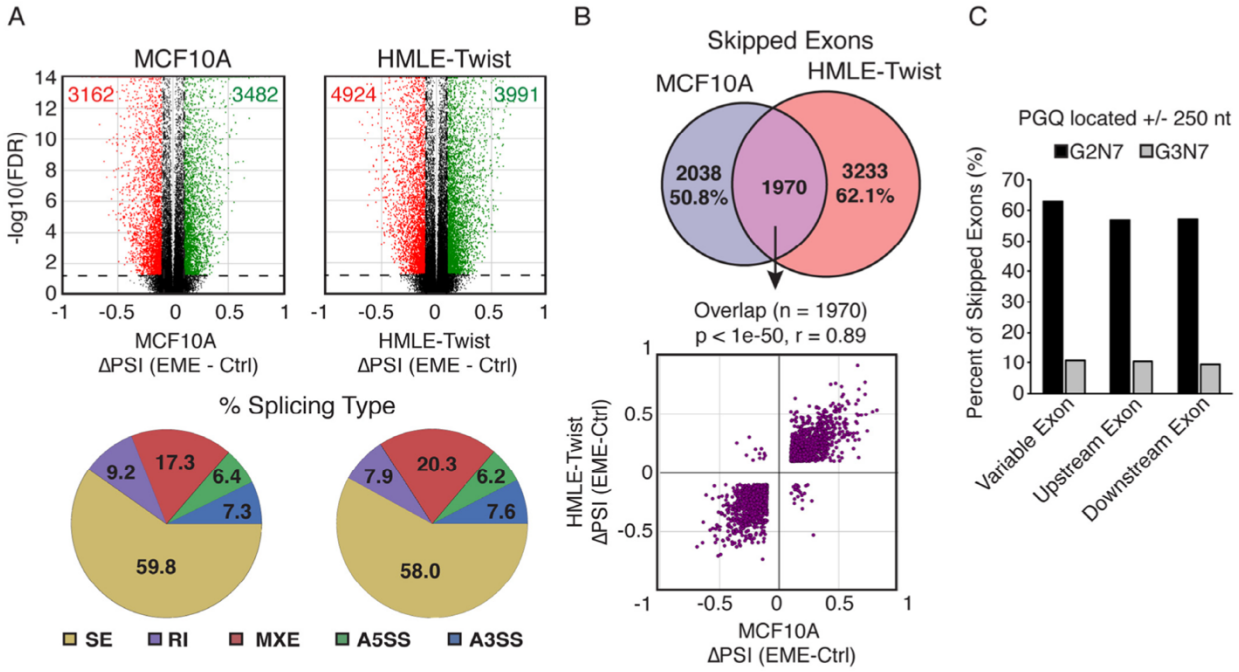

Figure 6. Emetine globally affects G-quadruplex associated alternative splicing. (A) Top Panel: Differential alternative splicing events identified after emetine treatment in MCF10A (Left) and HMLE-Twist (Right) cells. Alternative splicing events that show FDR ≤ 0.05 and $\Delta PSI \geq 0.2$ or ≤ -0.2 were colored as green and red dots, respectively. Bottom panel: Pie charts showing percentage of each splicing type represented in emetine differential splicing events. The majority of events are skipped exon (SE) events. (B) Top panel: Venn diagram showing common skipped exons regulated by emetine in both MCF10A and HMLE-Twist cells. Bottom panel: The common set of cassette exons show highly correlated regulation in both cells types. (C) Bar charts showing the percentage of skipped exon splicing events containing a G2N7 or G3N7 predicted G-quadruplex (PGQ) proximal to splice sites.

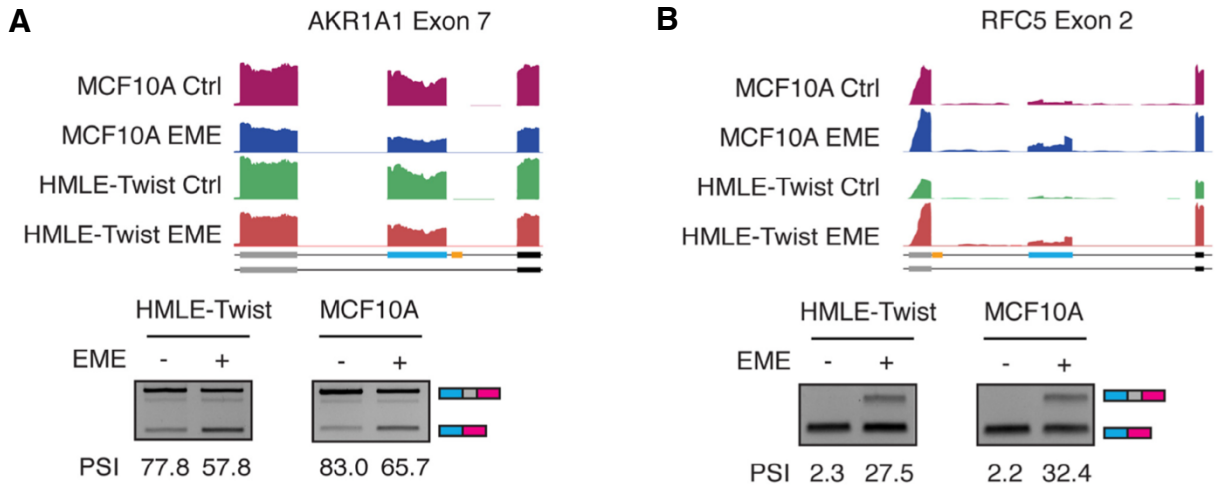

Figure 7. Genome browser tracks and validation of emetine regulated cassette exons. Genome browser tracts of RNA sequencing data and semi-quantitative RT-PCR validation showing emetine treatment inhibits exon inclusion of AKR1A1 Exon 7 (A) and promotes inclusion of RFC5 Exon 2 (B). The G-quadruplexes are depicted in yellow in the schematics of the exon annotation. *RT-PCR completed by Dr. Jing Zhang.

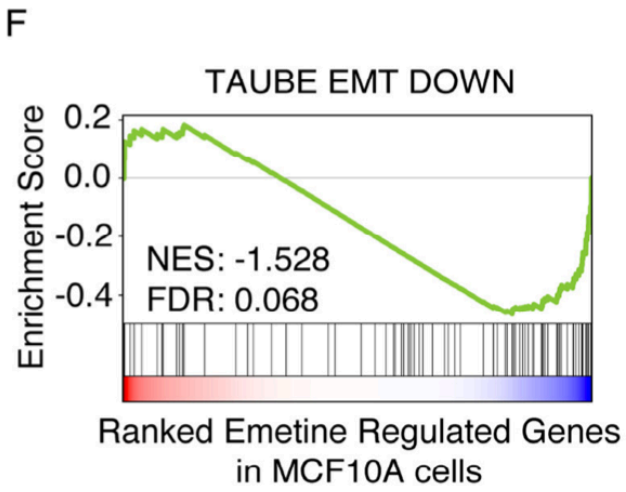

Figure 8. GSEA analysis of emetine regulated genes in epithelial MCF10A cells. Genes downregulated by emetine are also downregulated during EMT.

DISCUSSION

The critical role for alternative RNA splicing in normal development and pathological processes has been increasingly recognized. New technologies mapping the RNA binding sites of RNA binding proteins have greatly improved our understanding of protein-RNA interactions and their biological consequences. In addition to known linear RNA binding motifs, RNA secondary structures have been increasingly found to serve as RNA binding protein motifs in their own right in order to mediate RNA processing. G-quadruplexes are a particular form of RNA secondary structure, and are reported to regulate pre-mRNA processing, RNA turnover, and translation [84, 90, 106, 107]. Thus far, denomination of G-quadruplexes, their associated RNA binding proteins, and their biological functions remain largely unexplored.

G-tracts that consist of three or more consecutive guanines harboring the potential to form G-quadruplex structures are frequent splicing recognition elements enriched near splice sites [108, 109]. A number of splicing factors were found to interact with RNA motifs with G-quadruplex forming capacity, including hnRNPF/H, hnRNPA1, hnRNPA2B1, hnRNPU, and SR proteins [88, 90, 91, 93, 94, 110]. In this study, we found that potential G-quadruplex sequences are over-represented in the vicinity of hnRNPF binding sites compared to other hnRNPs. This observation is congruent with our published experimental results showing that hnRNPF binds to RNA containing G-quadruplexes, including the I-8 G-quadruplex important for CD44 variable exon inclusion, and that hnRNPF stimulates cassette exon inclusion in a manner dependent on G-quadruplexes

in cis [100]. Another member of the hnRNPF/H family, hnRNPH, is highly homologous to hnRNPF and is found to bind both G-quadruplex and G-tract motifs [90, 111].

It is worth noting that the amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) associated GGGGCC expansion of the C9ORF72 gene forms RNA G-quadruplexes. The G-quadruplex structure sequesters hnRNPH, resulting in local depletion of hnRNPH and thus disruption of hnRNPH-dependent splicing events that contain G-tracts [90]. These observations raise an interesting possibility for future investigation that hnRNPF and hnRNPH preferentially bind to G-quadruplex and G-tracts, respectively, and that the interplay between these two homologues may provide a precise mechanism for splicing regulation that concerns the dynamics between G-quadruplex formation and the underlying G-tracts. Another splicing regulator ESRP1 was previously shown to bind to the I-8 GU-rich motif to regulate CD44 alternative splicing. This GU-rich motif overlaps with the G-quadruplex structure. Interestingly however, hnRNPF and ESRP1 both stimulate CD44 variable exon inclusion, but in an independent manner [100]. Considering the dynamic nature of RNA G-quadruplex secondary structure formation, we speculate that the I-8-containing RNA fragment exists as a mixture of its linear form and G-quadruplex configuration. ESRP1 may occupy the linear GU-rich motif while hnRNPF binds to the G-quadruplex secondary structure, both of which promote inclusion of the CD44 variable exon. This possibility illustrates the complexity of alternative splicing regulation by suggesting a redundant mechanism to maintain CD44v splice isoform expression in epithelial cells and warrants future investigation.

hnRNPF has been reported to bind to G-quadruplexes as well as linear G-tracts [112-116]. For example, the G-quadruplexes at the 3' UTR of p53 recruits hnRNPF/H to execute its role in regulating p53 3'-end processing in response to DNA damage [112]. Interestingly however, another report suggested that hnRNPH/F may prefer linear G-rich sequences embedded in potential G-quadruplex motifs to modulate epithelial- and myoblast-specific splicing events [105], but it remains unclear whether hnRNPF preferentially binds to G-quadruplex secondary structure or linear G-tracts. Moreover, the three-dimensional structure of the quasi-RNA-recognition motif (qRRM), one of the three RRM RNA binding domains of hnRNPF, showed that qRRM binds to single-stranded G-tract RNAs to prevent the formation of G-quadruplexes [113, 114]. It will be of great interest to determine whether the full-length hnRNPF with three RRMs maintains interaction with G-tracts or preferentially interacts with G-quadruplexes in a manner different from the isolated qRRM. Our published work showed that hnRNPF exhibited the ability to bind G-quadruplex forming sequences, supporting the notion that hnRNPF is a G-quadruplex binding protein [100]. Regulation of the differential ability for hnRNPF to bind to G-quadruplex secondary structure versus linear G-tracts would be an interesting topic for future investigation. We speculate that the binding of hnRNPF to G-quadruplex and linear G-tracts could be influenced by other RNA binding proteins and that the preferential binding may determine the outcome of splicing products.

