Title

Interactions between RNA-binding proteins and transposable elements and their functional implication in post-transcriptional regulation

Author

Wang, Chengyang

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Interactions between RNA-binding proteins and transposable elements and their functional implication in post-transcriptional regulation

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Bioinformatics

by

Chengyang Wang

2022

ABSTRACT OF THE DISSERTATION

Interaction between RNA-binding proteins and transposable elements and its functional implication in post-transcriptional regulation

by

Chengyang Wang

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2022

Professor Yi Xing, Co-Chair

Professor Qing Zhou, Co-Chair

Gene-embedded transposable elements (TE) significantly influence RNA processing. A variety of RNA-binding proteins (RBPs) exert post-transcriptional regulation via TE binding. Transcriptome-wide identification of RBP binding sites can be accomplished by UV crosslinking and immunoprecipitation, followed by sequencing (CLIP-seq). However, the technical demands of CLIP and the repetitive nature of TEs present challenges to the large-scale investigation of the interplay between RBPs and TEs. Addressing these challenges requires the development of specialized computational approaches.

In the first part of the dissertation, we present a dedicated RBP-centric computational framework for the systematic study of RBP-TE interactions. In this framework, we use both multi-mapped

reads and uniquely mapped reads to recover RBP binding sites on transposable element. By applying this framework to a unified resource of 223 eCLIP-seq datasets from ENCODE, we observed extensive binding of a wide range of RNA-binding proteins to three major TE families: L1, L2 and Alu. For most RBPs, their motif frequencies in TE families with which they interact are higher than the average frequency of the motif over all TE families. Furthermore, we investigate the functional effects of RBP-TE interaction on TE exonization, a process of incorporation of intronic TEs into mature RNAs. This process usually have undesirable consequences, so mechanisms exist for repressing it. (e.g., MATR3 for repressing exonization of antisense L1 elements and HNRNPC for repressing exonization of antisense Alu elements). We identified two novel repressors for TE exonization: HNRNPM for antisense L1 and XRCC6(Ku70) for antisense Alu. XRCC6(Ku70) is previously known as a DNA-binding protein engaged in the DNA repair pathway. We found the selective repression of a set of antisense Alu exons by XRCC6(Ku70) and the strengthened XRCC6 binding in the close vicinity of 3' splice sites of these exons. More intriguingly, our analysis showed that XRCC6 can provide additional repressiveness for Alu exons which have a relatively short continuous U-tract in the proximal upstream of 3'SS, on which the effects of the global Alu repressor HNRNPC are compromised.

In the second chapter, we further disclose the functional implication of RBP-TE interactions on other post-transcriptional events, including RNA editing and RNA stability. By integrating RBP binding with differential RNA editing, we found that ILF3 can suppress RNA editing at sites in inverted repeat Alu elements. Besides, we showed that UPF1, the core factor of the pathway of nonsense-mediated mRNA decay, can enable RNA decay by binding to Alu elements on 3'UTR.

Taken together, our analysis improves our understanding of RBP-TE interplay and further illustrates functional implications of these interactions in post-transcriptional regulation.

The dissertation of Chengyang Wang is approved.

Kathrin Plath

Linda Liau

Yi Xing, Committee Co-Chair

Qing Zhou, Committee Co-Chair

University of California, Los Angeles

2022

To my parents and my grandparents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to express my deepest gratitude to my advisor, Dr. Yi Xing, who has been providing numerous supports for my research studies and career development. In his guidance, I have learnt how to creatively design a research topic, how to rigorously carry out research work, how to efficiently communicate with collaborators, and how to clearly deliver scientific ideas to others. In addition, I really appreciate his patience with me in the past years. All the things I have learned from my advisor will be priceless for my future career.

I really appreciate the enormous support and insightful advice from my committee members: Kathrin Plath, Linda Liau and Qing Zhou. It is my pleasure to have them in my committee.

I am thankful to my collaborators, Wankun Deng and Xinjun Ji from The Children's Hospital of Philadelphia, for their invaluable help and support. Wankun helped me in processing the eCLIP-seq data from ENCODE and improving figure quality. Xinjun not only carried out many experiments for the study, but also helped me in revising the thesis.

My life as a Ph.D student would not be so great without my friends and colleagues in the Xing lab: Zhicheng Pan, Xinjun Ji, Wankun Deng, Zijun Zhang, Juw Won Park, Yang Pan, Yida Zhang, Yuanyuan Wang, Yuan Gao, Yongbo Wang, Yan Gao, Ruijiao Xin, Samir Adhikari, Levon Demirdjian, Eddie Park, Yang Guo, Amal Katrib, Yang Xu, Zhixiang Lu, Yu-ting Tseng, Feng Wang and Yungang Xu.

Finally, I would like to thank my parents and my grandparents for their unconditional support throughout my life. Thank Shanxi Jiang and Yuchen Jin, who trust me all the time.

VITA

EDUCATION

- University of California, Los Angeles 2014 – 2022

  Ph.D. Candidate in Bioinformatics

  Ph.D. advisor: Dr. Yi Xing

- Tongji University, Shanghai 2011 – 2014

  Advisor: Dr. X. Shirley Liu and Dr. Cheng Li

  M.S. in Bioinformatics

- Nankai University, Tianjin 2007 – 2011

  B.S. in Mathematics

HONORS & AWARDS

- University Fellowship, UCLA, Jul. 2014 – Jun. 2015
- Guang Hua Scholarship, Tongji University, 2012

TEACHING EXPERIENCE

- Teaching Assistant. MIMG 180B: Scientific Analysis and Communication, UCLA

PUBLICATIONS

- **Chengyang Wang***, Wankun Deng*, Xinjun Ji, Yi Xing. Interaction between RNA-binding proteins and transposable elements and its functional implication in post-transcriptional regulation. (*Equal Contribution) Under review.

- Jing Zhang*, **Chengyang Wang***, Xi Chen, Mamoru Takada, Cheng Fan,Xingnan Zheng, Haitao Wen, Yong Liu,1 Chenguang Wang, Richard G. Pestell, Katherine M.Aird, William G. Kaelin Jr.,X. Shirley Liu, and Qing Zhang. EglN2 associates with the NRF1-PGC1α complex and controls mitochondrial function in breast cancer. (2015). *The EMBO Journal*. (*Equal Contribution)
- modENCODE Consortium. Comparative analysis of metazoan chromatin architecture. (2014). *Nature*.
- **Chengyang Wang**, Rui Tian, Qian Zhao, Han Xu, Clifford A. Meyer, Cheng Li, Yong Zhang and X. Shirley Liu. Computational inference of mRNA stability from histone modification and transcriptome profiles. (2012). *Nucleic Acids Research*

Chapter 1 Introduction

The process of the information in DNA being passed to RNA is called transcription. RNA molecules which have been transcribed are controlled by post-transcriptional regulation before being translated into proteins in the end[1]. The key players in post-transcriptional regulation are RNA-binding proteins, which can regulate a series of post-transcriptional events, such as alternative splicing, 3'polyadenylation, RNA editing, RNA modification and RNA stability[2,3]. There are approximately 1500 RNA-binding proteins in human cells[4]. The transcriptome-wide binding sites of RBPs can be determined by CLIP-seq, which combines UV cross-linking with immunoprecipitation followed by high-throughput sequencing[5,6]. Despite the importance of RBPs in cell biology, the function of most RBPs has yet to be fully deciphered. Of noted, emerging evidence has shown that an array of DNA-binding proteins also has binding affinities to RNA molecules and can affect the fate of their RNA targets[7-9]. Collectively, studying the interaction between proteins and RNAs and the functional implication of their interaction is fundamental in realizing the RNA world.

Transposable elements (TEs), also known as transposons, are DNA sequences that can change their locations from one place to another within a genome[10-13]. There are two classes of TEs, retrotransposons and DNA transposons. The former class can duplicate their copies on a genome through "the copy and paste" mechanism, where TE DNA is transcribed into RNA, and then the RNA is reverse transcribed back to DNA at another location. DNA transposons use the "cut and paste" mechanism to directly jump onto a different place without creating new copies.

TEs comprise half of human genome[14,15] and the majority of TEs are actually immobile any longer due to accumulated mutations or epigenetics silencing[16]. Recent studies have shown that a few TEs can gain new functions by co-opting with the host genome[10-13]. For examples, a large number of lncRNA contain TEs[17]; some of TEs can serve as cis-regulatory DNA elements[18]. Nevertheless, most TEs have neutral or deleterious effects in the host genome. More than 120 TE elements have been documented to be involved with human diseases[19].

Intragenic transposable elements are co-transcribed into pre-mRNA, and even kept as exons in the mature mRNA (a process called TE exonization)[20,21]. Given this, it is tempting to ask whether transcribed TEs can be recognized by RNA-binding proteins. If so, what is the implication of their interaction in post-transcriptional regulation. To this end, we systematically interrogate the interplay between RBPs and TEs in K562 cells and HepG2 cells using a dedicated analytical framework. The framework uses both multi-mapped reads and uniquely mapped reads of eCLIP-seq to recover RBP binding sites on TEs, which are highly repetitive and thus less amenable for short-reads mapping. We discovered a widespread interaction between RBPs and TEs. The recognized TEs are mainly composed of L1 family, L2 family and Alu family. In accord with the preferential binding, these TE families are highly enriched with binding motifs of their corresponding interacting RBPs.

In Chapter 2, we further disclose how the TE-interacting RBPs can repress TE exonization. A few repressors of TE exonization have been discovered. For examples, MATR3 suppresses exonization of L1/L2 elements by promoting PTBP1 binding on multivalent binding sites within

LINEs[22]. HNRNPC functions as a general repressor of antisense Alu exonization via binding to the continuous U-tract in the proximal upstream of Alu-encoded 3' splice sites[23].

Our analysis showed that L1-binder HNRNPM can repress exonization of antisense L1. The repression is highly consistent with the binding enrichment of HNRNPM around splice sites within antisense L1 elements. Additionally, we found that Alu-binder XRCC6(Ku70) can repress exonization of antisense Alu elements. XRCC6 was famous for its pivotal function of recognizing DNA damage break points[24]. At the same time, increasing evidence has recently demonstrated its direct binding to diverse RNA molecules[25-31]. The successful conduction of XRCC6 eCLIP-seq also confirms its RNA-binding ability. Furthermore, we found that XRCC6(Ku70) provides additional repressive safeguard for antisense Alu exons with relatively short continuous U-tract in proximal upstream of 3' splice sites, on which the effects of the global Alu repressor HNRNPC are compromised. To our knowledge, XRCC6 is the secondly discovered repressor for Alu exonization by now. Our work also manifests the multifunctionality of a canonical DNA binding protein.

In Chapter 3, we further discuss the functional implications of TE-interacting RBPs on other post-transcriptional events, including RNA editing and RNA stability. RNA editing can make sequence changes on an RNA molecule without changing its underlying DNA sequence[32-35]. The most prominent type of RNA editing is A-to-I editing, which is mainly catalyzed by ADAR enzyme family[36,37]. Here, we found that ILF3 can specifically suppress RNA editing levels at its binding sites in Alu-Alu duplexes, which are formed by base pairing of inverted repeat Alus. For

the effects on RNA stability, we showed that UPF1, the critical engaging factors of nonsense mediated decay pathway[38], can enable RNA decay by binding to Alu elements within 3'UTR.

References

1. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-3 (1970).

2. Keene, J.D. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* **8**, 533-43 (2007).

3. Glisovic, T., Bachorik, J.L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* **582**, 1977-86 (2008).

4. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat Rev Genet* **15**, 829-45 (2014).

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

6. Van Nostrand, E.L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711-719 (2020).

7. Holmes, Z.E. *et al.* The Sox2 transcription factor binds RNA. *Nat Commun* **11**, 1805 (2020).

8. Hudson, W.H. & Ortlund, E.A. The structure, function and evolution of proteins that bind DNA and RNA. *Nat Rev Mol Cell Biol* **15**, 749-60 (2014).

9. Cassiday, L.A. & Maher, L.J., 3rd. Having it both ways: transcription factors that bind DNA and RNA. *Nucleic Acids Res* **30**, 4118-26 (2002).

10. Chuong, E.B., Elde, N.C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**, 71-86 (2017).

11. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol* **19**, 199 (2018).

12. Burns, K.H. Transposable elements in cancer. *Nat Rev Cancer* **17**, 415-424 (2017).

13. Wells, J.N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet* **54**, 539-561 (2020).

14. de Koning, A.P., Gu, W., Castoe, T.A., Batzer, M.A. & Pollock, D.D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* **7**, e1002384 (2011).

15. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

16. Walsh, C.P., Chaillet, J.R. & Bestor, T.H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* **20**, 116-7 (1998).

17. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol* **13**, R107 (2012).

18. Chuong, E.B., Elde, N.C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083-7 (2016).

19. Hancks, D.C. & Kazazian, H.H., Jr. Roles for retrotransposon insertions in human disease. *Mob DNA* **7**, 9 (2016).

20. Stower, H. Alternative splicing: Regulating Alu element 'exonization'. *Nat Rev Genet* **14**, 152-3 (2013).

21. Schmitz, J. & Brosius, J. Exonization of transposed elements: A challenge and opportunity for evolution. *Biochimie* **93**, 1928-34 (2011).

22. Attig, J. *et al.* Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA Processing. *Cell* **174**, 1067-1081 e17 (2018).

23. Zarnack, K. *et al.* Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152**, 453-66 (2013).

24. Zahid, S. *et al.* The Multifaceted Roles of Ku70/80. *Int J Mol Sci* **22**(2021).

25. Kaczmarski, W. & Khan, S.A. Lupus autoantigen Ku protein binds HIV-1 TAR RNA in vitro. *Biochem Biophys Res Commun* **196**, 935-42 (1993).

26. Peterson, S.E. *et al.* The function of a stem-loop in telomerase RNA is linked to the DNA repair protein Ku. *Nat Genet* **27**, 64-7 (2001).

27. Dalby, A.B., Goodrich, K.J., Pfingsten, J.S. & Cech, T.R. RNA recognition by the DNA end-binding Ku heterodimer. *RNA* **19**, 841-51 (2013).

28. Ting, N.S., Yu, Y., Pohorelic, B., Lees-Miller, S.P. & Beattie, T.L. Human Ku70/80 interacts directly with hTR, the RNA component of human telomerase. *Nucleic Acids Res* **33**, 2090-8 (2005).

29. Lamaa, A. *et al.* A novel cytoprotective function for the DNA repair protein Ku in regulating p53 mRNA translation and function. *EMBO Rep* **17**, 508-18 (2016).

30. Shadrina, O. *et al.* Analysis of RNA binding properties of human Ku protein reveals its interactions with 7SK snRNA and protein components of 7SK snRNP complex. *Biochimie* **171-172**, 110-123 (2020).

31. Unfried, J.P. *et al.* Long Noncoding RNA NIHCOLE Promotes Ligation Efficiency of DNA Double-Strand Breaks in Hepatocellular Carcinoma. *Cancer Res* **81**, 4910-4925 (2021).

32. Li, S. & Mason, C.E. The pivotal regulatory landscape of RNA modifications. *Annu Rev Genomics Hum Genet* **15**, 127-50 (2014).

33. Song, C.X., Yi, C. & He, C. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nat Biotechnol* **30**, 1107-16 (2012).

34. Meyer, K.D. & Jaffrey, S.R. The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nat Rev Mol Cell Biol* **15**, 313-26 (2014).

35. Sun, W.J. *et al.* RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. *Nucleic Acids Res* **44**, D259-65 (2016).

36. Nishikura, K. A-to-I editing of coding and non-coding RNAs by ADARs. *Nat Rev Mol Cell Biol* **17**, 83-96 (2016).

37. Savva, Y.A., Rieder, L.E. & Reenan, R.A. The ADAR protein family. *Genome Biol* **13**, 252 (2012).

38. Kim, Y.K. & Maquat, L.E. UPFront and center in RNA decay: UPF1 in nonsense-mediated mRNA decay and beyond. *RNA* **25**, 407-422 (2019).

Chapter 2 Effects of RBP-TE interaction on TE exonization

2.1 Introduction

Transposable elements (TEs), also known as transposons, account for virtually 50% of the human genome, providing a substantial source of genetic variation, gene regulation, and genome evolution [1-4]. Despite their massive genomic proportion, only a tiny fraction (<0.05%) of TE elements remain the ability of mobilization. The activities of most TEs are reduced by accumulated mutations and/or repressive epigenetic modifications[5]. A minor fraction of TE insertions has evolved to become functional units of the host genome[6]; however, most TE insertions are usually detrimental[7]. Aberrant activation of these TEs can cause genetic disease or tumor[7].

Intragenic transposable elements are inevitably incorporated into the nascent transcripts of underlying resident genes. These elements are mostly in introns, so are usually spliced out from mature RNAs. However, TEs are replete with splice sites [8,9]. In some cases, these splice sites can be activated, thereby intronic TEs can be kept in the mature RNAs as alternatively spliced-in exons through a process referred to as 'exonization' [10,11]. In this way, certain human Alu elements have evolved to *bona fide* exons and are translated as part of encoded proteins [12-14]. Indeed, 2.7% and 3% of Ensembl-annotated exons in the human genome are derived from exonization of L1/L2 elements and Alu elements, respectively, greatly contributing to genome diversity.

To be fair, most of newly incorporated TE exons have undesirable consequences for the host. For example, antisense Alu elements derived exons usually introduce premature stop codons, generating aberrant proteins or causing resident mRNA degradation[15]. Recent studies indicates that RNA-binding proteins (RBPs) play an essential role in preventing transposable elements from undesired exonization. hnRNPC, as a global repressor for Alu exonization, represses exonization of antisense Alu elements by binding to the U-tract motif in the proximal upstream of Alu-encoded 3' splice sites and blocking U2AF65 binding [16]. MATR3 promotes binding of PTBP1 to multivalent binding sites within LINEs to hinder exonization of L1/L2 [17]. RBP DHX9 inhibits aberrant RNA processing by binding to dsRNAs formed by inverted repeat Alu elements[18].

Despite a few RBPs have been implicated in TE repression, systematic large-scale exploration of the impact of RBP binding to TEs has not been fully addressed for at least three reasons. While various assays have been proposed to profile RBP binding sites, including RIP-seq [19-21], HITS-CLIP [22], PAR-CLIP [23], iCLIP [24], and eCLIP [25], no approach has reached consensus in the community. In addition, large-scale RBP-RNA interaction datasets and differential RNA-seq datasets for most RBPs were not available. Finally, TEs are highly repetitive and thus lowly mappable for short reads (i.e. 30-50bp) of CLIP-seq, which makes inspection of the association between RBPs and TEs very challenging.

Hendrickson *et.al,* were first to conduct a large-scale investigation of RBP-TE interactions; they surveyed 75 CLIP-Seq datasets for 51 RBPs from 31 studies [26]. The heterogeneity of the data

sources complicates the analysis, and the lack of functional data for most investigated RBPs hindered the full interpretation of TE-derived binding events.

With recent efforts by ENCODE consortium, large-scale RBP datasets are generated [27], which comprises both RBP binding profiles (eCLIP-seq) and differential expression (RNA-seq). Thus, it is possible now to explore the questions, in large-scale, of what RBPs are interacting TEs and how these interaction affects TE activities. A recent such effort examined 223 eCLIP-seq datasets from ENCODE consortium [28]. However, their TE-centric approach may miss RBPs that selectively bind to a defined subset of elements of a certain TE family. Besides, their study did not thoroughly examine the effects of RBP-TE interaction on TE exonization.

In our study, we developed an RBP-centric framework, which utilizes both uniquely mapped reads and multi-mapped reads, to better interrogate the interaction between RBPs and TEs. We applied this framework to the 223 ENCODE eCLIP-seq datasets, which profiled 151 RBPs in either K562 cells or HepG2 cells. We revealed the extensive binding of RBPs on three major TE families: L1, L2 and Alu.

Furthermore, by integrative analysis with 752 differential RNA-seq datasets, we uncovered implications of these RBP-TE interplay in TE exonization. We identified two novel repressors of TE exonization: HNRNPM repressed exonization of antisense L1 elements, while XRCC6(Ku70) repressed exonization of antisense Alu elements. It is noteworthy that XRCC6 is previously known to play an essential role in recognizing DNA damage break points [29]. Our analysis identifies its new function in the world of RNA regulation.

2.2 Results

2.2.1 An analytical framework to recover RBP binding events on repetitive elements by leveraging multi-mapped reads

To better understand the binding modes of RNA binding proteins on TEs throughout the transcriptome, we developed a unified eCLIP-seq analysis workflow to identify TE-derived RBP binding sites by utilizing both uniquely mapped reads and multi-mapped reads (**Methods**) (**Figure 2.1B**). In brief, raw reads from eCLIP-seq are trimmed for adaptors, mapped against the human genome (hg19 assembly), and removed if located within rRNA regions. PCR duplicates are collapsed. Afterwards, the RBP binding sites are determined by CLAM [30], a peak calling tool previously developed in our lab, which rescues peaks on low-complexity regions by allocating multi-mapped reads probabilistically to their possible original loci. For each RBP dataset, eCLIP replicates are processed separately until the step of peak calling. An IDR approach was used to obtain reproducible binding sites across replicates [31]. A total of 223 eCLIP-seq datasets, profiling 151 RBPs in either K562 cells or HepG2 cells from the ENCODE consortium, were processed in this framework.

We found that incorporation of multi-mapped reads in the workflow greatly improved identification of RBP binding sites on transposable elements. To illustrate this point, we compared the difference in recognition of TE-derived peaks between using uniquely mapped reads alone and using both uniquely mapped reads and multi-mapped reads. We define ‘rescued peaks’ as new peaks identified using the combination of multi-mapped reads and uniquely mapped reads, as compared to peaks identified with uniquely mapped reads alone. ‘Common peaks’ are the peaks identified regardless of the inclusion of multi-mapped reads. In 83% of

datasets (186/223), we observed an increased fraction of TE-derived peaks in 'rescued peaks' relative to 'common peaks' (**Figure 2.1B**). For 21 datasets, the fractional increment was larger than 10% (**Figure 2.1B)**. Improved HNRNPC eCLIP signals over the intronic antisense Alu elements in NUP50 (**Figure 2.1C**) and MATR3 eCLIP signals over antisense L1 elements in DUXAP8 (**Figure 2.1D**) are shown.

2.2.2 Extensive binding of RNA-binding proteins on transposable elements in transcribed regions

A large percentage of RBP binding sites were found in transposable elements, maximizing at 80% of EXOSC5 binding sites (in K562 cells) and 75% of HNRNPC binding sites (in HepG2 cells) (**Figure 2.1E**). In general, TE-derived RBP binding sites were mostly composed of LINE class and SINE class, in agreement with the large copy number of these two classes in the human genome. The percentages of RBP binding sites within the four major TE classes are also varied across different RBPs.

Since the library of eCLIP-seq is strand-specific, we sought to portray a finer interaction landscape between RBPs and oriented TE families while correcting for binding distribution of individual RBPs on genic features (**Methods**). For each eCLIP-seq peak set, we calculated the distribution of all binding sites among the following genic features, in order of priority: 3'UTR> 5'UTR>CDS> lincRNA exon>proximal intron>distal intron. We created a control peak set by randomly selecting an equal number of sites from unbound transcribed areas as there are in the eCLIP peak set while preserving peak distribution over the six genic features. 100 such control sets were independently drawn. We use a binomial test to evaluate whether the fraction of

binding sites overlapping a certain oriented TE family from the eCLIP peak set is larger than the expected fraction as averaged across the 100 control sets. This process is repeated for each eCLIP peak set separately.

TE families in significant RBP-TE pairs were composed of L1, L2, Alu, MIR, and ERV1, and particularly of the first three TE families (**Figure 2.2A and Figure 2.2B**). Of 152 profiled RBPs, 49 RBPs were found to significantly interact with at least one TE family in one cell line (K562 and/or HepG2). These results highlight the prevalent interplay between RBPs and TEs.

One previous analysis applied a TE-centric approach to characterize RBP-TE interaction, based on the same eCLIP datasets, by comparing the number of eCLIP reads and input reads mapped to individual TE families. This approach favors TE families with a limited number of copies and preferentially recognizes RBPs that act as global binders for a certain TE family. In contrast, our RBP-centric method is designed to also emphasize RBPs that display selective binding on a defined subset of elements of a specific TE family.

Applying our new approach, we identified several novel Alu-interacting RBPs, including PTBP1, UPF1, AKAP1, XRCC6 (Ku70), SAFB, FAM120A, XPO5, and DDX59 (**Figure 2.2A and Figure 2.2B**), which were absent from the previous analysis. The PTBP1 binding motif (**Figure 2.2D**) is found on the U-tracts of the antisense Alu consensus sequence (**Figure 2.2E**), supporting its direct interaction with antisense Alu. UPF1 is required for Staufen1-mediated mRNA decay. Staufen1 recognizes RNA stem structures formed by a pair of partially complementary Alu elements, and then UPF1 is recruited to trigger RNA degradation [32]. Thus, it

is reasonable to speculate that the interaction between Alu and Staufen1 results in the identified interaction between Alu and UPF1 by our analysis.

Further functional analysis support some of these novel RBPs are Alu binders. For example, our analysis demonstrated that, if UPF1 or AKAP1 interacts with Alu elements within 3'UTRs, the expression level of the corresponding genes will be altered upon UPF1 or AKAP1 depletion, respectively (**Figure 3.2)**. The binding of XRCC6 on intronic antisense Alu well correlates with its repression of antisense Alu exonization (see below). Taken together, our RBP-centric analysis uncovered functionally important RBP-TE interactions.

L1 and L2, the two main families of LINE elements, appear to interact with the largest number of RNA-binding proteins (**Figure 2.2A and Figure 2.2B**), consistent with the notion that transcribed LINE can serve as RNP assembly [17]. L1/L2-derived RBP binding sites are mainly located in distal introns (**Figure 2.2A and Figure 2.2B**). While distal sites can regulate splicing of nearby exons through long-range RNA-RNA secondary structures [33], splicing events are mainly regulated by sequences in close vicinity of splice sites [34,35]. Therefore, the following analysis mainly examine effects of these TE-derived binding sites on exonization of the TEs themselves rather than on the regulation of splicing of nearby established exons.

2.2.3 TE sequence specifies interaction between TEs and RBPs

We evaluated the sequence specificity of TEs for RBP binding. Unlike transcription factor motifs, RBP motifs are typically short and degenerative [36]. Some RBPs recognize specific RNA secondary structures rather than RNA sequences. With these in mind, we only considered RBPs

whose binding specificities were primarily based on RNA sequences. We created a motif database for the TE-interacting RBPs identified above based on information from various databases and experimental assays [36-39]. The curated motif database excludes RBPs without established motifs, RBPs mainly recognizing structured elements instead of sequence motifs, and RBPs whose motifs are ambiguous from different sources. Only one motif was retained for each RBP. The curated motifs are shown in S**upplemental Table 1**.

We calculated the motif frequency of individual RBPs in each TE family. For most RBPs, the motif frequencies in TE families with which they interact are higher than the average frequency of the motif over all TE families (**Figure 2.2C**). This observation suggests that the sequence composition of TEs inherently encodes RBP binding. For instance, our eCLIP-seq analysis shows that LIN28B interacts with sense Alu, whereas hnRNPC and PTBP1 interact with antisense Alu (**Figure 2.2A and Figure 2.2B**). Consistently, their motifs were observed in the corresponding orientation in the Alu consensus sequence (**Figure 2.2D and Figure 2.2E**).

Five RBPs exhibit dual eCLIP enrichment for a certain TE family: HNRNPL and CSTF2 for both sense L1 and antisense L1; ILF3, UPF1, and AKAP1 for both sense Alu and antisense Alu (**Figure 2.2A and Figure 2.2B**). Of these, HNRNPL and CSTF2 have established motif. Unexpectedly, their motif enrichments are seen in only one L1 orientation (**Figure 2.2C)**. This could be due to the simplified assessment of RBP binding specificity by motif alone [40,41].

2.2.4 Identification of regulators for TE exonization by a large-scale screening analysis based on RNA-seq

TEs are replete with splice sites [8,9], which contribute to their ability to enter the mature RNA they occupy. Yet, TEs are more frequently found in introns, relative to exons, suggesting that most intragenic TE splice sites are actually suppressed. Intronic TE splice sites are considered an important source of cryptic splicing signals, whose activation give rise to cryptic exons.

RBPs play critical roles in repressing inappropriate exonization of TEs. MATR3 represses exonization of gene-embedded antisense L1 elements; HNRNPC represses exonization of antisense Alu elements. By combining eCLIP-seq and RNA-seq data, we aim to identify additional RBPs which can suppress exonization of TEs, with a focus on antisense L1 and antisense Alu elements.

To this end, we performed an RNA-seq based screening strategy to identify regulators of TE exonization by searching for genes whose depletion could activate or repress TE exonization. For each of the 752 differential RNA-seq datasets in ENCODE database, differential splicing events were determined by running rMATs-turbo with novel splicing sites detection option [42] (**Methods**). We combined RNA-seq replicates to enhance statistical power. Splicing events with an average read count of included junctions and skipped junctions >= 10 in both of combined knockdown (KD) samples and combined control (Ctrl) samples were retained for downstream analysis.

Exons were considered differentially spliced if they satisfied the following criteria: 1) P-value <= 0.05 and FDR <= 0.1; and 2) Change in percent-spliced-in (PSI) >= 0.1 in KD vs Ctrl for included exons and <= -0.1 in KD vs Ctrl for excluded exons. This relatively stringent PSI cutoff may underestimate the effects of gene perturbation on splicing, but rather ensures the faithful detection of differential splicing events.

Control exons were defined as follows: 1) FDR >= 0.5; 2) the absolute value of changes in PSI is less than 0.05; 3) Neither constitutively spliced-in exons (exons whose PSI in KD and Ctrl conditions are larger than 0.9) nor constitutively spliced-out exons (exons whose PSI in KD and Ctrl conditions are less than 0.1).

For each RNA-seq dataset, included exons and excluded exons were separately compared with control exons in the percentage of TE-derived exons to determine putative regulators of TE exonization (P-value cutoff = 0.05). Specifically, we identified 23 genes as repressors of antisense L1 exonization and 26 genes as repressors of antisense Alu exonization in K562 and/or HepG2 cells (**Figure 2.3A, Figure 2.3B, Figure 2.4A** and **Figure 2.4B**) (**Table 2.1** and **Table 2.2**). As expected, we discovered MATR3 as a repressor for antisense L1 exonization and HNRNPC as a repressor for antisense Alu exonization.

In contrast, we only found one gene as activator of antisense L1 exonization and two genes as activators of antisense Alu exonization, respectively (**Table 2.2** and **Table 2.4**). These data suggests that more genes can repress, rather than activate, exonization of antisense L1 and antisense Alu.

2.2.5 HNRNPM suppresses antisense L1 exonization

We grouped interactors of antisense L1 elements or repressors of antisense L1 exonization into three categories as denoted in **Figure 2.3A** and **Figure 2.3B**: 1) Proteins that significantly bind antisense L1 elements in eCLIP; 2) Proteins that do not bind antisense L1 elements in eCLIP-seq, but repress exonization of antisense L1; and 3) Proteins that repress antisense L1 exonization, but lack available eCLIP-seq from ENCODE.

Among RBPs interacting with antisense L1 elements, HNRNPM and MATR3 significantly repress exonization of antisense L1 (**Figure 2.3A** and **Figure 2.3B**). This result confirms a previous study showing MATR3 hinders exonization of intronic antisense L1 elements. Besides HNRNPM and MATR3, the remaining antisense L1 interactors do not seem to affect antisense L1 exonization (including two strong interactor-EXOSC5 and SUGP2) or affect antisense L1 exonization inconsistently across different KD/KO datasets (such as PTBP1 and TARDBP) (**Figure 2.3A** and **Figure 2.3B**).

It is established that HNRNPM can regulate alternative splicing [43,44]. However, HNRNPM's involvement in repression of antisense L1 exonization has yet to be fully characterized. The integrative RBPmap analysis reveals the enhanced binding of hnRNPM up to 2kb around splice sites of hnRNPM-repressed antisense L1-derived exons, as compared to control antisense L1-derived exons (**Figure 2.3C** and **Supplementary Figure 2.6**) (**Methods**). In accord, a significantly larger fraction of hnRNPM-repressed antisense L1 exons is actually bound by hnRNPM in eCLIP-seq, as compared to control antisense L1 exons (**Figure 2.3D**).

In addition, we found most of hnRNPM-repressed antisense L1 exons are cryptic exons (**Figure 2.3E**). A representative antisense L1-derived exon in SPATA7 is shown in **Figure 2.3F**. Collectively, our analysis demonstrates that hnRNPM can bind to antisense L1s and repress their exonization.

2.2.6 XRCC6 suppresses antisense Alu exonization

Using the same strategy as for L1 (above), we next sought to identify regulators that repress exonization of antisense Alu elements. Of the several RBPs which significantly interact with antisense Alu, we identified hnRNPC and XRCC6 (Ku70) as two major repressors of antisense Alu exonization (**Figure 2.4A** and **Figure 2.4B**).

hnRNPC has been demonstrated as a global repressor of exonization of antisense Alu elements through its high-affinity binding to uridine tracts (U-tracts) at the beginning and in the middle of antisense Alu elements (**Figure 2.2D and Figure 2.2E**) [16]. Consistent with its function as a general repressor, our analysis of differential RNA-seq identifies more than 3000 HNRNPC-repressed antisense Alu exons. In contrast, XRCC6 (Ku70) can repress a smaller number of antisense Alu exons (around 1000 Alu exons), suggesting the selective repression of Alu exonization by XRCC6.

XRCC6(Ku70) is primarily known as a DNA-binding protein involved in the DNA repair pathway. XRCC6 can dimerize with XRCC5 (Ku80) to form a protein complex Ku86 (a.k.a Ku complex). The Ku complex plays an essential role in nonhomologous end-joining repair by recognizing double-strand breaks. In addition to its DNA binding ability, the complex can also

directly bind RNA molecules. Kaczmarski W and colleagues first reported the RNA binding ability of the Ku complex by verifying its direct binding to HIV-1 transcripts through the recognition of an RNA harpin structure formed in the transactivation responsive (TAR) element in HIV virus[45]. Later, either the Ku protein or Ku70 alone were documented to interact with a variety of RNA molecules, such as yeast telomerase RNA TLC1[46,47], human telomerase RNA hTR[48], p53 mRNA[49], 7SK snRNA[50], LncRNA NIHCOLE[51] and some synthetic RNAs[50].