Emerging evidence has indicated the importance of G-quadruplexes in regulating fundamental biological activities, including translation regulation [81, 83, 117], alternative splicing [86, 90, 118], and 3' end processing [112]. These findings have stimulated

interest in profiling G-quadruplexes in mammalian cells. Using potassium ions (K+) and the small molecule pyridostatin (PDS), an RNA G-quadruplex specific stabilizing ligand, it was shown that widespread formation of G-quadruplex structures exists in the human transcriptome [87]. Interestingly, when dimethyl sulfate (DMS) was used to modify the N7 position of the guanine residue, which probes for unfolded G-quadruplexes, it was observed that most of the predicted G-quadruplex structures were modified by DMS and thus unfolded [119]. These seemingly contradictory results reveal the dynamic nature of RNA G-quadruplexes in live mammalian cells. G-quadruplex structures may be toggled between the folded and unfolded states. A transiently unfolded region could be accessed by DMS for N7G modification. The presence of this modification indicates that the G-quadruplex structures can be unfolded but does not exclude the existence of the higher-order G-quadruplex structures. Thus, novel approaches to define G-quadruplex structures in live cells and the regulatory mechanisms underlying the dynamics of G-quadruplex structures are of great interest to elucidate the location and function of these RNA secondary structures.

We previously identified emetine as a compound capable of regulating alternative splicing by interfering with the G-quadruplexes through a small molecular screen with a G-quadruplex sensitive splicing reporter [99]. From our RNAseq analysis in emetine treated cells, we found that many emetine-regulated exon skipping events contain G-quadruplexes proximal to the splice sites, suggesting that emetine may have a global role in regulating G-quadruplex-dependent alternative splicing. Our experiments have shown that emetine directly interacts with G-quadruplexes to regulate alternative splicing.

However, it is important to point out that emetine was previously shown to inhibit protein translation by binding to the 40S subunit of the ribosome [120]. In addition to its direct effect on alternative splicing, we cannot fully exclude the possibilities that the alternative splicing changes caused by emetine may be due to inhibiting the expression of splicing regulators [121]. Interestingly, using our RNAseq data, we compared the expression changes of RNA binding protein (RBPs) in response to emetine treatment to investigate emetine's indirect effects through RBPs. Significant differentially expressed genes were identified using the cutoffs as two-fold. None of the RBPs have been previously reported to bind to RNA G-quadruplexes. On the other hand, the expression level of the splicing factor hnRNPF, which affects alternative splicing in a G-quadruplex and emetine dependent manner, does not change in the presence of emetine.

To date, very few studies have evaluated the interaction of small molecules with RNA G-quadruplexes. TMPYP4 has been shown to disrupt G-quadruplex structures when binding to RNA [122, 123], but it has additional effects beyond potential disruption of RNA G-quadruplexes [124, 125]. Our study identifies emetine as a splicing modulatory compound. By disrupting RNA G-quadruplex structures, emetine may affect cellular functions by impacting G-quadruplex-mediated alternative splicing. Emetine can serve as a reagent to help understand the mechanisms underlying how RNA binding proteins regulate alternative splicing through the interaction with G-quadruplexes. It may also be used for the identification of splicing targets that are regulated by G-quadruplex structures. Thus, our studies may inspire researchers to consider the role of RNA secondary

structures in addition to the linear sequences in regulation mechanism of alternative splicing.

CHAPTER 4: AKAP8 antagonizes EMT alternative splicing through differential binding to the transcriptome in a cell-state specific manner

*This work was completed in collaboration with Dr. Xiaohui Hu and Rong Zheng

ABSTRACT

Tumor metastasis is the most lethal attribute of breast cancer, and the epithelial–mesenchymal transition (EMT) plays an important role in this process. Alternative splicing has been shown to causally contribute to EMT; however, the scope of critical splicing events and the regulatory network of splicing factors that govern them remain largely unexplored. Here we report the identification of A-Kinase Anchor Protein (AKAP8) as an EMT splicing regulatory factor that impedes EMT. AKAP8 binds differentially to the pre-mRNA in a cell-state specific manner linked to differential splicing outcomes. Genome-wide analysis revealed that AKAP8-mediated splicing events are oppositely regulated during EMT and that AKAP8 binding is enriched in introns proximal to its splicing regulatory exons. In addition, AKAP8 depletion accelerates an EMT transcriptomic signature, and the AKAP8 downstream target CLSTN1 stratifies breast cancer patients by survival. Thus, this study identifies AKAP8 as a stabilizer of the epithelial cell-state post-transcriptional landscape.

INTRODUCTION

Metastatic breast cancers remain the major challenge in clinical management, causing more than 40,000 breast cancer patients to succumb each year in the United States. One of the key mechanisms that facilitates cancer metastasis is the abnormal activation of a developmental process termed epithelial-mesenchymal transition (EMT) [126-128]. Aberrant activation of EMT enables primary epithelial cancer cells to acquire advantageous mesenchymal properties, including invasion, drug resistance, and blunted anoikis, ultimately allowing the survival of cancer cells within the circulatory system and subsequent colonization of distant organs [16, 129, 130]. Several transcription factors, including Twist, Snail, Slug, and Zeb1/2, and signaling pathways, including TGF-β, Wnt, and Notch, have been characterized as potent regulators that induce cells to undergo EMT [16, 131]. In addition to these important findings, growing evidence has suggested a new mode of action, i.e., alternative RNA splicing acts as a critical layer of regulation impinging on EMT [62, 132].

Alternative RNA splicing is a fundamental mechanism of post-transcriptional gene regulation. With 95% of the human multi-exon genes expressing more than one splice isoform, alternative splicing contributes to the diversity and complexity of human proteome, and thus organ development and tissue identity [6, 133, 134]. The regulation of alternative splicing relies on the precise binding of splicing factors to the RNA consensus motifs that are located in variable exons or their adjacent introns. As such, mutations in either the splicing factors or RNA motifs that perturbs the splicing factor

binding may result in developmental abnormalities and diseases [135]. One of the most dramatic examples of dysregulation of splicing in recent years was the finding that mutations in the SF3B1 splicing factor are found in 81% of cases of refractory anemia with ring sideroblasts and also in a distinct subtype of myelodysplastic syndrome. While the important observations between the splicing machinery and diseases are accumulating, our understanding on the mechanisms and functions of splicing regulation that impinges on diseases is still in its infancy.

The functional connection of alternative splicing to EMT and cancer metastasis was established through the study of the CD44 gene, whose alternative splicing generates two families of proteins, known as CD44v and CD44s. It was shown that epithelial cells that predominantly express CD44v demand an isoform switch to CD44s in order for cells to undergo EMT and for breast cancer cells to metastasis [10, 11, 18, 19, 100, 136-143]. A handful of additional alternative splicing events has subsequently been reported to play a functional role in EMT, suggesting that alternative splicing may serve as a general mechanism in controlling the EMT [144]. The EMT-associated splicing events are controlled by splicing factors and, to a large extent, these splicing factors do not function in isolation and instead act in a combinatorial manner to influence splicing [52, 62]. In the case of CD44 alternative splicing, the heterogeneous nuclear ribonucleoprotein M (hnRNPM) promotes the production of CD44s by binding to CD44 intronic splicing motifs, resulting in an EMT phenotype and an enhanced metastasis ability [18]. The splicing activity of hnRNPM is partially restricted by an epithelial-specific splicing factor ESRP1 through their competition on binding to the same RNA motifs, thus tightly

controlling the switch of CD44 splice isoforms and the transition of cell states during EMT [18, 65]. In addition to this mode of direct competition on binding to RNA substrates, it is conceivable that hnRNPM-interacting splicing factors could also influence hnRNPM's activity and thus its function in promoting EMT [52, 145]. In fact, several splicing factors were found to form a complex with hnRNPM, but the functional consequences in EMT and cancer metastasis have remained unexplored.

In the present study, we report the identification of the A kinase anchoring protein 8 (AKAP8) as an RNA binding protein that inhibits EMT through the regulation of alternative splicing. By integrating AKAP8 depletion RNA sequencing datasets obtained from epithelial and mesenchymal human breast cell lines, we show AKAP8 is required to maintain epithelial specific alternative splicing patterns. Cells with the loss of AKAP8 show an increased EMT transcriptional profile. Through investigation of AKAP8 binding to pre-mRNA in both epithelial and mesenchymal states we show that AKAP8 binds more strongly to pre-mRNA in the epithelial state and directly targets splicing events that accurately predict patient survival. These results shed new lights on the mechanisms of EMT and tumor metastasis that are regulated at the level of alternative RNA splicing.

METHODS

Cell cultures and EMT induction

HMLE/Twist-ER cells were grown in Mammary Epithelial Cell Growth Medium (Lonza, USA). To induce EMT in HMLE/Twist-ER cells, final concentration of 20 nM tamoxifen (TAM) was added to its culture medium, by splitting cells every other day, until the mesenchymal morphology was fully observed.

Plasmids and shRNAs

AKAP8 shRNA was cloned into pLKO.1 vector corresponding to the targeting sequence GCCAAGATCAACCAGCGTTTG. Nonspecific shRNA used as a control targeted the following sequence GCCCGAATTAGCTGGACACTCAA.

RNA sequencing and data analysis

Three biological replicates for control and AKAP8 knockdown HMLE/Twist-ER without TAM treatment and two biological replicates for control and AKAP8 knockdown HMLE/Twist-ER with TAM treatment cells were collected with 1ml TRIzol for a 10 cm dish. RNAs were extracted followed by the TRIzol Reagent kit from Invitrogen. The purified RNAs were submitted to Genomic Facility at University of Chicago for RNA quality validation, polyA selected RNA-seq library generation and paired-end sequencing on a HiSeq 4000. RNA-seq reads were aligned to the human genome (GRCh37, primary assembly) and transcriptome (Gencode version 24 backmap 37 comprehensive gene annotation) using STAR version 2.6.1a [32] with the following non-standard parameters

--outFilterMultimapNmax 1 --outSAMstrandField intronMotif --outFilterType BySJout --alignSJoverhangMin 8 --alignSJDBoverhangMin 3 --alignEndsType EndToEnd. Only uniquely aligned reads were retained for downstream analysis.