The repression of Alu exonization by XRCC6 well correlates with its binding to antisense Alu elements, as revealed by the strengthened XRCC6 occupancy in the close vicinity of 3' splice sites of included antisense Alu exons upon XRCC6 KD/KO (**Figure 2.4C** and **Supplementary Figure 2.8**). Consistently, a two-fold higher fraction of XRCC6-repressed antisense Alu exons was bound by XRCC6, as compared to control antisense Alu exons (**Figure 2.4D**). Moreover, the majority of XRCC6-repressed antisense Alu exons were cryptic exons (**Figure 2.4E**), indicating XRCC6 prevents Alu elements from aberrant exonization, to safeguard the transcriptome.

We carry out RT-PCR assays to validate 20 XRCC6-repressed antisense Alu exons determined by the above analysis of differential RNA-seq, including 14 cryptic Alu exons and 6 Ensembl-annotated Alu exons (**Methods**) (**Table 2.5** and **Table 2.6**). The signal tracks of XRCC6 eCLIP-seq and differential RNA-seq upon XRCC6 KD/KO for Alu exons in EXOSC9 (**Figure 2.5A**) and MDM2 are shown (**Figure 2.5B**). RT-PCR confirmed that inclusion ratios of all these 20 Alu exons in K562 cell were increased upon XRCC6 depletion (**Figure 2.5C, Figure 2.5D** and

Supplementary Figure 2.10). Knocking down efficiency are confirmed (**Supplementary Figure 2.9)**.

hnRNPC is the global repressor for antisense Alu exonization, so we next ask if hnRNPC also contributes to the repression of the 20 Alu exons. To this end, hnRNPC and XRCC6/hnRNPC (double knockdown) were depleted by siRNAs in K562 cells as well (**Methods**). Interestingly, most of the 20 Alu exons are included to some extent after hnRNPC knockdown and even more included when XRCC6 and hnRNPC were depleted simultaneously (**Figure 2.5C, Figure 2.5D** and **Supplementary Figure 2.10**). Knocking down efficiency are confirmed (**Supplementary Figure 2.9)**.

We then examined effects of hnRNPC depletion on all XRCC6-repressed antisense Alu exons identified by differential RNA-seq (**Figure 2.6E**). We found that hnRNPC depletion can de-repress large numbers of XRCC6-repressed Alu exons. On the contrary, XRCC6 depletion de-represses a subset of hnRNPC-repressed Alu exons. These analyses are consistent with the notion of HNRNPC functioning as a global repressor; in contrast, XRCC6 appears to be a more selective repressor.

There are two uridine tracts (U-tracts) on antisense Alu elements, which can serve as polypyrimidine tracts to facilitate activation of downstream cryptic 3' splice sites within Alu[52] (**Figure 2.2E, Supplementary Figure 2.11** and **Supplementary Figure 2.12**). HNRNPC fulfils its broad repression of Alu exonization via binding to the U-tracts[16]. The continuity of U-tract matters to HNRNPC binding, where HNRNPC binding strength on RNA decreases when the

length of continuous U-tracts shortens [16]. Thereby, mutations in U-tracts of Alu elements can compromise the protection of HNRNPC against Alu exonization. In accord with this, we observe an upward shift in the length of continuous U-tracts upstream of 3'ss of HNRNPC-repressed antisense Alu exons, as compared to control Alu exons (**Figure 2.5F**). On the other hand, antisense Alu exons repressed by XRCC6 do not prefer to have longer U-tract upstream of their 3'ss (**Figure 2.5F**).

We then ask if XRCC6 can complement HNRNPC in repressing Alu exons whose Alu-encoded 3'ss have shorter continuous upstream U-tracts (**Figure 2.5G**). To this end, we categorize included antisense Alu exons upon either HNRNPC depletion or XRCC6 depletion into groups with various length of longest continuous U-tract upstream of 3'SS. Each group is further dichotomized into exons more sensitive to XRCC6 depletion (Delta PSI in XRCC6 KD/KO vs Ctrl is larger than Delta PSI in HNRNPC KD/KO vs Ctrl) and exons more sensitive to HNRNPC depletion (Delta PSI in XRCC6 KD/KO vs Ctrl is less than Delta PSI in HNRNPC KD/KO vs Ctrl). We found that the fraction of Alu exons that are more sensitive to XRCC6 depletion becomes elevated as the continuous U-tract upstream of 3'SS becomes shorter (**Figure 2.5G**). This indicates that XRCC6 can provide additional safeguard against exonization of antisense Alu elements which are accompanied by shorter continuous U-tract in proximal upstream of Alu-encoded 3'SS.

2.3 Discussion

The repetitive nature of transposable elements, combined with the short length of CLIP-seq reads, hinders the characterization of RBP-TE association. To overcome these obstacles, we

developed an analytical framework to recover RBP binding on TEs by incorporating multi-mapped reads (**Figure 2.1A**). We found that the fractions of TE-derived peaks within 'rescued peaks' were much higher than within 'common peaks' for most datasets (**Figure 2.1B**). This demonstrates the improved peak calling on TE elements by our framework. Application of this new framework to ENCODE datasets allows us to generate a comprehensive interaction map between RNA-binding proteins (RBPs) and transposable elements (TEs).

With our RBP-centric strategy, we found that the profiled RBPs preferably interact with three TE families: L1, L2, and Alu (**Figure 2.2A** and **Figure 2.2B**). These three TE families are in high prevalence in transcribed regions (i.e., L1: 442,357 copies, L2: 254,283 copies, and Alu: 655,839 copies). On the other hand, these RBPs do not interact with other TE families that are also in high prevalence (e.g., hAT-Charlie: 142439 copies, ERVL-MaLR: 140477 copies and MIR: 333948 copies). Therefore, it seems that certain TE families probably have an inherent ability to interact with certain RBPs, regardless of the TE prevalence. In support of this notion, we found that RBP motif frequencies over the interacting TE families were higher than the overall frequency of that RBP motif over all TE families (**Figure 2.2C**). Besides RNA sequence motifs, RNA secondary structure can also shape RBP binding. It is therefore reasonable to further investigate whether transcribed TEs have structural elements that contribute to RBP recruitment.

RBPs have been implicated in the repression of exonization of TE elements. Our analysis recapitulates MATR3 as repressor of antisense L1 exonization (**Figure 2.3A** and **Figure 2.3B** and **Supplementary Figure 2.7**) and hnRNPC as repressor of antisense Alu exonization (**Figure**

2.4A and **Figure 2.4B**). Futhermore, we identified two novel repressors of TE exonization: hnRNPM and XRCC6 (**Figure 2.3A, Figure 2.3B, Figure 2.4A** and **Figure 2.4B**).

hnRNPM is a well-known repressor of alternative splicing. The physical interaction between hnRNPM and MATR3 has been described in mammalian cells [53]. hnRNPM and MATR3 are components of LASR, a multimeric assembly of splicing regulators. Here we identified a new role of hnRNPM in repressing exonization of gene embedded antisense L1 elements.

XRCC6 (Ku70), a subunit of Ku70–Ku80 complex, is engaged in the first step of the non-homologous end joining (NHEJ) pathway of DNA repair, i.e., recognition of double-strand breaks in DNA. Increasing evidence indicates that XRCC6(Ku70) alone or the Ku70-Ku80 complex can bind RNA. The successful identification of thousands of XRCC6 binding sites determined by eCLIP-seq also confirms the affinity of XRCC6 to RNA molecules. By integrating eCLIP-seq and RNA-seq data, we demonstrated that XRCC6 preferentially bind to gene-embedded antisense Alu elements and repress their exonization. Of note, all observations for XRCC6 occur in both K562 and HepG2 cells.

XRCC6 appears to interact with its RNA targets through recognition of specific hairpin structures, rather than specific RNA sequence [54]. The affinity of XRCC6 to hairpin structures is also supported by the *de novo* discovery of structural motifs in XRCC6 binding sites. Hairpin structures were consistently enriched in eCLIP-seq peaks relative to flanking regions in both K562 cells and HepG2 cells (**Supplementary Figure 2.11**).

Future studies are needed for delineating the functional significance of RBP-TE interaction we have uncovered in the current study. For instance, we observed that a very large fraction of EXOSC5 and SUGP2 binding sites overlap antisense L1 elements; however, depletion of those RBPs did not significantly de-repress exonization of antisense L1. A multitude of RBPs interact with L2 TEs, the effects of these interaction on L2 elements await for further investigation [55].

2.4 Methods

2.4.1 Cell culture

Human K562 cells (ATCC #CCL-243) were maintained at ~7 x 10^5 cells at 37 °C in a humidified 5 % CO_2 atmosphere in RPMI 1640 (Gibco), 10% fetal bovine serum (Corning), 1% penicillin/streptomycin (Gibco), and 1% Glutamax (Gibco), all of which were filtered through a 0.2μm PES membrane sterile filter (Nalgene rapid-flow). K562 cells were used at passages 7 to 14. Genomic DNA of the K562 cells was authenticated by STR profiling. Cells were confirmed to be mycoplasma-free by the Lonza MycoAlert assay.

2.4.2 siRNA transfection

K562 cells were counted on the Vi-Cell XR Cell Viability Analyzer (Beckman) and 1.0 x 10^6 cells were used for each transfection. Control siRNAs (4390843 and 4390846), siRNAs to XRCC6 (s5456 and s5457), and siRNAs to hnRNPC (s6719 and s6721) were purchased from ThermoFisher. siRNAs were transfected into K562 cells by 4D-Nucleofector (Lonza, Kit L) according to the manufacturer's instructions, at a final concentration of 8nM. The double knockdowns of XRCC6 and hnRNPC were performed by combining one XRCC6 siRNA with one hnRNPC siRNA. The transfected K562 cells were collected after 3 days and re-transfected

for another 3 days. At the end of day 6, the transfected cells were collected for RNA and protein preparation.

2.4.3 RNA extraction

K562 cells were spun down at 200 g and lysed using 1ml TRIzol (Invitrogen) according to manufacturer's instructions. 200µl of chloroform was added directly to the microcentrifuge tubes containing TRIzol-suspended cells. The tubes were briefly shaken and incubated at room temperature for 3-5 minutes. Then they were spun at 14,000 g at 4°C (Eppendorf) for 15 minutes. The aqueous layer was manually extracted and diluted in a 1:1 mixture with cold isopropanol (Fisher). The mixture was then incubated at room temperature for 10 minutes and was spun at 14,000 g at 4°C for 10 minutes. Next, the RNA pellets were washed once with freshly made 75% ethanol. The pellets were dried and re-suspended in RNA storage solution (ThermoFisher). RNA quality was assessed on the Agilent Tapestation 4200 (Agilent Technologies), RNA samples with an RNA integrity number >9.8 were retained for further downstream analysis.

2.4.4 qRT-PCR for measuring gene expression

Trizol-extracted total RNAs were treated with DNase I (Invitrogen, amplification grade) and then reverse transcribed using oligo-dT, Moloney murine leukemia virus reverse transcriptase (Promega), and 1x Moloney murine leukemia virus reverse transcription (RT) buffer (Promega), according to manufacturers' instructions. After incubation at 37°C for 1 hour, the samples were used as a template for qRT-PCR with PowerTrack™ SYBR Green Master Mix (Applied Biosystems™) on the QuantStudio™ 5 Real-Time PCR System (Applied Biosystems™),

according to manufacturer's instructions. XRCC6 and hnRNPC mRNA levels were normalized to that of Actin (**Supplementary Figure 2.9**).

2.4.5 Western blot analysis

One aliquot of the transfected K562 cells were lysed in RIPA buffer (Thermo Scientific) containing protease inhibitors (Pierce). Samples were placed on ice and sonicated (Fisherbrand 705 Sonic Dismembrator) by pulsing the machine for 10 seconds. Samples were then incubated on ice for 5 minutes and centrifuged at 14,000 g for 15 minutes to pellet debris. The supernatant was transferred to a new tube and protein concentrations were determined with the BCA™ protein assay kit (Pierce). Equal amounts of protein were separated via 4-15% SDS-PAGE gels (BioRad) and transferred to polyvinylidene difluoride (PVDF) membrane using TransBlot Turbo Transfer System (BioRad). Membranes were blocked for 1 hour in 5% milk in PBS+Tween20 (Life Technologies, BioRad). Western blots were performed with these primary antibodies: anti-XRCC6 antibody (Bethyl, ab83501, at 1:10,000), anti-hnRNP C1/C2 (Santa Cruz, sc-32308, at 1:1,000), and Anti-B-actin (Sigma-Aldrich, A5441-100UL, at 1:20,000). Secondary antibodies used were Anti-Mouse (ThermoFisher Scientific, 32430, 1:2,000) and Anti-Rabbit (Agilent Technologies, P0448, 1:2,000). After antibody incubation steps, band intensity was detected via chemiluminescence with Femto (Thermo Scientific) and imaged through the ChemiDoc MP Imaging System (BioRad). (**Supplementary Figure 2.9**).

2.4.6 RT-PCR for measuring exon inclusion ratio

Trizol-extracted total RNAs were treated with DNase I (Invitrogen, amplification grade) and then reverse transcribed using oligo-dT, Moloney murine leukemia virus reverse transcriptase

(Promega), and 1x Moloney murine leukemia virus reverse transcription (RT) buffer (Promega), according to manufacturers' instructions, as above. After incubation at 37°C for 1 hour, the samples were used as a template for RT-PCR. PCR products were separated by agarose gel, visualized by ChemiDoc MP imaging system, and the intensity of corresponding bands quantified by ChemiDoc image lab software (BioRad). The percentage of PCR product with exon inclusion (upper band) over the total PCR products (upper band + lower band) was calculated. (**Supplementary Figure 2.10**).

2.4.7 Analytic framework for peak identification

Raw data of 223 eCLIP experiments were downloaded from ENCODE [27]. After two round adapter trimming with Cutadapt v2.3 [56], reads were mapped to human genome (version GRCh37, GENCODE release 40) with STAR v2.5.3 [57] with the following parameters: --outSAMtype BAM Unsorted --alignEndsProtrude 15 ConcordantPair --twopassMode Basic --limitOutSJcollapsed 2000000. Mapped reads were filtered by BEDTools v2.27.1 [58] with parameter -f 0.90 to remove reads of rRNA origin. The rRNA annotation is obtained from Table Browser [59]. Read pairs exactly mapped to the same location are regarded as PCR duplicates. PCR duplicates are collapsed.

We use CLAM to call peaks in both eCLIP replicates (input sample is shared by two eCLIP replicates) at 3 steps: preprocessing, realigning, and peak calling [30]. In the preprocessing step, parameters '--read-tagger-method start --lib-type sense' are used to separate uniquely mapped reads and multi-mapped reads. In the realigning step, parameters are set to '--winsize 50 --max-tags -1 --read-tagger-method start --lib-type sense' to assign multi-mapped reads by EM

algorithm. In the peak calling step, Q-value cutoff is set to 1 with parameters '--binsize 50 --qval-cutoff 1' to identify less stringent peaks. These peaks from two replicates are sorted based on P-value, respectively. Finally, IDR algorithm was used to obtain reproducible peaks across replicates at IDR cutoff 0.01 [31].

We also run the pipeline with uniquely mapped reads alone. For an eCLIP-seq dataset, peaks identified only when multi-mapped reads are used are considered "rescue peaks", while peaks detected regardless of whether multi-mapped reads are used are considered "common peaks".

2.4.8 Identification of significantly interacting RBP-TE pairs in eCLIP-seq

For individual RBPs, we determine their significantly interacting TE families by a binomial test that evaluates if the fraction of binding sites overlapping a TE family in a certain orientation is larger than the expected fraction. The expected fraction is calculated based on 100 randomly selected control sets as described below.

1. Each binding site in the eCLIP peak set is annotated with a genic feature. Ambiguous annotations are settled in the following order of priority: 3'UTR > 5'UTR > CDS > lincRNA exon > proximal intron > distal intron. 100 independent control sets with the same number of sites as there are in the eCLIP peak set are independently generated. Each control set comprises randomly selected sites from transcribed regions without the binding of the corresponding RBP. At the same time, each control set is required to have identical distribution of genic features as the peak set.

2. For each control peak set, we calculate the fraction of sites overlapping individual TE families; the sense TE family and the antisense TE family are calculated separately. Such fractions for the 100 control sets are further averaged to obtain the 'expected' fraction.
3. We use a binomial test to evaluate if the fraction of binding sites in the eCLIP peak set overlapping a certain orientation of a TE family is larger than expected.
4. RBP-TE pairs are filtered out if the fraction of RBP peaks in the corresponding oriented TE family is less than 5% or the number of RBP peaks in the oriented TE family is less than 10.
5. RBP-TE pairs are considered significant if P-value is <=0.01.

2.4.9 Motif frequency calculation over TE families

For each motif from our curated motif database, we use FIMO to search for its occurrences over all similarly oriented copies of a TE family [60]. Motif frequency is calculated by dividing the total number of occurrences by the total length of all copies.

2.4.10 Alu consensus sequences

Alu consensus sequence is achieved by multiple alignment of consensus sequences of various Alu subfamilies from USCS Table Browser [61].

The consensus sequences of sense Alu is

GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGG
CGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCAACACGGTGAAACCCCGTCT
CTACTAAAAATACAAAAAATTAGCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGC
TACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGT

GAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTC

AAAAAAAAAAAAAAAAAAAAAAAAAAAAA

the consensus of antisense Alu is

TTTTTTTTTTTTTTTTTTTTTTTTTTTTGAGACGGAGTCTCGCTCTGTCGCCCAGGCT

GGAGTGCAGTGGCGCGATCTCGGCTCACTGCAAGCTCCGCCTCCCGGGTTCACGCCA

TTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTACAGGCGCCCGCCACCACGCCCG

GCTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTCACCGTGTTGCCAGGATGGTCTC

GATCTCCTGACCTCGTGATCCGCCCGCCTCGGCCTCCCAAAGTGCTGGGATTACAGG

CGTGAGCCACCGCGCCCGGCC

2.4.11 Large-scale screening analysis based on RNA-seq to identify potential regulators of TE exonization

752 differential RNA-seq datasets were available from ENCODE during our investigation. Raw sequencing data of technical replicates was combined to enhance sequencing depth. The combined reads were mapped to the human hg19 assembly. We applied rMATs-turbo with the following parameters to detect and quantify alternative splicing events: -t paired --variable-read-length --novelSS --statoff --nthread 8 --task both. The novel splice site detection in rMATS-turbo is turned on to identify both Ensembl-annotated splicing events and cryptic splicing events. For simplicity, we only consider the most common alternative splicing events, i.e., exon skipping events, in our following analysis.

Exon inclusion levels were measured as PSI (Percent Spliced In), which is the percentage of junction reads supporting the exon-including isoform. We use a custom python script to identify

differential splicing events in an experiment design without replicates. Splicing events with the average of 'including junction counts' and 'skipping junction counts' >= 10 in both the combined KD/KO sample and combined control sample were retained for downstream analysis.

Exons are considered differentially spliced if they satisfy the following the criteria. **1)** P-value<=0.05 and FDR<=0.1. **2)** The change in Percent-spliced-in (PSI) >=0.1 in KD/KO vs WT for included exons and <= -0.1 in KD/KO vs WT for excluded exons. The background exons are defined as follows: **1)** FDR>=0.5. **2)** the absolute value of change in PSI is less than 0.05. **3)** Excluding exons whose PSI in both conditions are less than 0.1, which are considered constitutively spliced-out exons. **4)** Excluding exons whose PSI in both conditions are larger than 0.9, which are regarded as constitutively spliced-in exons.

For each KD/KO dataset, the group of included exons and the group of excluded exons are separately compared with the group of control exons in the fraction of exons overlapping an oriented TE family to determine putative regulators of TE exonization (P-value cutoff=0.05).

2.4.12 Determination of transposable elements bound by a certain RBP

We assume eCLIP and Input read count of a certain RBP in a transposable element follows a Poisson distribution as follows.

$$P(T) = \frac{{\lambda_{eCLIP}}^{T} \times e^{-\lambda_{eCLIP}}}{T!}$$

$$P(t) = \frac{{\lambda_{Input}}^{t} \times e^{-\lambda_{Input}}}{t!}$$

Where T is observed read count from eCLIP and t is observed read count from Input. Both T and t are pre-normalized to sequencing depth and added by pseudo-count one. λ_{eCLIP} represents the true binding affinity in eCLIP. λ_{Input} is the baseline affinity in Input.

The prior of λ_{Input} is set as an uninformative uniform distribution. Then the posterior distribution of λ_{Input} is a gamma distribution with shape t+1 and scale 1 as indicated below.

$$Post(\lambda_{nput}) \propto \frac{{\lambda_{Input}}^{t} \times e^{-\lambda_{Input}}}{t!}$$

In null hypothesis where $\lambda_{eCLIP} = \lambda_{Input}$, the posterior distribution of λ_{Input} also serves as the prior distribution of λ_{eCLIP}. Then the compound distribution of T in null hypothesis can be modeled by a binomial distribution with size t and probability 0.5. The cumulative distribution function F of the compound distribution defines p-value as below.

$$F(T) = \int_0^T \left(\frac{t}{k}\right)\left(\frac{1}{2}\right)^t dk$$

$$pvalue = 1 - F(T)$$

2.4.13 RBPmap analysis

RBP occupancy for a given group of exons at a certain position relative to 3' and 5' splice site is measured as the percentage of exons with larger signal in eCLIP than that in Input. The measurement equally weights exons to correct for exon abundance and outlier events [62].

2.4.14 The identification of structure elements in RBP binding sites.

We use pyteiser[63] to recognize structure elements overly represented in the binding sites, relative to the flanking regions of the same length. The default parameters are used in every step of pyteiser implementation.

2.4.15 Alignment of individual antisense Alu elements to the consensus

We use PRANK-F[64,65] (default parameters for noncoding DNA sequences alignment) to align each of antisense Alu elements on the genome to the consensus antisense Alu sequence.

2.5 References

1. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol* **19**, 199 (2018).

2. Chuong, E.B., Elde, N.C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**, 71-86 (2017).

3. Wells, J.N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet* **54**, 539-561 (2020).

4. Lanciano, S. & Cristofari, G. Measuring and interpreting transposable element expression. *Nat Rev Genet* **21**, 721-736 (2020).

5. Walsh, C.P., Chaillet, J.R. & Bestor, T.H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* **20**, 116-7 (1998).

6. Chuong, E.B., Elde, N.C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083-7 (2016).

7. Hancks, D.C. & Kazazian, H.H., Jr. Roles for retrotransposon insertions in human disease. *Mob DNA* **7**, 9 (2016).

8. Huff, J.T., Zilberman, D. & Roy, S.W. Mechanism for DNA transposons to generate introns on genomic scales. *Nature* **538**, 533-536 (2016).

9. Alvarez, M.E.V. *et al.* Transposon clusters as substrates for aberrant splice-site activation. *RNA Biol* **18**, 354-367 (2021).

10. Stower, H. Alternative splicing: Regulating Alu element 'exonization'. *Nat Rev Genet* **14**, 152-3 (2013).

11. Schmitz, J. & Brosius, J. Exonization of transposed elements: A challenge and opportunity for evolution. *Biochimie* **93**, 1928-34 (2011).

12. Lin, L. *et al.* Diverse splicing patterns of exonized Alu elements in human tissues. *PLoS Genet* **4**, e1000225 (2008).

13. Shen, S. *et al.* Widespread establishment and regulatory impact of Alu exons in human genes. *Proc Natl Acad Sci U S A* **108**, 2837-42 (2011).

14. Lin, L. *et al.* The contribution of Alu exons to the human proteome. *Genome Biol* **17**, 15 (2016).

15. Ule, J. Alu elements: at the crossroads between disease and evolution. *Biochem Soc Trans* **41**, 1532-5 (2013).

16. Zarnack, K. *et al.* Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152**, 453-66 (2013).

17. Attig, J. *et al.* Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA Processing. *Cell* **174**, 1067-1081 e17 (2018).

18. Aktas, T. *et al.* DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. *Nature* **544**, 115-119 (2017).

19. Lerner, M.R. & Steitz, J.A. Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proc Natl Acad Sci U S A* **76**, 5495-9 (1979).

20. Tenenbaum, S.A., Carson, C.C., Lager, P.J. & Keene, J.D. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc Natl Acad Sci U S A* **97**, 14085-90 (2000).

21. Niranjanakumari, S., Lasda, E., Brazas, R. & Garcia-Blanco, M.A. Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo. *Methods* **26**, 182-90 (2002).

22. Licatalosi, D.D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464-9 (2008).

23. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129-41 (2010).

24. Konig, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**, 909-15 (2010).

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

26. Kelley, D.R., Hendrickson, D.G., Tenen, D. & Rinn, J.L. Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biol* **15**, 537 (2014).

27. Van Nostrand, E.L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711-719 (2020).

28. Van Nostrand, E.L. *et al.* Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biol* **21**, 90 (2020).

29. Zahid, S. *et al.* The Multifaceted Roles of Ku70/80. *Int J Mol Sci* **22**(2021).

30. Zhang, Z. & Xing, Y. CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. *Nucleic Acids Res* **45**, 9260-9271 (2017).

31. Li, Q., Brown, J.B., Huang, H. & Bickel, P.J. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics* **5**, 1752-1779, 28 (2011).

32. Gong, C. & Maquat, L.E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* **470**, 284-8 (2011).

33. Lovci, M.T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* **20**, 1434-42 (2013).

34. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465**, 53-9 (2010).

35. Bao, S., Moakley, D.F. & Zhang, C. The Splicing Code Goes Deep. *Cell* **176**, 414-416 (2019).

36. Lambert, N. *et al.* RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell* **54**, 887-900 (2014).

37. Paz, I., Kosti, I., Ares, M., Jr., Cline, M. & Mandel-Gutfreund, Y. RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res* **42**, W361-7 (2014).

38. Feng, H. *et al.* Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites. *Mol Cell* **74**, 1189-1204 e6 (2019).

39. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172-7 (2013).

40. Sun, L. *et al.* Predicting dynamic cellular protein-RNA interactions by deep learning using in vivo RNA structures. *Cell Res* **31**, 495-516 (2021).

41. Jolma, A. *et al.* Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. *Genome Res* **30**, 962-973 (2020).

42. 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* **111**, E5593-601 (2014).

43. 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* **63**, 2172-8 (2003).

44. West, K.O. *et al.* The Splicing Factor hnRNP M Is a Critical Regulator of Innate Immune Gene Expression in Macrophages. *Cell Rep* **29**, 1594-1609 e5 (2019).

45. Kaczmarski, W. & Khan, S.A. Lupus autoantigen Ku protein binds HIV-1 TAR RNA in vitro. *Biochem Biophys Res Commun* **196**, 935-42 (1993).

46. Peterson, S.E. *et al.* The function of a stem-loop in telomerase RNA is linked to the DNA repair protein Ku. *Nat Genet* **27**, 64-7 (2001).

47. Dalby, A.B., Goodrich, K.J., Pfingsten, J.S. & Cech, T.R. RNA recognition by the DNA end-binding Ku heterodimer. *RNA* **19**, 841-51 (2013).

48. Ting, N.S., Yu, Y., Pohorelic, B., Lees-Miller, S.P. & Beattie, T.L. Human Ku70/80 interacts directly with hTR, the RNA component of human telomerase. *Nucleic Acids Res* **33**, 2090-8 (2005).

49. Lamaa, A. *et al.* A novel cytoprotective function for the DNA repair protein Ku in regulating p53 mRNA translation and function. *EMBO Rep* **17**, 508-18 (2016).

50. Shadrina, O. *et al.* Analysis of RNA binding properties of human Ku protein reveals its interactions with 7SK snRNA and protein components of 7SK snRNP complex. *Biochimie* **171-172**, 110-123 (2020).

51. Unfried, J.P. *et al.* Long Noncoding RNA NIHCOLE Promotes Ligation Efficiency of DNA Double-Strand Breaks in Hepatocellular Carcinoma. *Cancer Res* **81**, 4910-4925 (2021).

52. Deininger, P. Alu elements: know the SINEs. *Genome Biol* **12**, 236 (2011).

53. Ramesh, N., Kour, S., Anderson, E.N., Rajasundaram, D. & Pandey, U.B. RNA-recognition motif in Matrin-3 mediates neurodegeneration through interaction with hnRNPM. *Acta Neuropathol Commun* **8**, 138 (2020).

54. Anisenko, A.N., Knyazhanskaya, E.S., Zatsepin, T.S. & Gottikh, M.B. Human Ku70 protein binds hairpin RNA and double stranded DNA through two different sites. *Biochimie* **132**, 85-93 (2017).

55. Sela, N., Mersch, B., Hotz-Wagenblatt, A. & Ast, G. Characteristics of transposable element exonization within human and mouse. *PLoS One* **5**, e10907 (2010).

56. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3 (2011).

57. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

58. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).

59. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-6 (2004).

60. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-8 (2011).

61. Fernandes, J.D. *et al.* The UCSC repeat browser allows discovery and visualization of evolutionary conflict across repeat families. *Mob DNA* **11**, 13 (2020).

62. Yee, B.A., Pratt, G.A., Graveley, B.R., Van Nostrand, E.L. & Yeo, G.W. RBP-Maps enables robust generation of splicing regulatory maps. *RNA* **25**, 193-204 (2019).

63. Fish, L. *et al.* A prometastatic splicing program regulated by SNRPA1 interactions with structured RNA elements. *Science* **372**(2021).

64. Loytynoja, A. Phylogeny-aware alignment with PRANK. *Methods Mol Biol* **1079**, 155-70 (2014).

65. Loytynoja, A. Phylogeny-Aware Alignment with PRANK and PAGAN. *Methods Mol Biol* **2231**, 17-37 (2021).

2.6 Figures

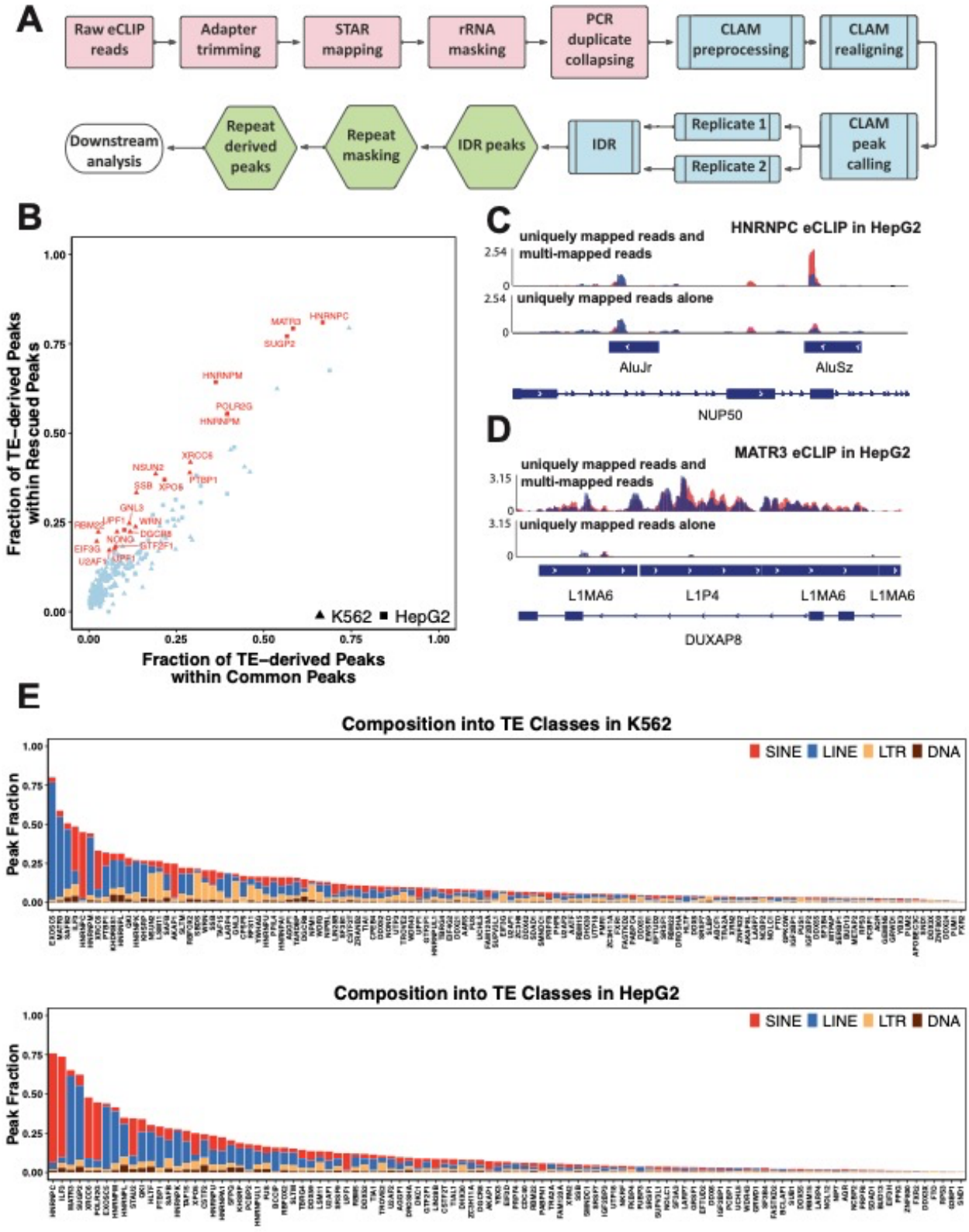

A
Raw eCLIP reads
Adapter trimming
STAR mapping
rRNA masking
PCR duplicate collapsing
CLAM preprocessing
CLAM realigning
CLAM peak calling
Replicate 1
Replicate 2
IDR
IDR peaks
Repeat masking
Repeat derived peaks
Downstream analysis
B
Fraction of TE−derived Peaks within Rescued Peaks
Fraction of TE−derived Peaks within Common Peaks
K562
HepG2
C
HNRNPC eCLIP in HepG2
uniquely mapped reads and multi-mapped reads
uniquely mapped reads alone
AluJr
AluSz
NUP50
D
MATR3 eCLIP in HepG2
uniquely mapped reads and multi-mapped reads
uniquely mapped reads alone
L1MA6
L1P4
L1MA6
L1MA6
DUXAP8
E
Composition into TE Classes in K562
SINE
LINE
LTR
DNA
Peak Fraction
Composition into TE Classes in HepG2
SINE
LINE
LTR
DNA
Peak Fraction

Figure 2.1 An analytical framework to recover RBP binding events on repetitive elements by leveraging multi-mapped reads

A. Schematic chart of eCLIP-seq processing workflow with incorporation of multi-mapped reads.

B. Comparison of the fraction of transposable element (TE)-derived peaks, between 'rescued peaks' versus 'common peaks'. eCLIP-seq datasets are highlighted in red if the increment in the fraction of TE-derived peaks is larger than 10% when including multi-mapped reads.