Differential alternative splicing was quantified using rMATS version 4.0.2 [33] using the following non-default parameters --readLength 100 --cstat 0.01 --libType fr-secondstrand. To identify significant differential splicing events, we set up the following cutoffs: FDR < 0.05, |ΔPSI| ≥ 0.1, and average junction reads per event per replicate ≥ 20. Differential gene expression analysis was performed by counting reads over genes from the same annotation as alignment using featureCounts version 1.5.0 with the following non-default parameters -s 2 -p -C -B. Differential gene expression analysis was conducted using DESeq2 performed on genes with at least 10 counts present in the library with the lowest sequencing depth [67]. Significantly regulated genes were defined as genes with an | log2FC | > 1 and FDR < 0.05.

Regulon overlap analysis between differential alternative splicing events and differentially expressed genes was assessed with Pearson correlation and hypergeometric testing.

Gene set enrichment analysis (GSEA) was conducted using the GSEA pre-rank method where differentially expressed genes were ranked by log2FC before conducting GSEA analysis using gene set level permutation 10000 times [36, 37].

eCLIP assay and data analysis

AKAP8 single end enhanced crosslinking and immunoprecipitation (seCLIP or eCLIP) was performed in HMLE/Twist-ER cells without or with TAM treated, in two biological replicates, by following exactly the protocol previously published [146]. Specifically, input and AKAP8 antibody IP-ed RNAs were cut from 95 KD to 180 KD on the membrane for RNA isolation and following steps of the protocol through library generation. eCLIP libraries were sent to Genomic Facility at University of Chicago for single-end sequencing. eCLIP data processing was conducted using the public eCLIP pipeline version 0.2.1a (https://github.com/YeoLab/eclip/releases/tag/0.2.1a) and public merge-peaks pipeline version 0.0.6 (https://github.com/YeoLab/merge_peaks/releases /tag/0.0.6), derived from the previously published eCLIP pipeline [146]. High confidence eCLIP peaks for each cell state were called by selecting AKAP8 binding peaks with a minimum log2FC IP / Input signal > 2 among the replicates and an adjusted p-value < 0.05, resulting in 21,668 and 26,230 AKAP8 binding peaks in the epithelial and mesenchymal states, respectively. These were the peaks used for downstream analysis. De novo motif analysis was conducted using HOMER v4.10 findMotifsGenome.pl script with the following non-default parameters -p 4 -rna -S 10 -len 4,5,6 -size 100 -chopify. De novo motifs were computed compared to a background of shuffled human introns. Metagenes and other analyses were computed using custom R and python scripts.

Statistical analyses

All data were presented as mean ± standard deviation, unless specifically indicated. Correlation was assessed by pearson correlation. Statistical significance tests included two tailed, unpaired Student's t-tests, Fisher's exact tests, hypergeometric tests, and log-rank tests. p-value < 0.05 was considered statistically significant. $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***) where indicated.

RESULTS

AKAP8 promotes epithelial-state-associated alternative splicing

AKAP8 was only recently reported as an RNA binding protein capable of binding to and regulating alternative splicing [145]. Given our observation that AKAP8 interacts with hnRNPM and promotes CD44 minigene splicing, we sought to interrogate whether AKAP8 regulates alternative splicing, especially EMT-related alternative splicing, on a global scale. Thus, we performed deep RNA sequencing using the HMLE/Twist-ER cell lines that expressed control or AKAP8 shRNA in both epithelial and mesenchymal states. We identified 144 and 228 significant AKAP8-regulated alternative splicing events in both epithelial (Fig. 1A) and mesenchymal states (Fig. 1B), respectively (FDR < 0.05, |ΔPSI| ≥ 0.1, average junction reads per cassette event ≥ 20, FPKM ≥ 5). Classification of the AKAP8-mediated alternative splicing showed that the majority of the events fell into cassette skipped exon events, the most common form of alternative splicing (Fig. 1C-D).

Since SEs are the most common type of AKAP8 regulated alternative splicing event, we next overlapped AKAP8 regulated SEs with those regulated during EMT. Our results revealed that more than half of AKAP8-regulated SEs in epithelial cells are also altered during EMT (58 overlapping events out of 104 AKAP8-regulated SEs). Remarkably, the vast majority (96.6%) of the 58 common events are regulated in a discordant direction, indicating that AKAP8 antagonizes EMT-associated alternative splicing (Fig. 2A, Left Panel). This inverse direction of regulation by AKAP8 is unique in the epithelial state. In mesenchymal cells, 52 out of the 151 AKAP8-regulated events overlapped with the SEs that are altered during EMT. However, they showed roughly half

concordant (46.2%) and half discordant (53.8%) directions compared to EMT-associated SEs (Fig. 2A, Right Panel). The fraction of discordantly regulated events in the epithelial state is significantly higher than that in the epithelial state (Fig. 2B, $p < 0.0001$ per Fisher's exact test). These data show that AKAP8 strongly suppresses EMT-associated alternative splicing in a manner that is epithelial state-specific. Interestingly, the gene expression levels of AKAP8 splicing events were not altered in response to AKAP8 knockdown in either of the epithelial or mesenchymal state, indicating that AKAP8-mediated splicing changes are not a secondary effect of transcription (Fig. 3A). This dichotomy is supported by the fact that only 10 SEs were found to be regulated by AKAP8 in both the epithelial and mesenchymal states (Fig. 3B). These 10 overlapping events are all regulated in the same direction (Fig. 3C). Interestingly, seven out of the 10 events overlapped with the EMT-associated SEs and all showed discordant direction with the EMT-associated SEs. These results imply that AKAP8 consistently antagonizes a set of EMT-associated SEs in both cell states (Fig. 3D).

Comparing the transcriptome of control and AKAP8 knockdown cells, we found that genes that are upregulated during EMT show increased expression upon AKAP8 knockdown in the epithelial state, while genes that are downregulated during EMT show decreased expression in the AKAP8-silenced mesenchymal cells (Fig. 2C). Together, these results show that AKAP8 antagonizes EMT-associated alternative splicing across the transcriptome to maintain an epithelial cell-state.

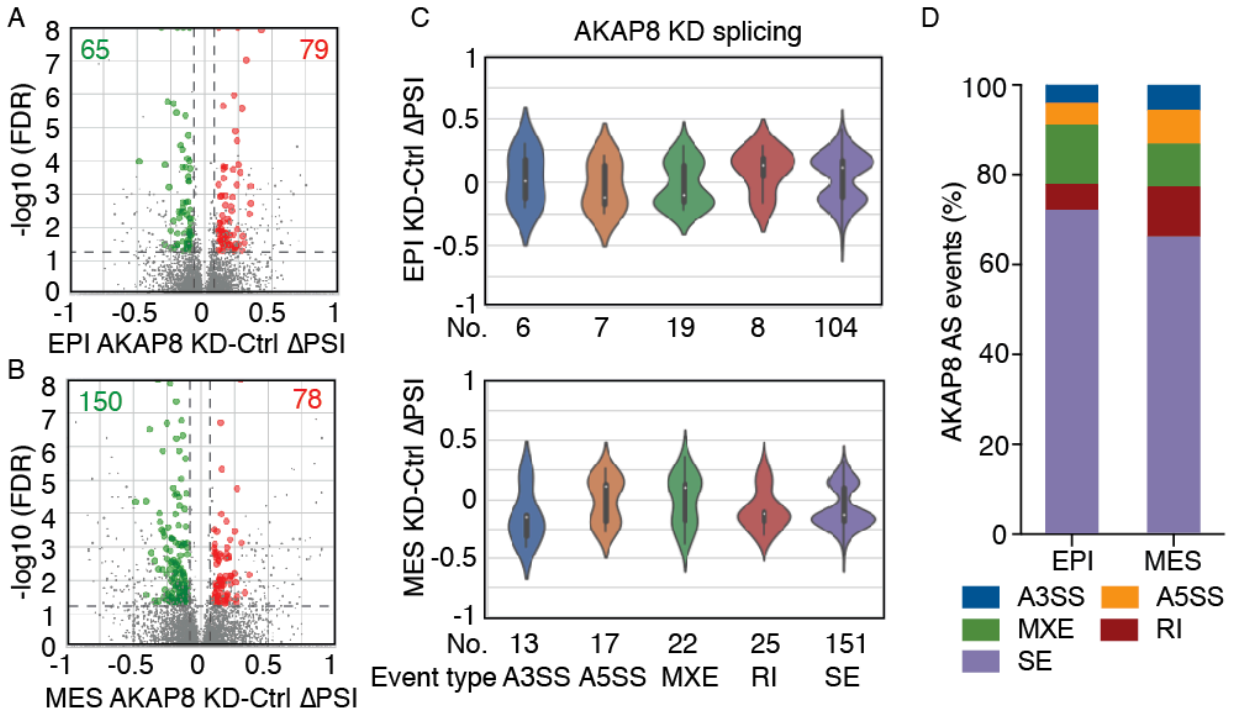

Figure 1. AKAP8 regulates alternative splicing events in epithelial and mesenchymal cell states. (A) Volcano plots showing differential alternative splicing events identified upon knockdown of AKAP8 in epithelial cells (HMLE/Twist-ER Untreated). (B) Differential alternative splicing events identified upon knockdown of AKAP8 in mesenchymal cells (HMLE/Twist-ER Tamoxifen-treated). Significant alternative splicing events (FDR < 0.05 and |deltaPSI| > 0.1) were colored as green and red dots if AKAP8 knockdown resulted in skipping or inclusion, respectively. (C) Violin plot of differential alternative splicing distribution identified upon knockdown of AKAP8 in epithelial cells (top panel) and mesenchymal cells (bottom level) grouped by alternative splicing type. (D) Stacked bar graph of the proportion of different types of AKAP8-dependent alternative splicing events in epithelial and mesenchymal cells. The proportions of each type are largely similar.