C. Improved eCLIP-seq signal of hnRNPC in HepG2 over one antisense Alu element (AluSz) within gene NUP50, when combining multi-mapped reads with uniquely mapped reads. eCLIP-seq replicates are colored in red and blue, respectively. eCLIP-seq signals are normalized by sequencing depth. Two Alu elements in this region are shown with arrows denoting orientation. Gene structure with exons (thick bars) and introns (narrow lines) is also displayed below.

D. Improved eCLIP-seq signal of MATR3 in HepG2 over some consecutive antisense L1 elements within gene DUXAP8, when combining multi-mapped reads with uniquely mapped reads, as in Figure 2.1C.

E. Stacked barplot showing the fraction of RBP binding sites that overlap four different TE classes. SINE, LINE, and LTR belong to the retroTE class; DNA signifies DNA TEs.

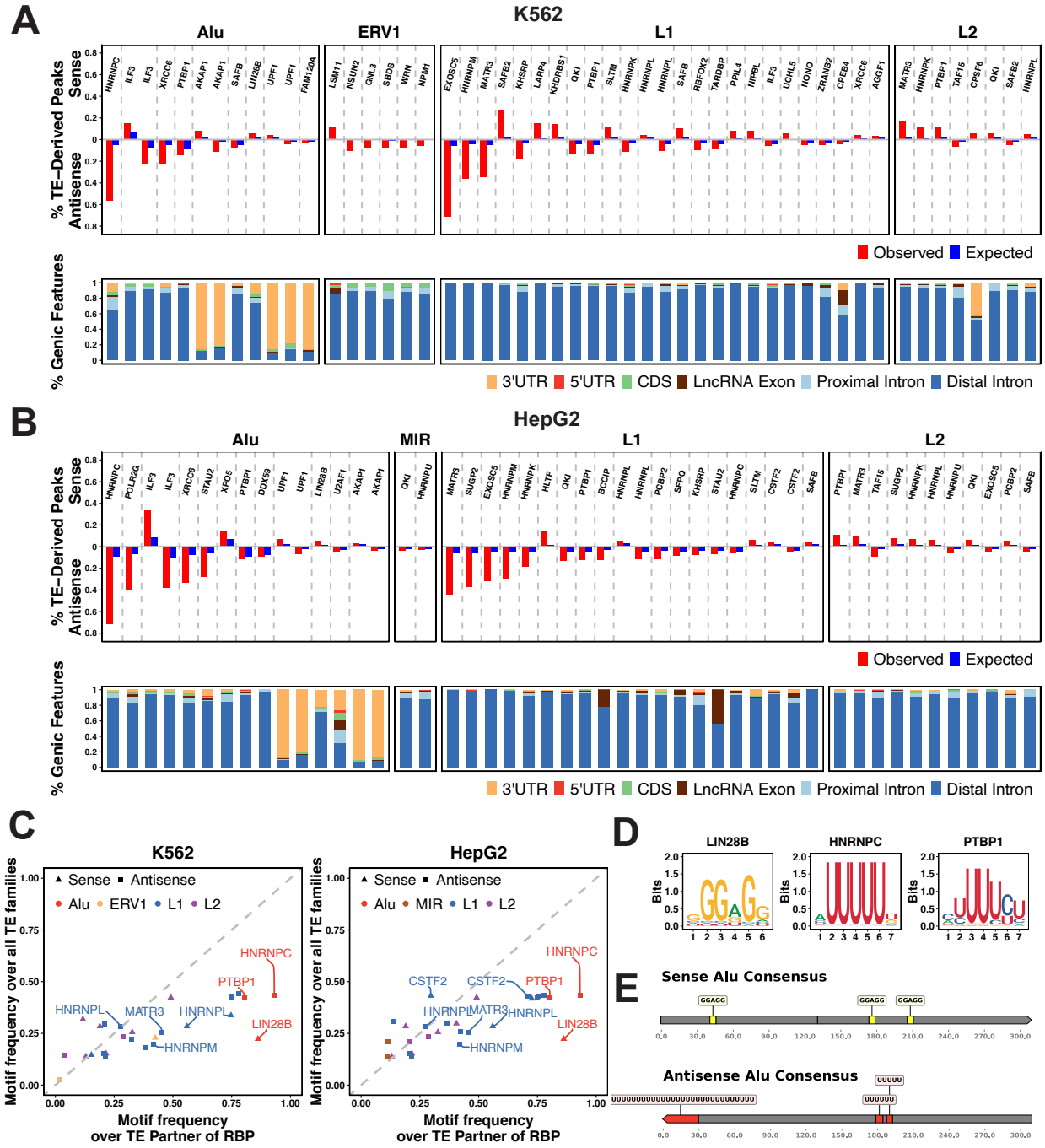


Figure 2. 2 The interaction between RNA binding proteins and transposable element families

A and B. RBPs that significantly interact with TE families in K562 cells (A) and HepG2 cells (B) determined by the analysis of eCLIP-seq binding sites. Top: The observed and expected

fractions of eCLIP-seq peaks overlapping one sense/antisense TE family. The expected fraction was calculated as described in the **Methods**. Bottom: The distribution of TE-derived binding events over different transcribed regions.

C. The motif frequency of each RBP over the TE family with which it interacts (x-axis) is plotted against the frequency of the same motif over all TE families (y-axis). Only RBPs recorded in our curated motif database are included. Left: RBP-TE interacting pairs identified in K562 cells. Right: RBP-TE interacting pairs identified in HepG2 cells. Orientation of TE families is indicated by shapes (triangle: sense; rectangle: antisense).

D Motif logo of LIN28B, hnRNPC, and PTBP1. These three RBPs bind to Alu elements in orientation specific manner according to the analysis of eCLIP-seq.

E. The binding motifs for these three RBPs are highlighted along the consensus sequences of sense Alu and antisense Alu, consistent with their actual binding on oriented Alu elements.

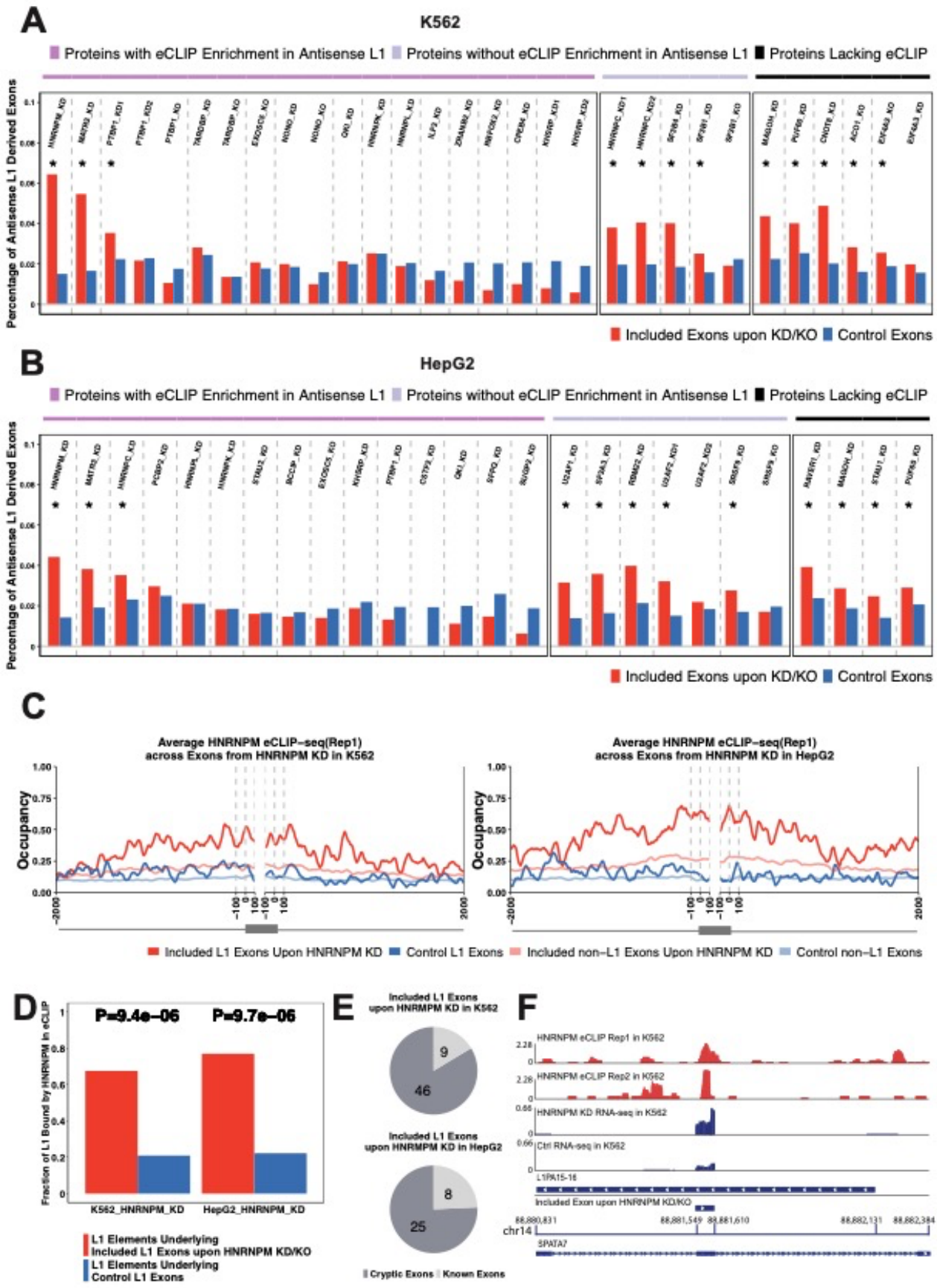

A
K562
Proteins with eCLIP Enrichment in Antisense L1
Proteins without eCLIP Enrichment in Antisense L1
Proteins Lacking eCLIP
Percentage of Antisense L1 Derived Exons
Included Exons upon KD/KO
Control Exons
B
HepG2
Proteins with eCLIP Enrichment in Antisense L1
Proteins without eCLIP Enrichment in Antisense L1
Proteins Lacking eCLIP
Percentage of Antisense L1 Derived Exons
Included Exons upon KD/KO
Control Exons
C
Average HNRNPM eCLIP−seq(Rep1)
across Exons from HNRNPM KD in K562
Average HNRNPM eCLIP−seq(Rep1)
across Exons from HNRNPM KD in HepG2
Occupancy
Included L1 Exons Upon HNRNPM KD
Control L1 Exons
Included non−L1 Exons Upon HNRNPM KD
Control non−L1 Exons
D
P=9.4e−06
P=9.7e−06
Fraction of L1 Bound by HNRNPM in eCLIP
K562_HNRNPM_KD
HepG2_HNRNPM_KD
L1 Elements Underlying
Included L1 Exons upon HNRNPM KD/KO
L1 Elements Underlying
Control L1 Exons
E
Included L1 Exons
upon HNRMPM KD in K562
9
46
Included L1 Exons
upon HNRMPM KD in HepG2
8
25
Cryptic Exons
Known Exons
F
HNRNPM eCLIP Rep1 in K562
HNRNPM eCLIP Rep2 in K562
HNRNPM KD RNA-seq in K562
Ctrl RNA-seq in K562
L1PA15-16
Included Exon upon HNRNPM KD/KO
chr14
SPATA7

Figure 2. 3 HNRNPM suppresses exonization of antisense L1 elements

A. RNA-seq based screening for repressors of antisense L1 exonization in K562 cells. The interactors of antisense L1 elements or repressors of antisense L1 exonization are categorized into three groups: 1) Proteins that bind antisense L1 elements in eCLIP; 2) Proteins that do not bind antisense L1 elements in eCLIP, but repress exonization of antisense L1; and 3) Proteins that repress antisense L1 exonization, but lack available eCLIP from ENCODE. Included exons upon RBP KD/KO and control exons are defined in **Methods**. Asterisk indicates a significantly higher fraction of antisense-L1-element-derived exons in the set of included exon upon gene depletion than in the control exon set. KD/KO datasets within each of the three categories are sorted based on p-value. RBP KD/KO datasets with the same depleted target gene are juxtaposed together.

B. RNA-seq based screening for repressors of antisense L1 exonization in HepG2 cells. The same strategy as in A.

C. RBPmap shows the high binding occupancy of HNRNPM over included antisense L1 exons upon HNRNPM depletion. HNRNPM occupancy on four different exon sets were calculated: included antisense L1 exons, control antisense L1 exons, included non-L1 exons, and control non-L1 exons. For a given exon set, HNRNPM occupancy at a certain position is measured as the percentage of exons which have larger normalized eCLIP-seq signal than normalized input signal. One eCLIP-seq replicate is used in this figure. RBPmap based on the other eCLIP-seq replicate are shown in **Supplemental Figure 2.6**.

D For included antisense L1 exons upon HNRNPM KD, a higher fraction of their underlying L1 elements is actually bound by HNRNPM, as compared to control antisense L1 exons. A Bayes

model applied to eCLIP-seq data is used to determine bound antisense L1 elements, as described in **Methods**. P-values are evaluated by one-sided proportion test.

E. The majority of included antisense L1 exons upon HNRNPM depletion are cryptic exons.

F. One antisense L1 derived exons from SPATA7 is shown. Tracks of HNRNPM eCLIP-seq (two replicates), HNRNPM KD RNA-seq, and Ctrl RNA-seq in K562 cells are displayed around an antisense L1 derived exons in SPATA7 that are included after HNRNPM KD.

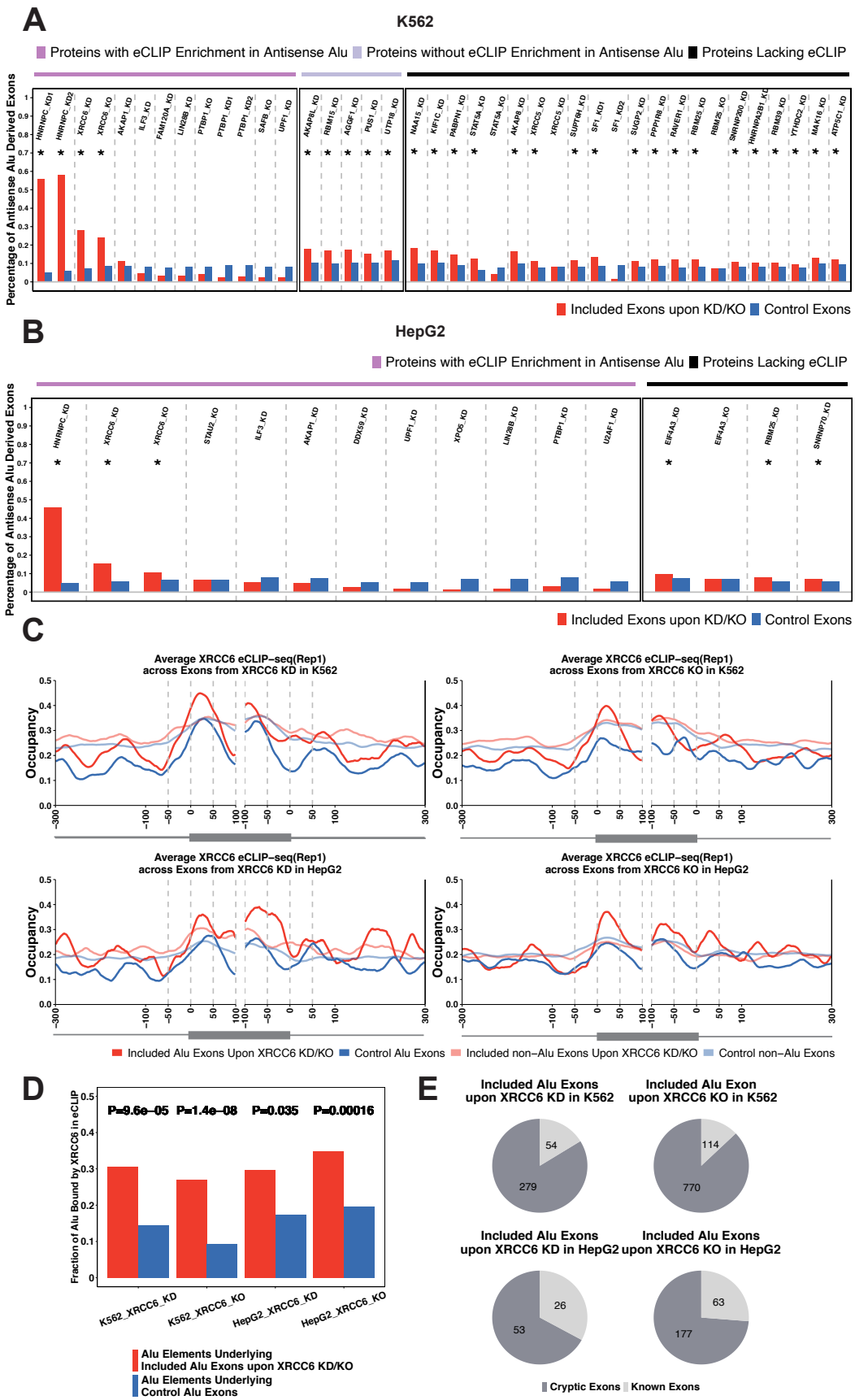

A
K562
Proteins with eCLIP Enrichment in Antisense Alu
Proteins without eCLIP Enrichment in Antisense Alu
Proteins Lacking eCLIP
Percentage of Antisense Alu Derived Exons
HNRNPC_KD1
HNRNPC_KD2
XRCC6_KD
XRCC6_KO
AKAP1_KD
ILF3_KD
FAM120A_KD
LIN28B_KD
PTBP1_KO
PTBP1_KD1
PTBP1_KD2
SAFB_KO
UPF1_KD
AKAP8L_KD
RBM15_KD
AGGF1_KD
PUS1_KO
U2TP8_KD
NAA15_KD
KIF1C_KD
PABPN1_KD
STAT5A_KD
STAT5A_KO
AKAP8_KD
XRCC5_KO
XRCC5_KD
SUPT6H_KD
SF1_KD1
SF1_KD2
SUGP2_KD
PPIP1R8_KO
RAVER1_KD
RBM25_KO
SNRNP200_KD
HNRNPA2B1_KD
RBM39_KD
YTHDC2_KD
MAK16_KD
ATP5C1_KD
Included Exons upon KD/KO
Control Exons
B
HepG2
Proteins with eCLIP Enrichment in Antisense Alu
Proteins Lacking eCLIP
Percentage of Antisense Alu Derived Exons
HNRNPC_KD
XRCC6_KD
XRCC6_KO
STAU2_KO
ILF3_KD
AKAP1_KD
DDX59_KD
UPF1_KD
XPO5_KD
LIN28B_KD
PTBP1_KD
U2AF1_KD
EIF4A3_KD
EIF4A3_KO
RBM25_KD
SNRNP70_KD
Included Exons upon KD/KO
Control Exons
C
Average XRCC6 eCLIP−seq(Rep1) across Exons from XRCC6 KD in K562
Average XRCC6 eCLIP−seq(Rep1) across Exons from XRCC6 KO in K562
Average XRCC6 eCLIP−seq(Rep1) across Exons from XRCC6 KD in HepG2
Average XRCC6 eCLIP−seq(Rep1) across Exons from XRCC6 KO in HepG2
Occupancy
Included Alu Exons Upon XRCC6 KD/KO
Control Alu Exons
Included non−Alu Exons Upon XRCC6 KD/KO
Control non−Alu Exons
D
P=9.6e−05
P=1.4e−08
P=0.035
P=0.00016
Fraction of Alu Bound by XRCC6 in eCLIP
K562_XRCC6_KD
K562_XRCC6_KO
HepG2_XRCC6_KD
HepG2_XRCC6_KO
Alu Elements Underlying Included Alu Exons upon XRCC6 KD/KO
Alu Elements Underlying Control Alu Exons
E
Included Alu Exons upon XRCC6 KD in K562
Included Alu Exon upon XRCC6 KO in K562
54
279
114
770
Included Alu Exons upon XRCC6 KD in HepG2
Included Alu Exons upon XRCC6 KO in HepG2
26
53
63
177
Cryptic Exons
Known Exons

Figure 2. 4 XRCC6(Ku70) suppresses exonization of antisense Alu elements

A. Screening for repressors of antisense Alu exonization in K562 cells. The same strategy as described in **Figure 2.3A** is employed for antisense Alu elements.

B. Screening for repressors of antisense Alu exonization in HepG2 cells, as in A, above.

C. RBPmap shows the high binding occupancy of XRCC6 over included antisense Alu exons upon XRCC6 depletion. XRCC6 occupancy on four different exon sets were calculated: included antisense Alu exons, control antisense Alu exons, included non-Alu exons, and control non-Alu exons. One of the eCLIP-seq replicate is used in this figure. RBPmap based on the other eCLIP-seq replicate are shown in **Supplemental Figure 2.8.**

D For included antisense Alu exons upon XRCC6 KD/KO, a higher fraction of their underlying Alu elements is actually bound by XRCC6, as compared to control antisense Alu exons.

E. The majority of included antisense Alu exons upon XRCC6 depletion are cryptic exons.

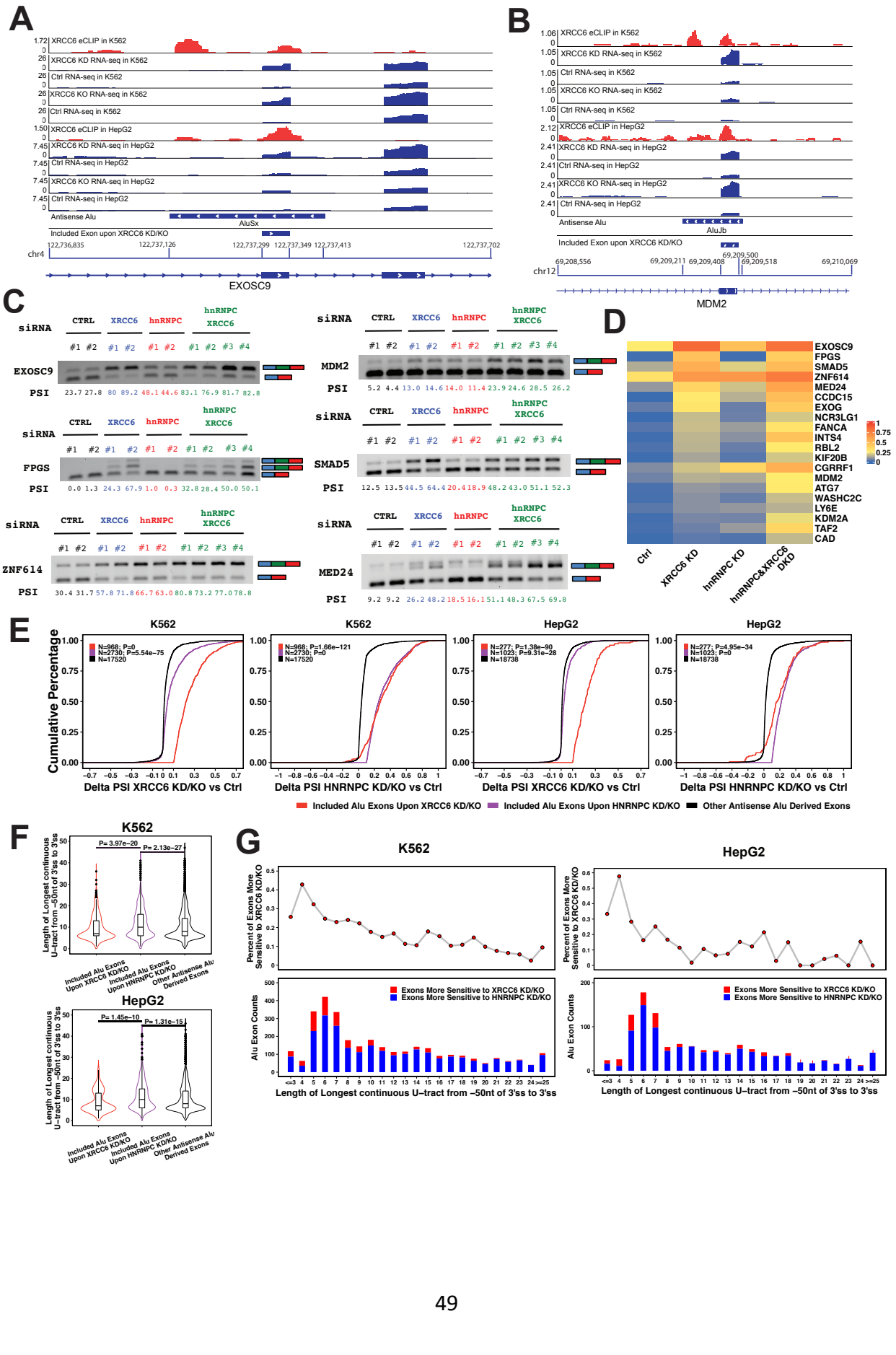

A
XRCC6 eCLIP in K562
XRCC6 KD RNA-seq in K562
Ctrl RNA-seq in K562
XRCC6 KO RNA-seq in K562
Ctrl RNA-seq in K562
XRCC6 eCLIP in HepG2
XRCC6 KD RNA-seq in HepG2
Ctrl RNA-seq in HepG2
XRCC6 KO RNA-seq in HepG2
Ctrl RNA-seq in HepG2
Antisense Alu
AluSx
Included Exon upon XRCC6 KD/KO
chr4
EXOSC9
B
XRCC6 eCLIP in K562
XRCC6 KD RNA-seq in K562
Ctrl RNA-seq in K562
XRCC6 KO RNA-seq in K562
Ctrl RNA-seq in K562
XRCC6 eCLIP in HepG2
XRCC6 KD RNA-seq in HepG2
Ctrl RNA-seq in HepG2
XRCC6 KO RNA-seq in HepG2
Ctrl RNA-seq in HepG2
Antisense Alu
AluJb
Included Exon upon XRCC6 KD/KO
chr12
MDM2
C
siRNA
CTRL
XRCC6
hnRNPC
hnRNPC XRCC6
EXOSC9
PSI
FPGS
ZNF614
MDM2
SMAD5
MED24
D
EXOSC9
FPGS
SMAD5
ZNF614
MED24
CCDC15
EXOG
NCR3LG1
FANCA
INTS4
RBL2
KIF20B
CGRRF1
MDM2
ATG7
WASHC2C
LY6E
KDM2A
TAF2
CAD
Ctrl
XRCC6 KD
hnRNPC KD
hnRNPC&XRCC6 DKD
E
K562
HepG2
Cumulative Percentage
Delta PSI XRCC6 KD/KO vs Ctrl
Delta PSI HNRNPC KD/KO vs Ctrl
Included Alu Exons Upon XRCC6 KD/KO
Included Alu Exons Upon HNRNPC KD/KO
Other Antisense Alu Derived Exons
F
K562
HepG2
Length of Longest continuous U-tract from −50nt of 3'ss to 3'ss
Included Alu Exons Upon XRCC6 KD/KO
Included Alu Exons Upon HNRNPC KD/KO
Other Antisense Alu Derived Exons
G
K562
HepG2
Percent of Exons More Sensitive to XRCC6 KD/KO
Alu Exon Counts
Exons More Sensitive to XRCC6 KD/KO
Exons More Sensitive to HNRNPC KD/KO
Length of Longest continuous U-tract from −50nt of 3'ss to 3'ss

Figure 2. 5 XRCC6 provides an additional safeguard against exonization of antisense Alu exons with shorter continuous U-tract in proximal upstream of 3' splice sites

A and B. Tracks of XRCC6 eCLIP-seq, XRCC6 KD/KO RNA-seq, Ctrl RNA-seq in K562 and HepG2 cells are shown for the XRCC6-repressed antisense Alu exons in EXOSC9 (A) and MDM2 (B), respectively.

C. RT-PCR validation of 20 antisense-Alu-derived exons which are included in XRCC6 KD/KO RNA-seq. RT-PCR assays of the cellular RNA with control siRNA knockdown (in black), XRCC6 siRNA knockdown (in blue), hnRNPC siRNA knockdown (in red), and double XRCC6/hnRNPC knockdown (in green) in K562 cells were carried out to determine their impact on the inclusion of the indicated antisense Alu exon. The PSI value measure by RT-PCR is shown at the bottom. RT-PCR results for the other 14 exons are displayed in **Supplemental Figure 2.10**.

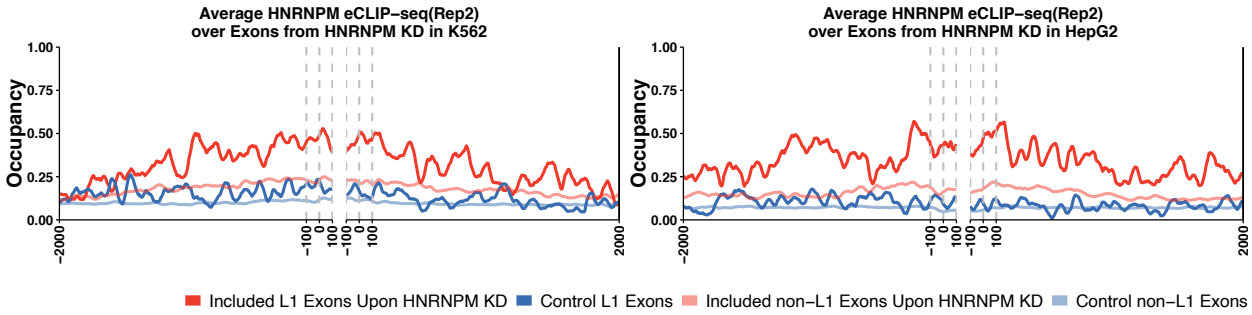
D) The heatmap displaying the average PSI value of the 20 antisense-Alu-derived exons measured by RT-PCR in Ctrl, XRCC6 KD, hnRNPC KD, and hnRNPC & XRCC6 double KD.

E) Most XRCC6-repressed Alu exons are de-repressed upon HNRNPC KD/KO, but not the other way around. For RBPs with two KD/KO RNA-seq datasets available, "included Alu exons upon KD/KO" are defined as the union of included antisense-Alu-derived exons of the two datasets. The remaining antisense-Alu-derived exons which are spliced-in in any XRCC6/HNRNPC depletion datasets or control datasets serve as background. Delta PSI in KD/KO versus Ctrl is the largest delta PSI of the two KD/KOs (Delta PSI refers to the change in PSI values). The cumulative distribution of delta PSI (delta PSI in XRCC6 KD/KO versus Ctrl or delta PSI in HNRNPC KD/KO versus Ctrl) in the three exon sets is depicted in K562 or HepG2 cells. The evaluated significance between the specified exon set and the background set (One-tailed

Wilcoxon test), as well as exon counts in the specified exon set, are indicated next to the colored square symbols.

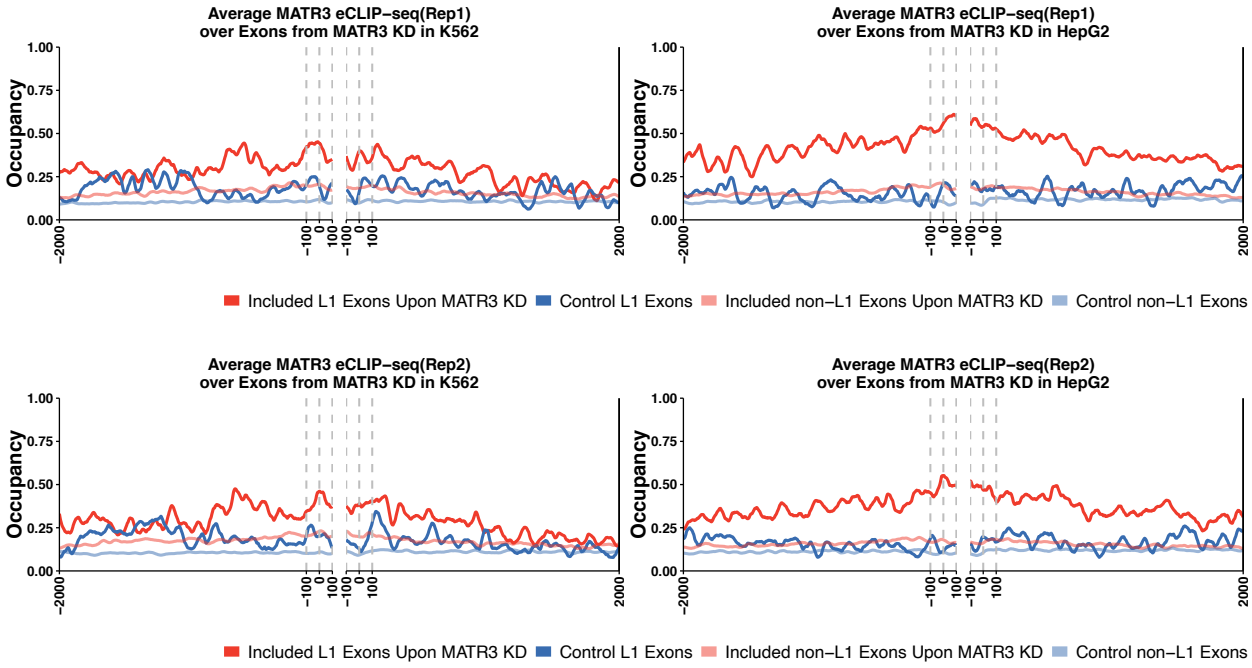
F) The length distribution of the longest continuous U-tract between -50nt of 3' splice sites and 3' splice sites are compared among the three Alu exon sets as defined in E.

G) The union of included Alu exons upon XRCC6 KD/KO and included Alu exons upon HNRNPC KD/KO (as defined in E) are categorized according to the length of the longest U-tract upstream of Alu-encoded 3' splice sites. In each category, exons more sensitive to XRCC6 depletion are those whose delta PSI in XRCC6 KD/KO versus Ctrl are larger than delta PSI in HNRNPC KD/KO versus Ctrl. The exons more sensitive to HNRNPC depletion are analogously defined. Percentage of exons more sensitive to XRCC6 depletion in individual U-tract categories is indicated on the top panel. Alu exon counts are displayed below.