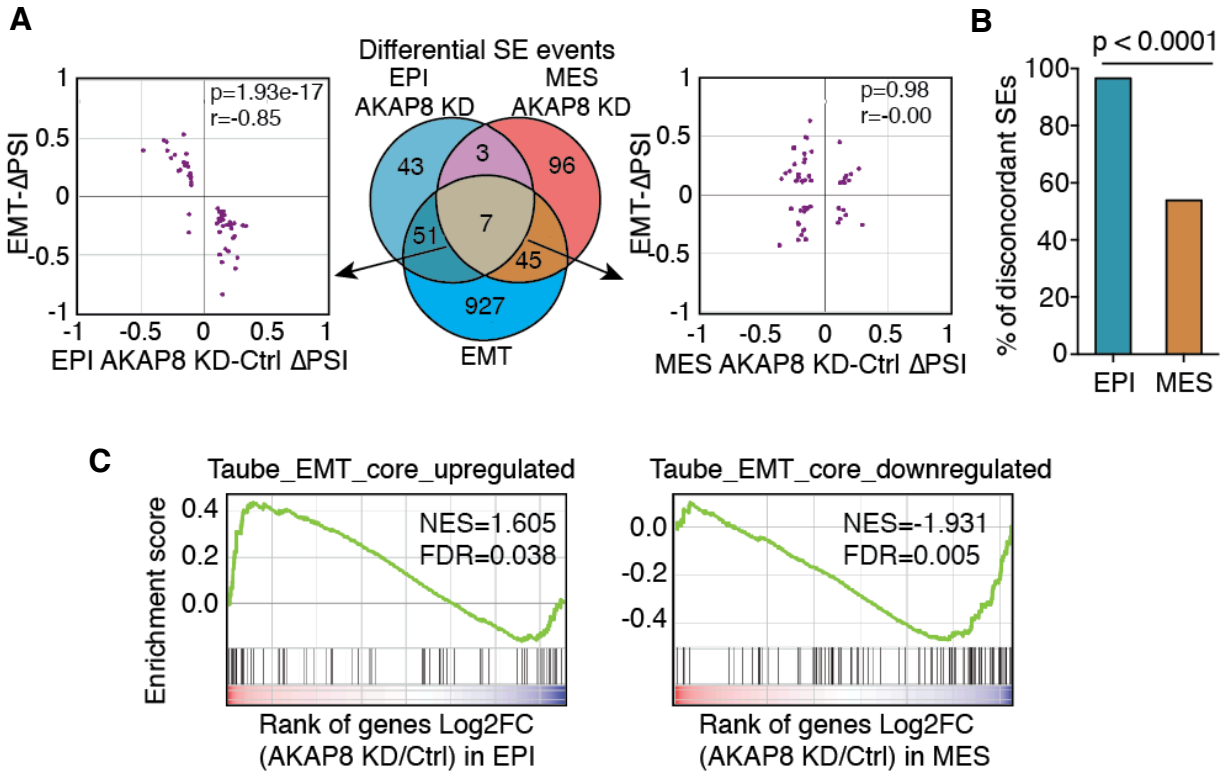

Figure 2. AKAP8 regulates EMT in a cell-state specific manner through alternative splicing. (A) AKAP8 dependent skipped exon events overlap with skipped exons altered during EMT (Middle Panel). Scatterplots of ΔPSI values between EMT-regulated skipped exons and AKAP8-dependent differential skipped exon events identified in epithelial cells (left panel) and mesenchymal cells (right panel). Skipped exons regulated during EMT and by AKAP8 in epithelial cells are highly negatively correlated (r=-0.85, p=1.93e-17) while those in mesenchymal cells are not correlated (r=0.00, p=0.98). (B) 96.6% (56/58) of epithelial AKAP8 differential alternative splicing events are discordant with EMT compared to 53.8% (28/52) of mesenchymal events, which is a significant difference (p computed by Fisher's exact test). (C) (Left panel) AKAP8 knockdown in epithelial cells causes upregulation of genes upregulated during EMT. (Right panel) Similarly, AKAP8 knockdown in the mesenchymal state results in downregulation of genes downregulated during EMT. These data support the observation that AKAP8 knockdown promotes EMT.

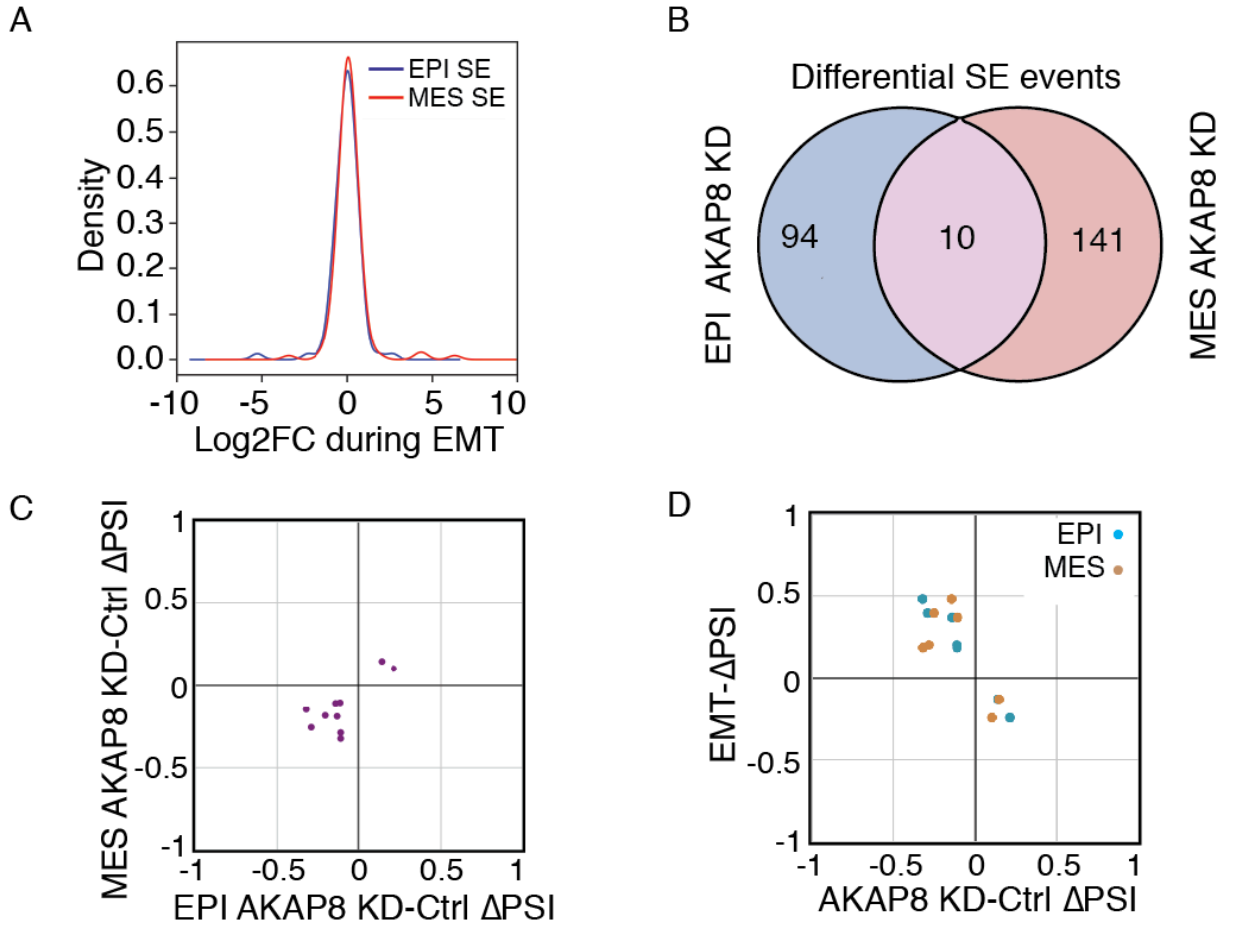

Figure 3. AKAP8 cell-state specific splicing regulons partially overlap and show similar regulatory direction. (A) Kernel Density Estimate plot showing the log2 transformed fold change distribution of genes that undergo significant SE in epithelial (blue line) and mesenchymal (red line) cells. (B) Venn diagram showing overlap between splicing events differentially spliced upon AKAP8 knockdown in epithelial versus mesenchymal cells. (C) Scatterplot showing the ΔPSI values distribution of the shared 10 splicing events regulated by AKAP8 in epithelial and mesenchymal cells. (D) Scatterplot showing the ΔPSI values distribution of the 10 shared splicing events of AKAP8 regulated in epithelial and mesenchymal, as well as EMT. (E) Semi-quantitative RT-PCR analysis of 6 splicing events regulated during EMT that is also regulated by AKAP8 in at least one cell state.

AKAP8 binds to RNA with a consensus motif

To better understand the function of AKAP8 in regulating alternative splicing, we sought to identify high-confidence binding sites for AKAP8 across the transcriptome. We performed single nucleotide resolution enhanced cross-linking and immunoprecipitation (eCLIP) for AKAP8 in the epithelial and mesenchymal cell states. Two eCLIP biological replicates were performed in each cell state, and immunoprecipitation efficiency was unaffected by UV crosslinking (Fig. 4A). High confidence peaks were called using the previously published seCLIP and Irreproducible Discovery Rate analysis pipelines [147]. Biological replicates in each cell state showed a high degree of correlation, highlighting the reproducibility of our assay (Fig. 4B). By normalizing IP signal with size-matched input eCLIP libraries, we obtained quantitative estimations of AKAP8 binding intensity, resulting in 21,665 and 26,228 high confidence AKAP8 binding sites in the epithelial and mesenchymal state, respectively (minimum log2FC(IP/Input) per replicate > 2, -log (adjusted p value) > 3). Mapping the location of AKAP8 binding sites across the gene body revealed that the majority of binding sites are located in distal introns greater than 500 nucleotides from splice sites (Fig. 5A). Interestingly, while AKAP8 showed less binding to distal introns in the epithelial state compared to the mesenchymal state (Fig. 5B. p=9.38e-174 per Fisher's exact test), it binds to proximal intronic regions more significantly in the epithelial state (Fig. 5B. p=6.60e-116 per Fisher's exact test). As binding sites closer to splice sites are more highly correlated with splicing activities, these results suggest that AKAP8 functions more directly in regulating alternative splicing through binding to pre-mRNAs in the epithelial state compared to the mesenchymal state.

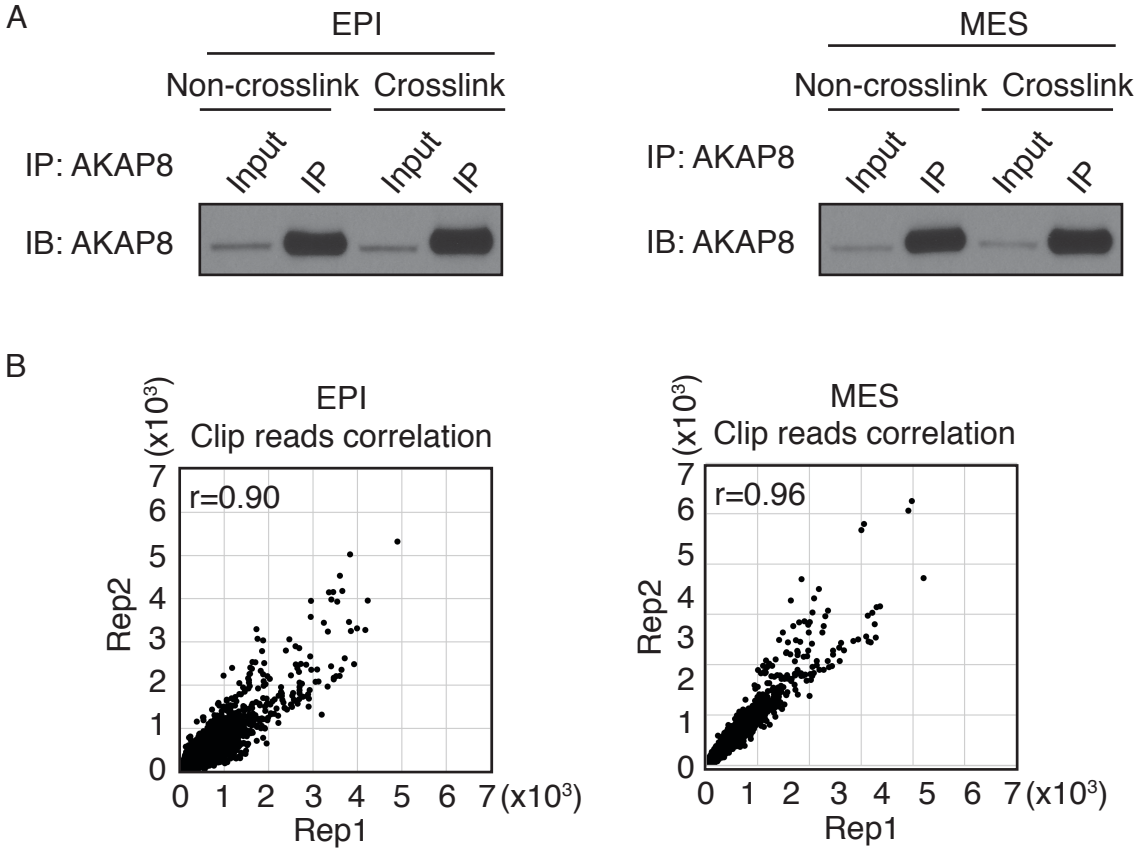

Figure 4. AKAP8 eCLIP reads in epithelial and mesenchymal cell states are highly correlated. (A) Western blot analysis for IP-ed AKAP8 in both crosslinked and non-crosslinked control cells. *Figure completed by Dr. Xiaohui Hu. (B) Scatter plots showing correlation between AKAP8 eCLIP replicates in HMLE Twist-ER untreated epithelial cells (left panel) and tamoxifen treated mesenchymal cells (right panel).

Differences in binding to other gene regions in the 5' and 3' UTR or the coding region showed no statistical differences (Fig. 5B).

We also performed metagene analysis to locate the AKAP8 binding intensity across all human introns and exons (Fig. 5C). We observed significantly stronger binding of AKAP8 to introns compared to exons. Interestingly, AKAP8 intronic binding in the epithelial state appears skewed towards the 5' splice site while binding in the mesenchymal state is distributed more evenly across the intron (Fig. 5C, compare top and bottom plots in Left Panel). Quantification of the binding ability in the proximal introns revelated significantly higher binding in the epithelial state (Fig. 5D), reinforcing our observations that AKAP8 binds more frequently (Fig. 5B) and more strongly (Fig. 5D) to proximal intronic regions in the epithelial state.

Taking advantage of the single-nucleotide precision of eCLIP, we identified AKAP8 high fidelity binding motifs. We took the center of each AKAP8 binding interval and extended the length 100 nucleotides upstream and downstream of each center. Given that RNA binding proteins typically recognize short degenerate motifs and AKAP8 primarily binds introns, we restricted our motif search to 4- 6 nucleotides long against a shuffled background of human intronic regions of 200-nucleotide length. This analysis identified highly significant AKAP8 binding motifs containing guanine stretches of at least three nucleotides flanked by one or two uridine or adenine nucleotides (Fig. 5E).
As expected, AKAP8 shares very similar recognition motifs in epithelial and mesenchymal states.

One example of the AKAP8 binding site is located at the upstream proximal intron of CLSTN1 variable exon 11, where AKAP8 binds in both epithelial and mesenchymal states but with a greater than two-fold binding affinity in the epithelial state (Fig. 5F). Our de novo motif analysis identified four AKAP8 motifs in this peak that are three to four Guanine nucleotides adjacent to Adenine or Uridine nucleotides. Because our RNA sequencing analysis revealed that AKAP8 knockdown in epithelial cells resulted in a decrease in the production of the short isoform of CLSTN1, the isoform that is also reduced during EMT, these results suggest that AKAP8 binds to CLSTN1 and inhibits EMT-mediated alternative splicing, which may cause a blockage of an EMT phenotype. As an EMT-regulated splicing event and is regulated by AKAP8, we interested to see if CLSTN1 isoform level could predict breast cancer patient survival using the TCGA BRCA dataset. Indeed, CLSNT1 stratifies breast cancer patient survival where the short, AKAP8-promoted isoform predicts worse survival (Fig. 6).

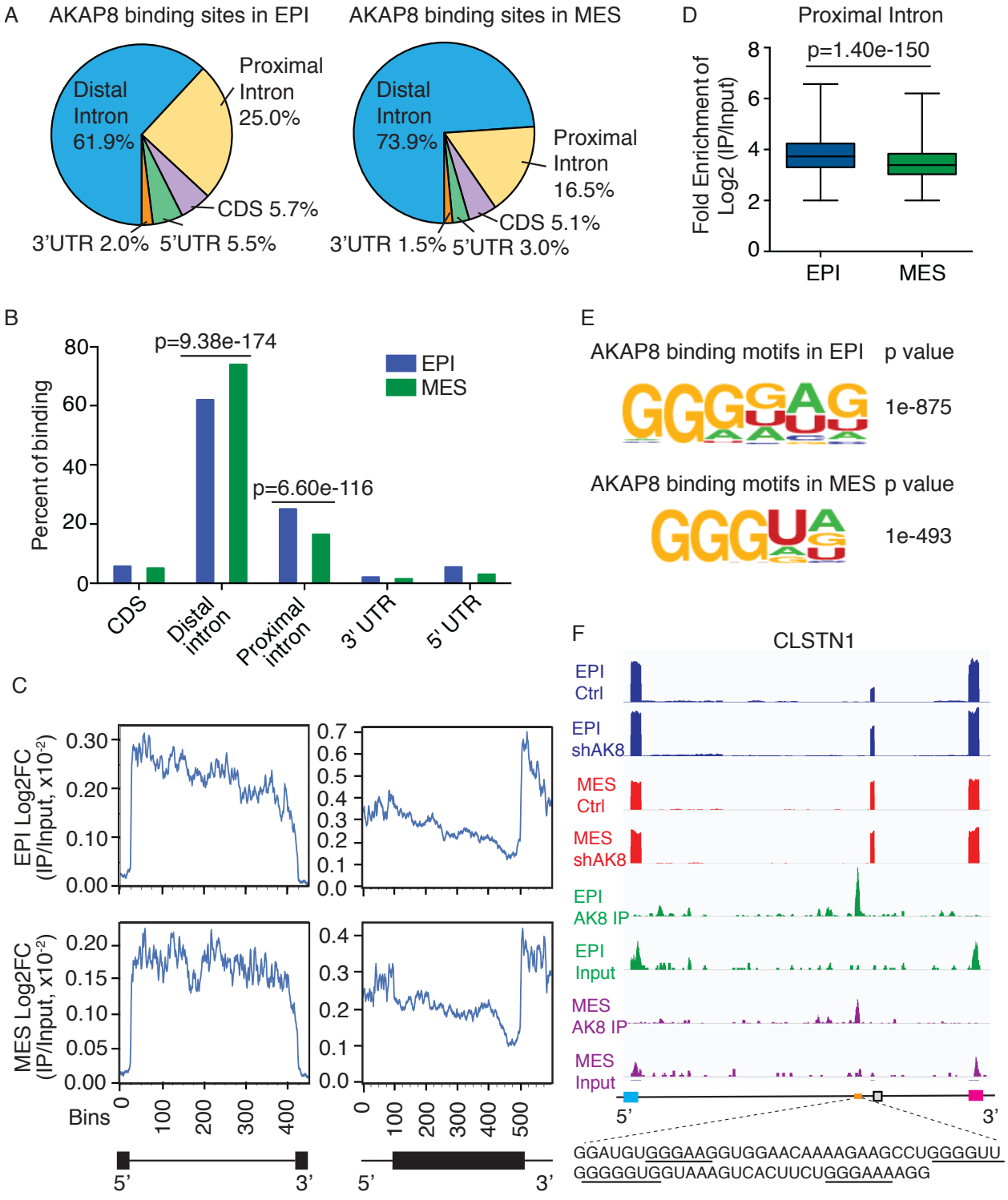

Figure 5: AKAP8 eCLIP identifies cell-state specific binding properties and alternative splicing targets. Figure legends on next page:

Figure 5 Legend: AKAP8 eCLIP identifies cell-state specific binding properties and alternative splicing targets. (A) Distribution of AKAP8 eCLIP binding sites across different regions of gene bodies. AKAP8 shows increased binding to the proximal intron in the epithelial cell state (HMLE/Twist-ER Untreated) compared to the mesenchymal cell state (HMLE/Twist-ER Tamoxifen-treated). (B) Bar plot comparing the percentage of AKAP8 binding sites in each gene region between epithelial and mesenchymal cell states. The proportion of epithelial binding sites binding to distal/proximal intronic regions is significantly lower/greater compared to mesenchymal binding sites. (p computed by Fisher's Exact Test). (C) Metagenes plotting AKAP8 binding intensity (average log2FC IP/Input) in epithelial state (top panels) or mesenchymal state (bottom panels) across all introns (left panels) or exons (right panels). AKAP8 shows a trend of stronger binding downstream of the 5' splice site in the epithelial state compared to mesenchymal state. (D) Boxplots comparing the binding intensity (log2FC IP/Input) of all AKAP8 binding sites located in proximal introns showing that AKAP8 binds more strongly to proximal introns in the epithelial state compared to the mesenchymal state (p computed by Student's t-test, two-tailed). (E) Weblogos depicting the most significant AKAP8 eCLIP binding motif in epithelial and mesenchymal cells identified through de novo motif analysis. The two motifs are similar with guanine stretches of at least 3 nucleotides flanked by uridine or adenine nucleotides. (F) Integrated genome viewer tracks centered on a cassette exon of CLSTN1 which undergoes exon inclusion upon AKAP8 knockdown selectively in the epithelial state. This cassette also undergoes inclusion during EMT. AKAP8 binds comparatively more strongly to the proximal intron upstream of the cassette exon in the epithelial state which is correlated the effect of AKAP8 KD on exon inclusion. Top four tracks represent autoscaled RPM-normalized RNA seq reads. Bottom four tracks represent IP/Input normalized eCLIP signal. Bottom cartoon shows CLSTN1 cassette exon in light blue flanked by upstream (dark blue) and downstream (pink) constitutive exons. Yellow bar indicates AKAP8 eCLIP binding site. Zoomed out sequence indicates nucleotide sequence covered by binding site with G stretches > 3 nucleotides highlighted red and flanking adenine or uridine nucleotides highlighted in green.