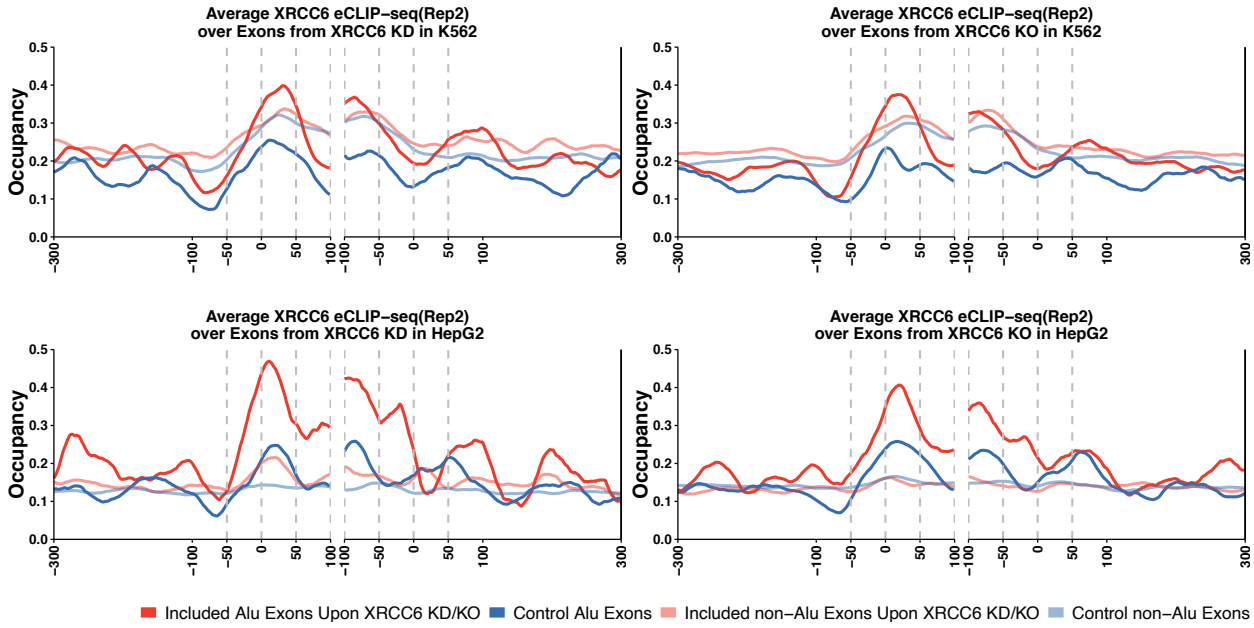
Supplementary Figure 2. 6 Enhanced HNRNPM eCLIP-seq signal over included antisense L1 exons upon HNRNPM KD (using the second eCLIP-seq replicate).

RBPmap based on the second HNRNPM eCLIP-seq replicate. The figure description is as in **Figure 2.3C** and **Methods**.



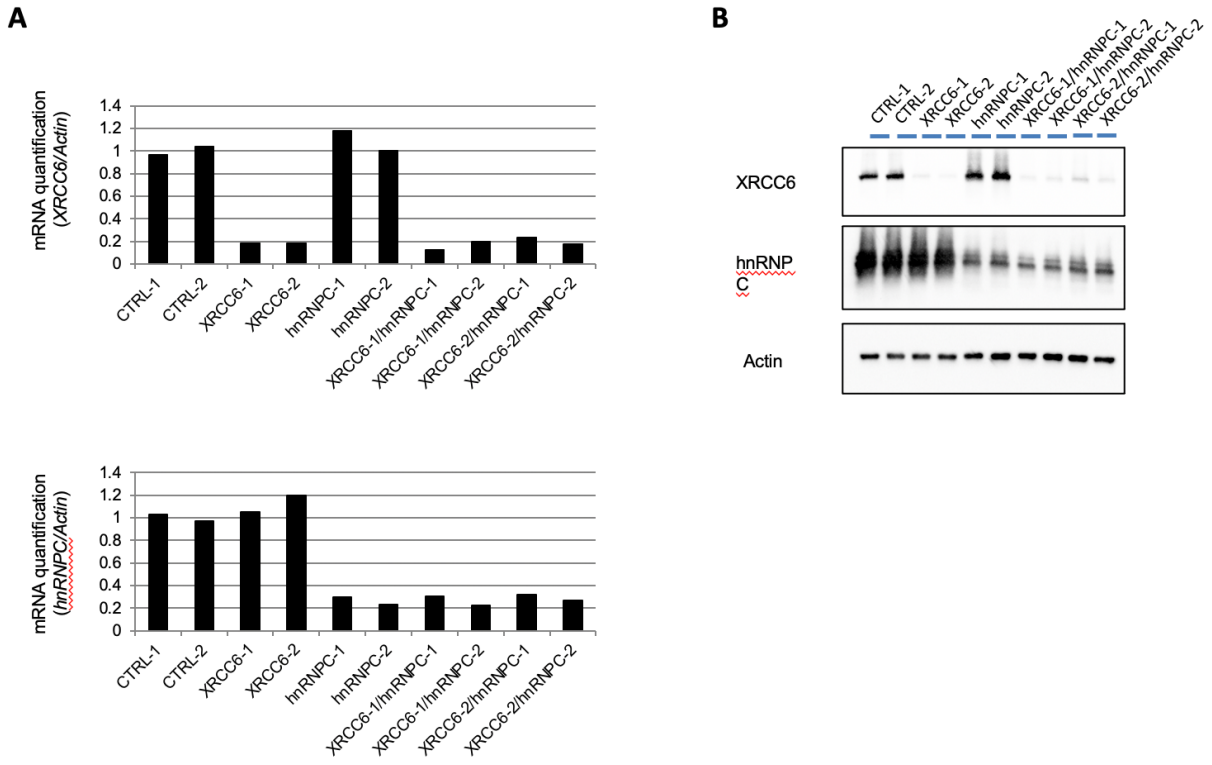
Supplementary Figure 2. 7 Enhanced MATR3 eCLIP-seq signal over included antisense L1 exons upon MATR3 KD

RBPmap based on MATR3 eCLIP-seq over included exons after MATR3 depletion. The figure description is as in **Figure 2.3C** and **Methods**.



Supplementary Figure 2. 8 Enhanced XRCC6 eCLIP-seq signal over included antisense Alu exons upon XRCC6 depletion (using the second eCLIP-seq replicate)

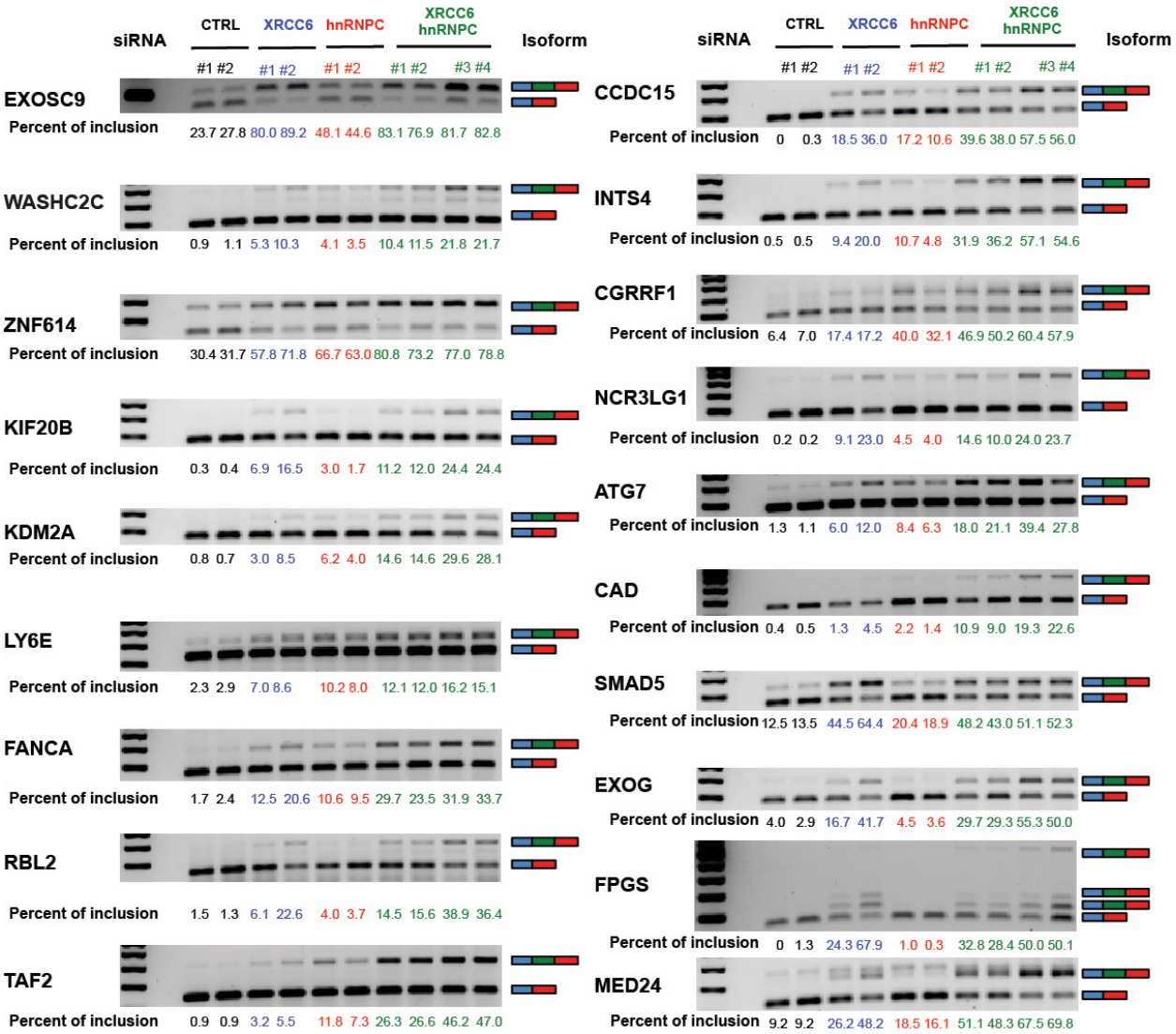
RBPmap based on the second XRCC6 eCLIP-seq replicate. The figure description is as in **Figure 2.3C** and **Methods**.



Supplementary Figure 2. 9 Depletion of XRCC6 and hnRNPC in K562 cells

A. Depletions of XRCC6 and hnRNPC mRNAs in K562 cells by siRNAs. siRNAs targeting XRCC6 (2 independent siRNAs), hnRNPC (2 independent siRNAs), or both (double knockdown, 4 combinations), along with control siRNAs (CTRL, 2 independent siRNAs) were transduced into K562 cells. The impact of the targeted depletions was quantified 6 days after transfection by qRT-PCR (values normalized to corresponding Actin mRNA levels).

B. Western analysis. The protein analysis confirms efficient depletion of XRCC6 and hnRNPC proteins in siRNA transfected K562 cells.

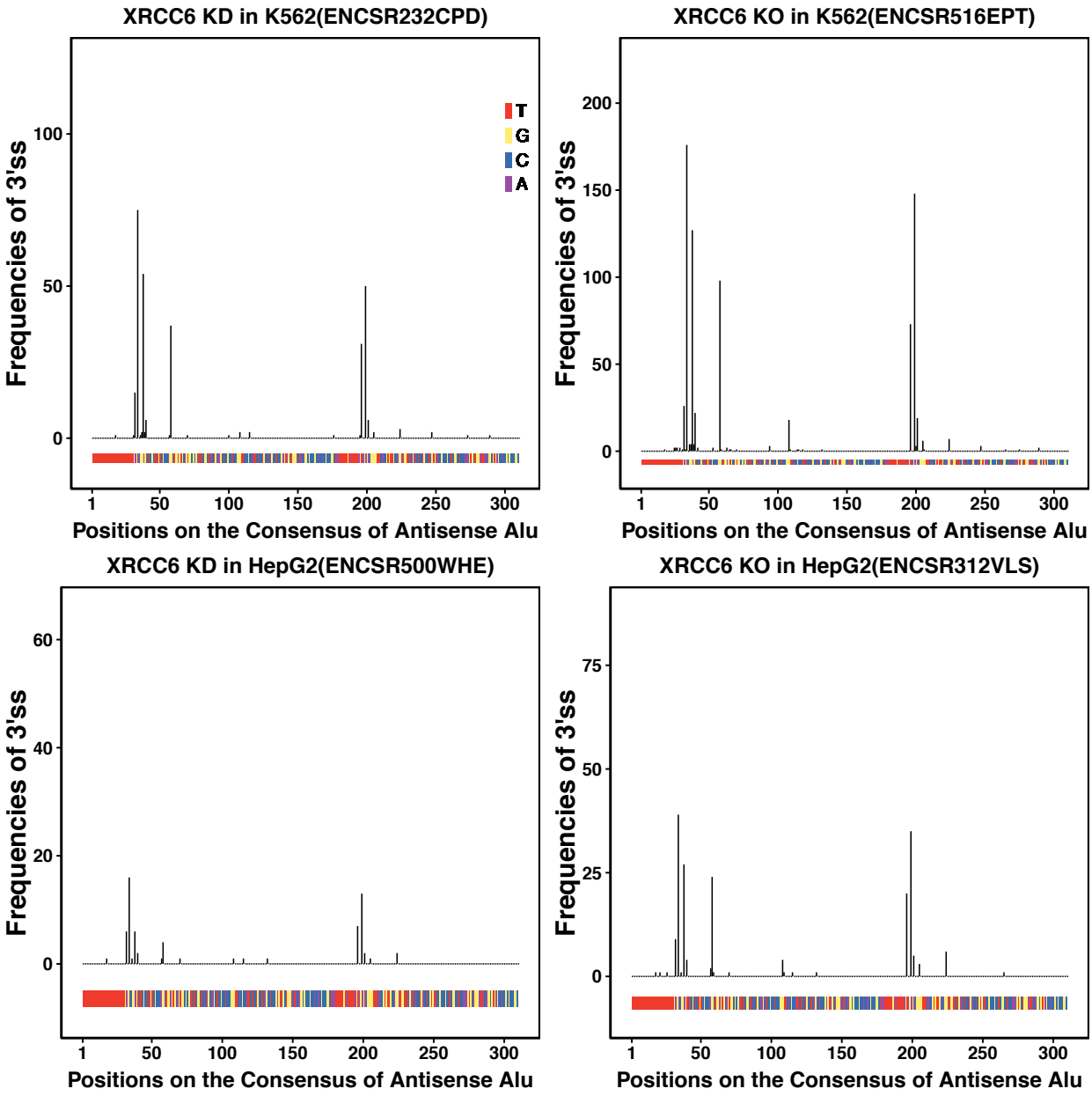


Supplementary Figure 2. 10 Validation of 20 XRCC6-repressed antisense Alu exons by RT-PCR

RT-PCR analysis of the cellular RNA with control siRNA knockdown (in black), XRCC6 siRNA knockdown (in blue), hnRNPC siRNA knockdown (in red), and double XRCC6/hnRNPC knockdown (in green) was carried out to determine their impact on the inclusion of Alu

elements. The percentage of PCR product with Alu exon inclusion (upper bands) over the total PCR products (upper bands + the lowest band) was calculated and shown at the bottom.