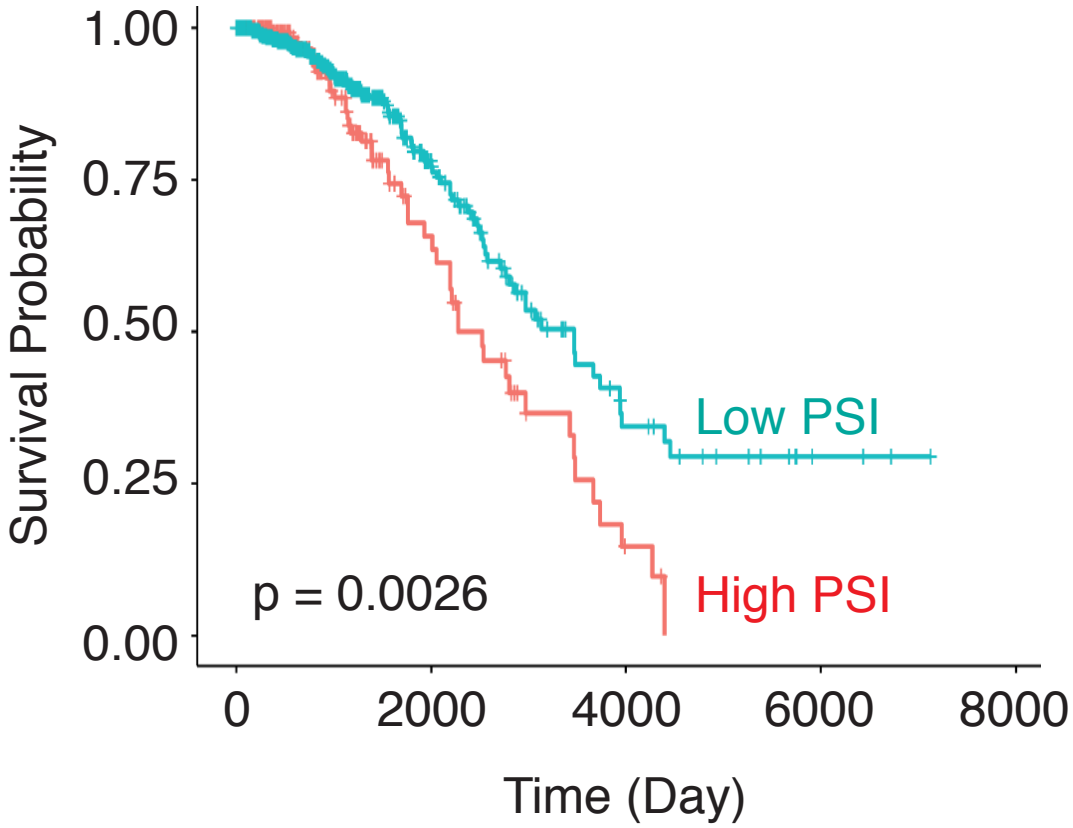

Figure 6: AKAP8 splicing target CLSTN1 predicts breast cancer patient survival in an isoform specific manner. Kaplan Meier plot showing TCGA breast cancer patient survival difference between patients stratified by CLSTN1 splice isoform levels defined by 2-means clustering.

DISCUSSION

Alternative splicing, as an essential step in gene expression, contributes significantly to an ever-growing number of human diseases, especially cancer [135, 148]. In this study, we identified an RNA binding protein AKAP8 as an alternative splicing modulator to inhibit cells from undergoing EMT. AKAP8 is capable of binding RNA consensus sequences such as within the intron of CLSTN1 which decrease in intensity during EMT. These results nominate AKAP8 as a splicing regulator antagonizing EMT-associated alternative splicing

AKAP8 was originally identified as a kinase anchoring protein that recruits protein kinase A (PKA) to nuclear matrix and chromatin structures [149-151]. Interestingly, AKAP8 serves as a DNA binding protein with a preference for GC-rich sequences [152] as well as an RNA binding protein that influence RNA stability and pre-mRNA splicing [145, 153, 154]. Our study presented here provided a functional role for AKAP8 in RNA metabolism during, highlighting the importance of RNA binding protein in regulating cancer processes.

Our genome-wide analysis of AKAP8 regulated RNA demonstrated that AKAP8 regulates alternative splicing, predominantly the most common form of splicing, skipped exons. Most of the AKAP8 regulated alternative splicing events in the epithelial state show opposite directions of splicing regulation compared to those occurring during EMT, suggesting a functional role for AKAP8 in negatively regulating alternative splicing that occurs during EMT. This observation is supported by AKAP8-dependent gene signatures, and importantly, by our experimental examinations. We performed isoform specific

depletion of one of AKAP8 downstream targets CLSTN1 and showed that the CLSTN1-S splice isoform that AKAP8 promotes is decreased during EMT and that depletion of this isoform accelerates EMT, resembling the phenotype caused by AKAP8 depletion. These results provide better understanding towards the mechanism of RNA binding proteins to regulate EMT.

In vivo cross-linking coupled with immunoprecipitation analysis and RNA sequencing revealed the binding sites of AKAP8 across the transcriptome. We found that AKAP8 preferentially binds to introns with less frequent binding to the coding regions or 5' or 3' UTRs. Given the precise resolution of the eCLIP method used in this study, we were able to resolve a short 5-6mer consensus motif consisting of guanine stretches of at least 3 nucleotides long preferentially flanked by adenine or uridine nucleotides. The top motifs were many orders of magnitude more significant than secondary motifs identified by the de novo motif analysis pipeline and were consistent in both Epithelial and Mesenchymal cells, suggesting that AKAP8 binds a similar motif regardless of cell-state and that our motif analysis method is reproducible. A study by Hu et al., 2016 also resolved an AKAP8 motif through de novo motif analysis, where two 12-mer motifs of equal significance were obtained. While these motifs are longer than the ones we report, our findings are somewhat consistent with an AGGAGGA motif identified in the second listed motif in that study. This commonality notwithstanding, we suspect that the motifs derived from our study resolved a more precise AKAP8 motif with a higher level of statistical significance due to the single-nucleotide resolution of our eCLIP methodology and the integration of input normalization into our AKAP8 binding site calling compared

to the less precise RIP-seq method used in the aforementioned study. In addition, we resolved the motifs using a background control of shuffled human introns as this was the most abundant gene region bound by AKAP8; motif normalization was not listed for the methods of the previous study.

Overall, this study identifies AKAP8 as an RNA binding protein that maintains epithelial-cell state specific splicing in the context of increased proximal intron binding in the epithelial state. This study further highlights the critical role RNA binding protein play in the maintenance of cell-state as well as the nuanced role RNA binding proteins play in cell-plasticity and phenotype change.

REFERENCES

1. The Encode Consortium, An integrated encyclopedia of DNA elements in the human genome. Nature, 2012. 489(7414): p. 57-74.

2. Harrow, J., et al., GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res, 2012. 22(9): p. 1760-74.

3. Barash, Y., et al., Deciphering the splicing code. Nature, 2010. 465(7294): p. 53-9.

4. Pan, Q., et al., Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics, 2008. 40(12): p. 1413-5.

5. Wang, E.T., et al., Alternative isoform regulation in human tissue transcriptomes. Nature, 2008. 456(7221): p. 470-476.

6. Liu, S. and C. Cheng, Alternative RNA splicing and cancer. Wiley Interdisciplinary Reviews: RNA, 2013. 4(5): p. 547-66.

7. Cieply, B. and R.P. Carstens, Functional roles of alternative splicing factors in human disease. Wiley Interdiscip Rev RNA, 2015. 6(3): p. 311-26.

8. Shapiro, I.M., et al., An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. PLoS Genetics, 2011. 7(8): p. e1002218.

9. Yang, Y., et al., Determination of a comprehensive alternative splicing regulatory network and combinatorial regulation by key factors during the epithelial-to-mesenchymal transition. Molecular and Cellular Biology, 2016. 36(11): p. 1704-19.

10. Brown, R.L., et al., CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. Journal of Clinical Investigation, 2011. 121(3): p. 1064-74.

11. Reinke, L.M., Y. Xu, and C. Cheng, Snail represses the splicing regulator epithelial splicing regulatory protein 1 to promote epithelial-mesenchymal transition. Journal of Biological Chemistry, 2012. 287(43): p. 36435-42.

12. Warzecha, C.C., et al., An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. The EMBO Journal, 2010. 29(19): p. 3286-300.

13. Thiery, J., Epithelial-mesenchymal transitions in development and pathologies. Current Opinion in Cell Biology, 2003. 15(6): p. 740-6.

14. Nieto, A.M., et al., EMT: 2016. Cell, 2016. 166(1): p. 21-45.

15. Thiery, J.P., et al., Epithelial-mesenchymal transitions in development and disease. Cell, 2009. 139(5): p. 871-90.

16. Yang, J. and R.A. Weinberg, Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. Developmental Cell, 2008. 14(6): p. 818-29.

17. Zhao, P., et al., The CD44s splice isoform is a central mediator for invadopodia activity. J Cell Sci, 2016. 129(7): p. 1355-65.

18. Xu, Y., et al., Cell type-restricted activity of hnRNPM promotes breast cancer metastasis via regulating alternative splicing. Genes & Development, 2014. 28(11): p. 1191-203.

19. Hernandez, J.R., et al., Alternative CD44 splicing identifies epithelial prostate cancer cells from the mesenchymal counterparts. Medical Oncology, 2015. 32(5): p. 159.

20. Lu, H., et al., Exo70 isoform switching upon epithelial-mesenchymal transition mediates cancer cell invasion. Developmental Cell, 2013. 27(5): p. 560-73.

21. Braeutigam, C., et al., The RNA-binding protein Rbfox2: an essential regulator of EMT-driven alternative splicing and a mediator of cellular invasion. Oncogene, 2013. 33(9): p. 1082-92.

22. Warzecha, C.C., et al., The epithelial splicing factors ESRP1 and ESRP2 positively and negatively regulate diverse types of alternative splicing events. RNA Biology, 2009. 6(5): p. 546-562.

23. Passacantilli, I., et al., hnRNPM guides an alternative splicing program in response to inhibition of the PI3K/AKT/mTOR pathway in Ewing sarcoma cells. Nucleic Acids Res, 2017. 45(21): p. 12270-12284.

24. Grille, S.J., et al., The protein kinase Akt induces epithelial mesenchymal transition and promotes enhanced motility and invasiveness of squamous cell carcinoma lines. Cancer Res, 2003. 63(9): p. 2172-8.

25. Warzecha, C.C., et al., ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. Molecular Cell, 2009. 33(5): p. 591-601.

26. Bebee, T.W., et al., The splicing regulators Esrp1 and Esrp2 direct an epithelial splicing program essential for mammalian development. eLife, 2015. 4: p. e08954.

27. Dittmar, K.A., et al., Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing. Molecular and Cellular Biology, 2012. 32(8): p. 1468-82.

28. Huelga, S.C., et al., Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Rep, 2012. 1(2): p. 167-78.

29. Lleres, D., et al., Direct interaction between hnRNP-M and CDC5L/PLRG1 proteins affects alternative splice site choice. EMBO Rep, 2010. 11(6): p. 445-51.

30. Harvey, S.E. and C. Cheng, Methods for characterization of alternative RNA splicing. Methods Mol Biol, 2016. 1402: p. 229-241.

31. Livak, K.J. and T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method. Methods, 2001. 25(4): p. 402-8.

32. Dobin, A., et al., STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England), 2013. 29(1): p. 15-21.

33. Shen, S., et al., rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A, 2014. 111(51): p. E5593-601.

34. Coelho, M.B., et al., Nuclear matrix protein Matrin3 regulates alternative splicing and forms overlapping regulatory networks with PTB. The EMBO Journal, 2015. 34(5): p. 653-68.

35. Cancer Genome Atlas, N., Comprehensive molecular portraits of human breast tumours. Nature, 2012. 490(7418): p. 61-70.

36. Mootha, V.K., et al., PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet, 2003. 34(3): p. 267-73.

37. Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

38. Huang da, W., B.T. Sherman, and R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc, 2009. 4(1): p. 44-57.

39. Huang da, W., B.T. Sherman, and R.A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res, 2009. 37(1): p. 1-13.

40. Kang, Y., et al., A multigenic program mediating breast cancer metastasis to bone. Cancer Cell, 2003. 3(6): p. 537-49.

41. Minn, A.J., et al., Genes that mediate breast cancer metastasis to lung. Nature, 2005. 436(7050): p. 518-24.

42. Weyn-Vanhentenryck, S.M., et al., HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. Cell Rep, 2014. 6(6): p. 1139-1152.

43. Jangi, M., et al., Rbfox2 controls autoregulation in RNA-binding protein networks. Genes Dev, 2014. 28(6): p. 637-51.

44. Taube, J.H., et al., Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. Proceedings of the National Academy of Sciences, 2010. 107(35): p. 15449-54.

45. Korpal, M. and Y. Kang, The emerging role of miR-200 family of microRNAs in epithelial-mesenchymal transition and cancer metastasis. RNA Biol, 2008. 5(3): p. 115-9.

46. Park, S.M., et al., The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. Genes Dev, 2008. 22(7): p. 894-907.

47. Blick, T., et al., Epithelial mesenchymal transition traits in human breast cancer cell lines parallel the CD44 (hi/+) CD24 (lo/-) stem cell phenotype in human breast cancer. J Mammary Gland Biol Neoplasia, 2010. 15(2): p. 235-52.

48. Prat, A., et al., Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res, 2010. 12(5): p. R68.

49. Vanharanta, S., et al., Loss of the multifunctional RNA-binding protein RBM47 as a source of selectable metastatic traits in breast cancer. eLife, 2014. 3(0): p. e02734.

50. Wang, W., et al., Internalized CD44s splice isoform attenuates EGFR degradation by targeting Rab7A. Proceedings of the National Academy of Sciences, 2017. 114(31): p. 8366-8371.

51. Weise, A., et al., Alternative splicing of Tcf7l2 transcripts generates protein variants with differential promoter-binding and transcriptional activation properties at Wnt/beta-catenin targets. Nucleic Acids Res, 2010. 38(6): p. 1964-81.

52. Damianov, A., et al., Rbfox proteins regulate splicing as part of a large multiprotein complex LASR. Cell, 2016. 165(3): p. 606-19.

53. Venables, J.P., et al., RBFOX2 Is an Important Regulator of Mesenchymal Tissue-Specific Splicing in both Normal and Cancer Tissues. Molecular and Cellular Biology, 2013. 33(2): p. 396-405.

54. Sun, H., et al., HnRNPM and CD44s expression affects tumor aggressiveness and predicts poor prognosis in breast cancer with axillary lymph node metastases. Genes Chromosomes Cancer, 2017. 56(8): p. 598-607.

55. Ueda, J., et al., Epithelial splicing regulatory protein 1 is a favorable prognostic factor in pancreatic cancer that attenuates pancreatic metastases. Oncogene, 2014. 33(36): p. 4485-95.

56. Lu, Z.X., et al., Transcriptome-wide landscape of pre-mRNA alternative splicing associated with metastatic colonization. Mol Cancer Res, 2015. 13(2): p. 305-18.

57. Yae, T., et al., Alternative splicing of CD44 mRNA by ESRP1 enhances lung colonization of metastatic cancer cell. Nat Commun, 2012. 3: p. 883.

58. Yao, J., et al., Altered expression and splicing of ESRP1 in malignant melanoma correlates with epithelial-mesenchymal status and tumor-associated immune cytolytic activity. Cancer Immunol Res, 2016. 4(6): p. 552-61.

59. Jeong, H.M., et al., ESRP1 is overexpressed in ovarian cancer and promotes switching from mesenchymal to epithelial phenotype in ovarian cancer cells. Oncogenesis, 2017. 6(11): p. e391.

60. Manning, K.S. and T.A. Cooper, The roles of RNA processing in translating genotype to phenotype. Nat Rev Mol Cell Biol, 2017. 18(2): p. 102-114.

61. Gerstberger, S., M. Hafner, and T. Tuschl, A census of human RNA-binding proteins. Nat Rev Genet, 2014. 15(12): p. 829-45.

62. Yang, Y., et al., Determination of a comprehensive alternative splicing regulatory network and combinatorial regulation by key factors during the epithelial-to-mesenchymal transition. Mol Cell Biol, 2016. 36(11): p. 1704-19.

63. Li, J., et al., An alternative splicing switch in FLNB promotes the mesenchymal cell state in human breast cancer. Elife, 2018. 7.

64. Pillman, K.A., et al., miR-200/375 control epithelial plasticity-associated alternative splicing by repressing the RNA-binding protein quaking. EMBO J, 2018. 37(13).

65. Harvey, S.E., et al., Coregulation of alternative splicing by hnRNPM and ESRP1 during EMT. RNA, 2018. 24(10): p. 1326-1338.

66. Yuan, Y., et al., Cell type-specific CLIP reveals that NOVA regulates cytoskeleton interactions in motoneurons. Genome Biol, 2018. 19(1): p. 117.

67. Love, M.I., W. Huber, and S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 2014. 15(12): p. 550.

68. Flynn, R.A., et al., Dissecting noncoding and pathogen RNA-protein interactomes. RNA, 2015. 21(1): p. 135-43.

69. Shah, A., et al., CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. Bioinformatics, 2017. 33(4): p. 566-567.

70. Damianov, A., et al., Rbfox Proteins regulate splicing as part of a large multiprotein complex LASR. Cell, 2016. 165(3): p. 606-19.

71. Huelga, S.C., et al., Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Reports, 2012. 1(2): p. 167-78.

72. Fu, X.D. and M. Ares, Jr., Context-dependent control of alternative splicing by RNA-binding proteins. Nat Rev Genet, 2014. 15(10): p. 689-701.

73. Feng, H., et al., Modeling RNA-binding protein specificity in vivo by precisely registering protein-RNA crosslink sites. bioRxiv, 2018: p. 428615.

74. Van Nostrand, E.L., et al., A Large-scale binding and functional map of human RNA binding proteins. bioRxiv, 2017: p. 179648.

75. Jangi, M. and P.A. Sharp, building robust transcriptomes with master splicing factors. Cell, 2014. 159(3): p. 487-98.

76. Graveley, B.R., Alternative splicing: increasing diversity in the proteomic world. Trends Genet, 2001. 17(2): p. 100-7.

77. Johnson, J.M., et al., Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science, 2003. 302(5653): p. 2141-4.

78. Sultan, M., et al., A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 2008. 321(5891): p. 956-60.

79. Zhang, J. and J.L. Manley, Misregulation of pre-mRNA alternative splicing in cancer. Cancer Discov, 2013. 3(11): p. 1228-37.

80. Rosenberg, A.B., et al., Learning the sequence determinants of alternative splicing from millions of random sequences. Cell, 2015. 163(3): p. 698-711.

81. Kumari, S., et al., An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation. Nat Chem Biol, 2007. 3(4): p. 218-21.

82. Agarwal, T., et al., RNA G-quadruplexes: G-quadruplexes with "U" turns. Curr Pharm Des, 2012. 18(14): p. 2102-11.

83. Bugaut, A. and S. Balasubramanian, 5′-UTR RNA G-quadruplexes: translation regulation and targeting. Nucleic Acids Res, 2012. 40(11): p. 4727-41.

84. Millevoi, S., H. Moine, and S. Vagner, G-quadruplexes in RNA biology. Wiley Interdiscip Rev RNA, 2012. 3(4): p. 495-507.

85. Blice-Baum, A.C. and M.R. Mihailescu, Biophysical characterization of G-quadruplex forming FMR1 mRNA and of its interactions with different fragile X mental retardation protein isoforms. RNA, 2014. 20(1): p. 103-14.

86. Marcel, V., et al., G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms. Carcinogenesis, 2011. 32(3): p. 271-8.

87. Kwok, C.K., et al., rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. Nat Methods, 2016. 13(10): p. 841-4.

88. Liu, X., et al., Structure-dependent binding of hnRNPA1 to telomere RNA. Journal of the American Chemical Society, 2017. 139(22): p. 7533-7539.

89. Zhang, K., et al., The C9orf72 repeat expansion disrupts nucleocytoplasmic transport. Nature, 2015. 525(7567): p. 56-61.

90. Conlon, E.G., et al., The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains. Elife, 2016. 5.

91. von Hacht, A., et al., Identification and characterization of RNA guanine-quadruplex binding proteins. Nucleic Acids Res, 2014. 42(10): p. 6630-44.

92. Lattmann, S., et al., Role of the amino terminal RHAU-specific motif in the recognition and resolution of guanine quadruplex-RNA by the DEAH-box RNA helicase RHAU. Nucleic Acids Res, 2010. 38(18): p. 6219-33.

93. Garneau, D., et al., Heterogeneous nuclear ribonucleoprotein F/H proteins modulate the alternative splicing of the apoptotic mediator Bcl-x. J Biol Chem, 2005. 280(24): p. 22641-50.

94. Khateb, S., et al., The tetraplex (CGG)n destabilizing proteins hnRNP A2 and CBF-A enhance the in vivo translation of fragile X premutation mRNA. Nucleic Acids Res, 2007. 35(17): p. 5775-88.

95. Thiery, J.P. and J.P. Sleeman, Complex networks orchestrate epithelial-mesenchymal transitions. Nat Rev Mol Cell Biol, 2006. 7(2): p. 131-42.

96. Soret, J., et al., Selective modification of alternative splicing by indole derivatives that target serine-arginine-rich protein splicing factors. Proceedings of the National Academy of Sciences of the United States of America, 2005. 102(24): p. 8764-8769.

97. Arslan, A.D., et al., A high throughput assay to identify small molecule modulators of alternative pre-mRNA splicing. Journal of biomolecular screening, 2013. 18(2): p. 180-190.

98. Stoilov, P., et al., A high-throughput screening strategy identifies cardiotonic steroids as alternative splicing modulators. Proc Natl Acad Sci U S A, 2008. 105(32): p. 11218-23.

99. Zhang, J., S.E. Harvey, and C. Cheng, A high-throughput screen identifies small molecule modulators of alternative splicing by targeting RNA G-quadruplexes. Nucleic Acids Res, 2019. 47(7): p. 3667-3679.

100. Huang, H., et al., RNA G-quadruplex secondary structure promotes alternative splicing via the RNA-binding protein hnRNPF. Genes & development, 2017. 31(22): p. 2296-2309.

101. Huang, H., Y. Xu, and C. Cheng, Detection of alternative splicing during epithelial-mesenchymal transition. J Vis Exp, 2014(92): p. e51845.

102. Trapnell, C., et al., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 2010. 28(5): p. 511-5.

103. Katz, Y., et al., Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods, 2010. 7(12): p. 1009-15.

104. Kikin, O., L. D'Antonio, and P.S. Bagga, QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic Acids Res, 2006. 34(Web Server issue): p. W676-82.

105. Dardenne, E., et al., RNA helicases DDX5 and DDX17 dynamically orchestrate transcription, miRNA, and splicing programs in cell differentiation. Cell Rep, 2014. 7(6): p. 1900-13.

106. Ji, X., et al., Research progress of RNA quadruplex. Nucleic Acid Ther, 2011. 21(3): p. 185-200.

107. Kikin, O., et al., GRSDB2 and GRS_UTRdb: databases of quadruplex forming G-rich sequences in pre-mRNAs and mRNAs. Nucleic Acids Res, 2008. 36(Database issue): p. D141-8.

108. Huppert, J.L., et al., G-quadruplexes: the beginning and end of UTRs. Nucleic Acids Res, 2008. 36(19): p. 6260-8.

109. Eddy, J. and N. Maizels, Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. Nucleic Acids Res, 2008. 36(4): p. 1321-33.

110. Kiledjian, M. and G. Dreyfuss, Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. Embo J, 1992. 11(7): p. 2655-64.

111. Xiao, X., et al., Splice site strength-dependent activity and genetic buffering by poly-G runs. Nat Struct Mol Biol, 2009. 16(10): p. 1094-100.

112. Decorsiere, A., et al., Essential role for the interaction between hnRNP H/F and a G quadruplex in maintaining p53 pre-mRNA 3'-end processing and function during DNA damage. Genes Dev, 2011. 25(3): p. 220-5.

113. Samatanga, B., et al., The high kinetic stability of a G-quadruplex limits hnRNP F qRRM3 binding to G-tract RNA. Nucleic Acids Res, 2013. 41(4): p. 2505-16.

114. Dominguez, C., et al., Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. Nat Struct Mol Biol, 2010. 17(7): p. 853-61.

115. Dominguez, C. and F.H. Allain, NMR structure of the three quasi RNA recognition motifs (qRRMs) of human hnRNP F and interaction studies with Bcl-x G-tract RNA: a novel mode of RNA recognition. Nucleic Acids Res, 2006. 34(13): p. 3634-45.

116. Matunis, M.J., J. Xing, and G. Dreyfuss, The hnRNP F protein: unique primary structure, nucleic acid-binding properties, and subcellular localization. Nucleic Acids Res, 1994. 22(6): p. 1059-67.

117. Wolfe, A.L., et al., RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. Nature, 2014. 513(7516): p. 65-70.

118. Gomez, D., et al., Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing. Nucleic Acids Res, 2004. 32(1): p. 371-9.

119. Guo, J.U. and D.P. Bartel, RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. Science, 2016. 353(6306).

120. S Akinboye, E. and O. Bakare, Biological activities of emetine. The Open Natural Products Journal, 2011. 4(1).

121. Boon-Unge, K., et al., Emetine regulates the alternative splicing of Bcl-x through a protein phosphatase 1-dependent mechanism. Chem Biol, 2007. 14(12): p. 1386-92.

122. Zamiri, B., et al., TMPyP4 porphyrin distorts RNA G-quadruplex structures of the disease-associated r(GGGGCC)n repeat of the C9orf72 gene and blocks interaction of RNA-binding proteins. J Biol Chem, 2014. 289(8): p. 4653-9.

123. Morris, M.J., et al., The porphyrin TmPyP4 unfolds the extremely stable G-quadruplex in MT3-MMP mRNA and alleviates its repressive effect to enhance translation in eukaryotic cells. Nucleic Acids Res, 2012. 40(9): p. 4137-45.

124. Cammas, A. and S. Millevoi, RNA G-quadruplexes: emerging mechanisms in disease. Nucleic Acids Research, 2017. 45(4): p. 1584-1595.

125. Bugaut, A., et al., Small molecule-mediated inhibition of translation by targeting a native RNA G-quadruplex. Org Biomol Chem, 2010. 8(12): p. 2771-6.

126. Frixen, U.H., et al., E-cadherin-mediated cell-cell adhesion prevents invasiveness of human carcinoma cells. J Cell Biol, 1991. 113(1): p. 173-85.

127. Sabbah, M., et al., Molecular signature and therapeutic perspective of the epithelial-to-mesenchymal transitions in epithelial cancers. Drug Resist Updat, 2008. 11(4-5): p. 123-51.

128. Yang, J., et al., Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. Cell, 2004. 117(7): p. 927-39.

129. Pattabiraman, D.R., et al., Activation of PKA leads to mesenchymal-to-epithelial transition and loss of tumor-initiating ability. Science, 2016. 351(6277): p. aad3680.

130. Tam, W.L. and R.A. Weinberg, The epigenetics of epithelial-mesenchymal plasticity in cancer. Nat Med, 2013. 19(11): p. 1438-49.

131. Lamouille, S., J. Xu, and R. Derynck, Molecular mechanisms of epithelial-mesenchymal transition. Nat Rev Mol Cell Biol, 2014. 15(3): p. 178-96.

132. Warzecha, C.C. and R.P. Carstens, Complex changes in alternative pre-mRNA splicing play a central role in the epithelial-to-mesenchymal transition (EMT). Semin Cancer Biol, 2012. 22(5-6): p. 417-27.

133. Kim, M.S., et al., A draft map of the human proteome. Nature, 2014. 509(7502): p. 575-81.

134. Baralle, F.E. and J. Giudice, Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol, 2017. 18(7): p. 437-451.

135. Scotti, M.M. and M.S. Swanson, RNA mis-splicing in disease. Nat Rev Genet, 2016. 17(1): p. 19-32.

136. Cheng, C. and P.A. Sharp, Regulation of CD44 alternative splicing by SRm160 and its potential role in tumor cell invasion. Mol Cell Biol, 2006. 26(1): p. 362-70.

137. Zhang, H., et al., CD44 splice isoform switching determines breast cancer stem cell state. Genes Dev, 2019. 33(3-4): p. 166-179.

138. Bhattacharya, R., et al., Mesenchymal splice isoform of CD44 (CD44s) promotes EMT/invasion and imparts stem-like properties to ovarian cancer cells. J Cell Biochem, 2018. 119(4): p. 3373-3383.

139. Chen, L., et al., Snail driving alternative splicing of CD44 by ESRP1 enhances invasion and migration in epithelial ovarian cancer. Cell Physiol Biochem, 2017. 43(6): p. 2489-2504.

140. Miwa, T., et al., Isoform switch of CD44 induces different chemotactic and tumorigenic ability in gallbladder cancer. Int J Oncol, 2017. 51(3): p. 771-780.

141. Preca, B.T., et al., A self-enforcing CD44s/ZEB1 feedback loop maintains EMT and stemness properties in cancer cells. Int J Cancer, 2015. 137(11): p. 2566-77.

142. Sakuma, K., et al., HNRNPLL, a newly identified colorectal cancer metastasis suppressor, modulates alternative splicing of CD44 during epithelial-mesenchymal transition. Gut, 2018. 67(6): p. 1103-1111.

143. Zhang, F.L., et al., Cancer-associated MORC2-Mutant M276I regulates an hnRNPM-mediated CD44 splicing switch to promote invasion and metastasis in triple-negative breast cancer. Cancer Res, 2018. 78(20): p. 5780-5792.

144. Pradella, D., et al., EMT and stemness: flexible processes tuned by alternative splicing in development and cancer progression. Mol Cancer, 2017. 16(1): p. 8.

145. Hu, J., et al., AKAP95 regulates splicing through scaffolding RNAs and RNA processing factors. Nat Commun, 2016. 7: p. 13347.

146. Van Nostrand, E.L., et al., Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods, 2016. 13(6): p. 508-14.

147. Van Nostrand, E.L., et al., Robust, cost-effective profiling of RNA binding protein targets with single-end enhanced crosslinking and immunoprecipitation (seCLIP). Methods Mol Biol, 2017. 1648: p. 177-200.

148. Srebrow, A. and A.R. Kornblihtt, The connection between splicing and cancer. J Cell Sci, 2006. 119(Pt 13): p. 2635-41.

149. Coghlan, V.M., et al., Cloning and characterization of AKAP 95, a nuclear protein that associates with the regulatory subunit of type II cAMP-dependent protein kinase. J Biol Chem, 1994. 269(10): p. 7658-65.

150. Eide, T., et al., Protein kinase A-anchoring protein AKAP95 interacts with MCM2, a regulator of DNA replication. J Biol Chem, 2003. 278(29): p. 26750-6.

151. Kubota, S., et al., Role for tyrosine phosphorylation of A-kinase anchoring protein 8 (AKAP8) in its dissociation from chromatin and the nuclear matrix. J Biol Chem, 2015. 290(17): p. 10891-904.

152. Marstad, A., et al., A-kinase anchoring protein AKAP95 is a novel regulator of ribosomal RNA synthesis. Febs J, 2016. 283(4): p. 757-70.

153. Jungmann, R.A. and O. Kiryukhina, Cyclic AMP and AKAP-mediated targeting of protein kinase A regulates lactate dehydrogenase subunit A mRNA stability. J Biol Chem, 2005. 280(26): p. 25170-7.

154. Kvissel, A.K., et al., Involvement of the catalytic subunit of protein kinase A and of HA95 in pre-mRNA splicing. Exp Cell Res, 2007. 313(13): p. 2795-809.