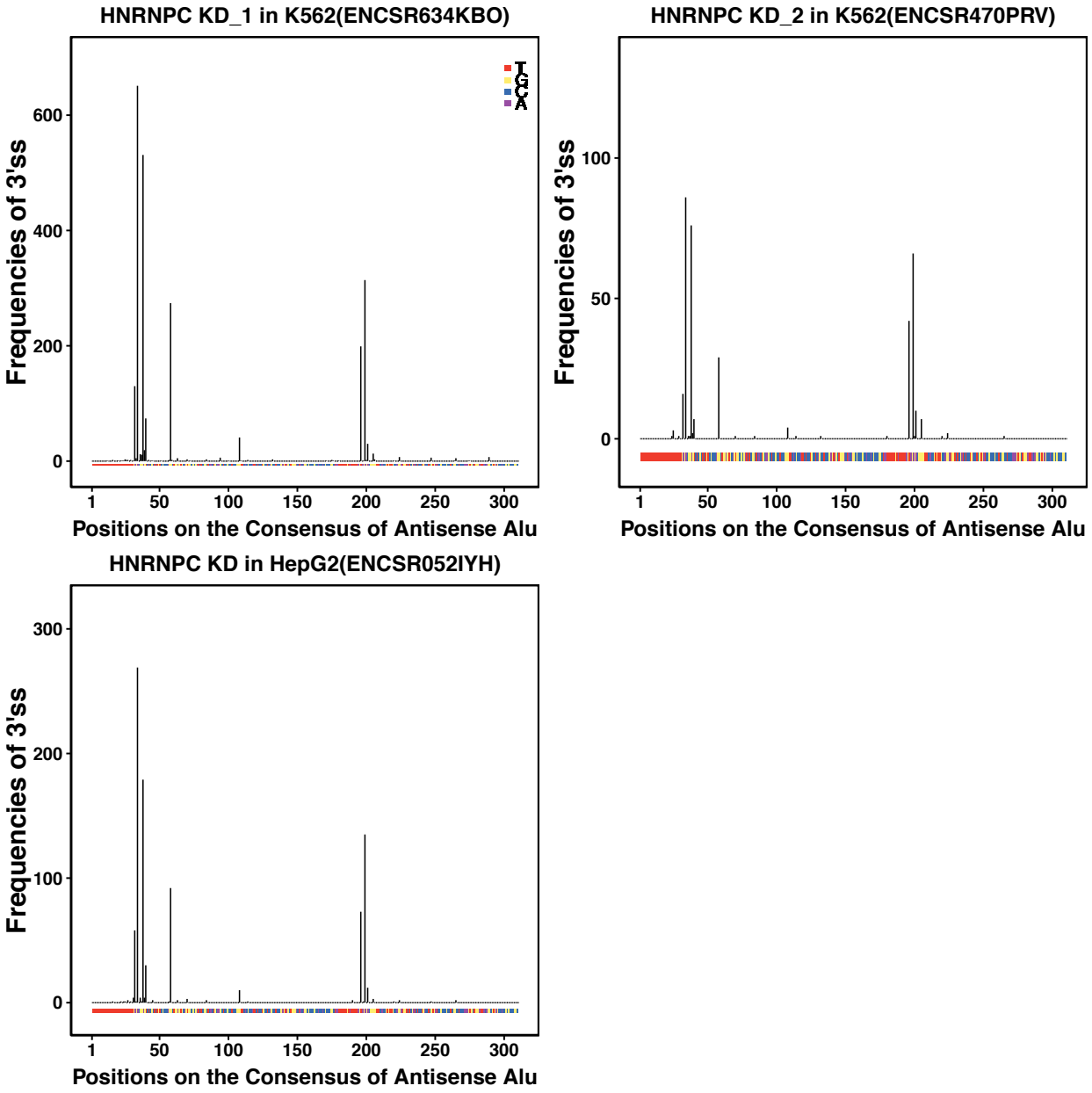
An RT-PCR schematic is displayed to the right of individual Gels. Upstream exon (in blue) and downstream exon (in red) are indicated. Exonized antisense Alu (in dark green) is in the middle.



Supplementary Figure 2. 11 De-repressed Alu-encoded 3' splice sites upon XRCC6 depletion on the consensus sequence of antisense Alu elements

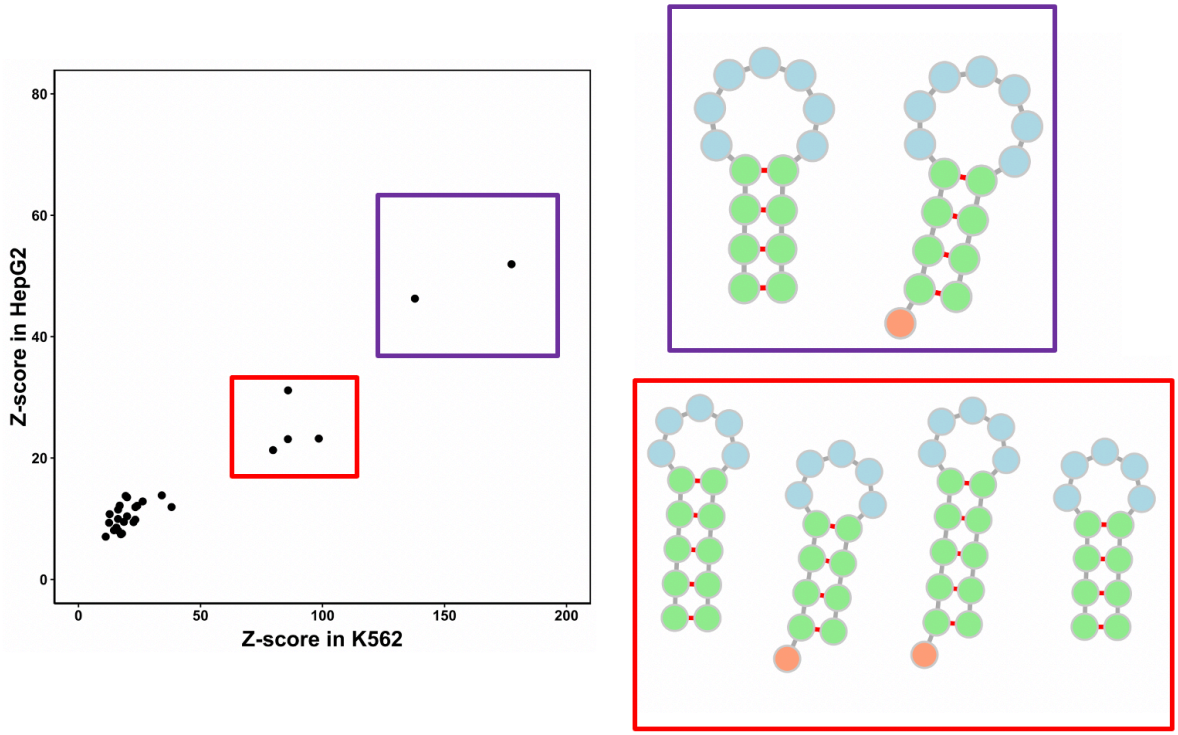
3' splice sites of included antisense Alu exons upon XRCC6 depletion are mapped onto positions on the consensus sequence of antisense Alu. Individual antisense Alu elements are aligned with

the consensus using pairwise alignment by PRANK-F (default parameters for noncoding DNA sequences alignment are used) (**Methods**). The alignment allows for the mapping of 3'ss of included Alu exons onto a position within the consensus sequence. The nucleotides of the consensus sequence are colored as indicated.



Supplementary Figure 2. 12 De-repressed Alu-encoded 3' splice sites upon HNRNPC depletion on the consensus sequence of antisense Alu elements

3' splice sites of included antisense Alu exons upon HNRNPC depletion are mapped onto the consensus sequence of antisense Alu, using the same strategy as described in **Supplementary Figure 2.11**.



Supplementary Figure 2. 13 Hairpin structures are enriched in XRCC6 binding sites in two cell lines

We carry out the discovery of de nove structural elements in XRCC6 eCLIP-seq peaks using algorithm pyteiser with the default parameters. The flanking regions of the same length as peaks are chosen as background. Significant Z-scores for structural elements overly represented in peaks relative to the background are displayed for both K562 cells and HepG2 cells. Two clusters of hairpin structures are enriched by pyteiser, with colored surrounding rectangles highlighted in the figure and on the right side.

2.6 Tables

Table 2. 1 Differential RNA-seq datasets where antisense-L1-derived exons are significantly enriched in included exons upon gene depletion

Datasets_of_depletion	#Included	#TE-derived_included	#Control	#TE-derived_control	P-value
K562_HNRNPM-ENCSR746NIM	730	47	1100	8	4.16E-15
K562_MATR3-ENCSR792XFP	1080	59	467	6	1.09E-11
HepG2_HNRNPM-ENCSR995JMS	678	30	735	1	5.04E-07
K562_HNRNPC-ENCSR634KBO	4474	170	4381	18	5.18E-05
HepG2_MATR3-ENCSR492UFS	1987	76	1440	14	8.10E-05
HepG2_U2AF1-ENCSR372UWV	1520	48	6883	35	0.00015
K562_MAGOH-ENCSR849STR	1007	44	2872	13	0.00015
HepG2_SF3A3-ENCSR374NMJ	921	33	5557	26	0.00020
HepG2_RBM22-ENCSR330KHN	1180	47	1910	8	0.00037
K562_PUF60-ENCSR558XNA	2341	94	6036	63	0.00039
HepG2_U2AF2-ENCSR622MCX	1460	47	4990	30	0.00047
K562_SF3B4-ENCSR081XRA	647	26	5780	23	0.00070
K562_HNRNPC-ENCSR470PRV	642	26	739	3	0.0021
K562_CNOT8-ENCSR312HJY	205	10	1037	9	0.0043
K562_PTBP1-ENCSR527IVX	1753	62	2062	19	0.0058
HepG2_RAVER1-ENCSR576GOW	1044	41	2315	7	0.0067
K562_ACO1-ENCSR237QLO	884	25	2705	11	0.0082
HepG2_MAGOH-ENCSR746EKS	2226	64	3727	21	0.0094

HepG2_HNRNPC-ENCSR052IYH	2071	73	2399	16	0.011
HepG2_SRSF9-ENCSR597XHH	1049	29	1145	6	0.021
K562_SF3B1-ENCSR047QHX	995	25	2957	19	0.033
K562_EIF4A3-ENCSR264MSX	2579	66	2988	24	0.037
HepG2_STAU1-ENCSR124KCF	444	11	148	1	0.045
HepG2_PUF60-ENCSR648BSC	1617	47	4439	49	0.049

Table 2. 2 Differential RNA-seq datasets where antisense-L1-derived exons are significantly enriched in excluded exons upon gene depletion

Datasets_of_depletion	#Excluded	#TE-derived_excluded	#Control	#TE-derived_control	P-value
K562_FASTKD2-ENCSR608IAI	3658	79	109	7	0.0017

Table 2. 3 Differential RNA-seq datasets where antisense-Alu-derived exons are significantly enriched in included exons upon gene depletion

Datasets_of_depletion	#Included	#TE-derived_included	#Control	#TE-derived_control	P-value
K562_HNRNPC-ENCSR634KBO	4474	2491	4381	45	0
HepG2_HNRNPC-ENCSR052IYH	2071	945	2399	12	6.1E-201
K562_HNRNPC-ENCSR470PRV	642	374	739	8	1.1E-176
K562_XRCC6-ENCSR232CPD	1097	309	1718	63	3.60E-73
K562_XRCC6-ENCSR516EPT	3386	817	1364	84	3.06E-69
HepG2_XRCC6-ENCSR500WHE	465	72	654	10	1.67E-14
HepG2_XRCC6-ENCSR312VLS-2	1062	136	867	38	5.57E-14
K562_NAA15-ENCSR945GUR	757	139	3692	71	1.27E-10
K562_AKAP8L-ENCSR809ISU	1013	181	3660	74	2.81E-10
K562_RBM15-ENCSR385UPQ	864	147	4473	101	1.65E-09
HepG2_XRCC6-ENCSR312VLS	1962	212	1329	62	2.01E-09
K562_AGGF1-ENCSR812TLY	1009	175	4166	83	5.26E-09
K562_KIF1C-ENCSR823WTA	756	129	3140	57	6.64E-08
K562_PABPN1-ENCSR416ZJH	1310	192	3327	63	1.30E-07
K562_STAT5A-ENCSR174FUO	438	55	644	13	2.37E-07
K562_AKAP8-ENCSR000YYN	562	92	1572	25	1.39E-06
K562_XRCC5-ENCSR276GMG	2387	264	2494	134	1.59E-06
K562_PUS1-ENCSR618IQH	1220	187	6718	159	1.22E-05
K562_UTP18-ENCSR165VBD	837	142	4040	77	1.25E-05

K562_SUPT6H-ENCSR530BOP	1769	208	2601	39	3.17E-05
K562_SF1-ENCSR562CCA	981	131	5713	116	8.73E-05
K562_SUGP2-ENCSR192BPV	1661	187	2562	69	0.00021
K562_PPP1R8-ENCSR844QNT	1483	178	2658	56	0.00026
K562_RAVER1-ENCSR904BCZ	446	55	593	6	0.00028
K562_RBM25-ENCSR149DMY	455	56	1356	22	0.0018
HepG2_EIF4A3-ENCSR957EEG	1866	179	2645	25	0.0035
K562_SNRNP200-ENCSR943LIB	1299	142	3766	75	0.0035
K562_HNRNPA2B1-ENCSR794NUE	1551	163	4089	104	0.0065
K562_RBM39-ENCSR678WOA	1274	134	5354	132	0.0078
K562_YTHDC2-ENCSR843LYF	1632	155	2704	60	0.011
K562_MAK16-ENCSR517JHY	543	70	1690	43	0.013
HepG2_RBM25-ENCSR610AEI	803	64	2440	34	0.024
HepG2_SNRNP70-ENCSR635BOO	1823	130	1893	23	0.025
K562_ATP5C1-ENCSR231DXJ	542	65	1692	45	0.027

Table 2. 4 Differential RNA-seq datasets where antisense-Alu-derived exons are significantly enriched in excluded exons upon gene depletion

Datasets_of_depletion	#Excluded	#TE-derived_excluded	#Control	#TE-derived_control	P-value
K562_RPS3-ENCSR642GBC	2536	202	930	119	6.91E-06
HepG2_SART3-ENCSR011BBS	3191	208	347	35	0.0063

Table 2. 5 XRCC6-repressed antisense Alu exons validated by RT-PCR

Gene	Chrom	Strand	ExonStart	ExonEnd
ATG7	chr3	+	11402783	11402901
CAD	chr2	+	27448219	27448409
CCDC15	chr11	+	124909968	124910075
CGRRF1	chr14	+	54981993	54982111
EXOSC9	chr4	+	122737298	122737349
FANCA	chr16	-	89826043	89826161
INTS4	chr11	-	77637511	77637688
INTS4	chr11	-	77637511	77637685
KIF20B	chr10	+	91515535	91515680
KDM2A	chr11	+	66948314	66948410
MDM2	chr12	+	69209407	69209500
MED24	chr17	-	38176780	38176894
MED24	chr17	-	38176780	38176872
MED24	chr17	-	38176780	38176872
NCR3LG1	chr11	+	17389385	17389642
RBL2	chr16	+	53488137	53488259
TAF2	chr8	-	120773847	120774012
WASHC2C	chr10	+	46259096	46259292
ZNF614	chr19	-	52520828	52520950
SMAD5	chr5	+	135488363	135488447
EXOG	chr3	+	38565040	38565136
LY6E	chr8	+	144101698	144101794
LY6E	chr8	+	144101678	144101794
LY6E	chr8	+	144101678	144101794
FPGS	chr9	+	130574990	130575041
FPGS	chr9	+	130574990	130575375
FPGS	chr9	+	130574990	130575382
FPGS	chr9	+	130574990	130575085
FPGS	chr9	+	130574990	130575386
FPGS	chr9	+	130574990	130575155

Table 2. 6 Primers for RT-PCR validation

Gene	Forward Primer	Reverse Primer
RBL2	GAACCTGGGAACTTTGGAGAGA	GAACCAGCGTTCAGACACCT
KDM2A	GCGACGACGCTATGAAGATG	CAGAGGATCTCTCAAGCCACC
LY6E	CTGCTGGTACCTGCGTCC	CACGCAGTAGTTGTCCTGGT
CAD	GCACACCAGTGGAGACCATT	ACTTGGCTGGTATGGGCAAA
KLF20B	ACTCAAGCGAAAGAAGCAGAGA	GCTGCAACCAGTTGATCTCG
ZNF614	GTTGGCACATGGACAAGAACC	TGTCCATTGCATTGCTGCAC
ATG7	GGGGACTTGTGTCCAAACCA	CAGTCCTGGACGACTCACAG
EXOSC9	GGCGGTGGTGATCAAGCTAT	TTTGGATTCTTGTCTGGTTCCAA
INTS4	CACCCTCCGAGAAGATCAGC	TCCCTATCAGTAGGGTACTTGGT
NCR3LG1	GACCCTGGGACTGTCTACCA	ACCAGTCCAACACCAATGAATG
CGRRF1	GCTCTACCCAGAGCAAGACC	TGTGCTGAACTGGGTCTCTT
CCDC15	CCACTATGCTGTTGTGGTCCT	GCCTGTGTGCAGAAGCAAAAT
WASHC2C	AAAAGGTGCATCTCTGCTGC	TGCTGAACAACAGGGCAGAT
FANCA	AGACTGGTTACACCTGGAGC	CAAGAATGGTACACGCAGCC
MDM2	TGTTGGTGCACAAAAAGACACT	TCACAGAGAAGCTTGGCACG
TAF2	TCCTGATGTGCGACTCATTCTT	ATCACTTGGCACATGTCCGT
MED24	TGCGACTGCTGAGCTCTAAT	AGGACTCGGTTCAGAGGGT
EXOG	TGGGCCTTTGACCTTACCTC	GCCGATGGCTTCATTGGGTA
SMAD5	TTAAAATGTCCAGAAATCTGCCTC	GTCAGTGGCTACCGAAAGAAC
FPGS	TGCCAGTTTGACTATGCCGT	CTCTTCGTCCAGGTGGTTCC

Chapter 3 Effects of RBP-TE interaction on RNA editing and RNA stability

3.1 Introduction

RNA editing is a widespread post-transcriptional event where specific RNA sequences are changed without changing the underlying DNA[1,2]. This event occurs across many RNA types, such as mRNA, miRNA and rRNA, etc. The most prevalent form of RNA editing is A-to-I editing (Adenosine-to-inosine editing), which comprises 90% of editing events in RNA. Adenosine deaminases (ADAR) catalyzes the chemical conversion of adenosines to inosines[3,4].

Most of A-to-I editing events take place within transcribed Alu elements[5,6]. A few Alu-interactors are discovered by the analysis in Chapter 2. We ask if any of them can contribute to the biogenesis of RNA editing events. To this end, we first recognizing sites whose RNA editing levels are changed upon RBP depletion by re-analyzing the differential RNA-seq data. The ratio of A to G mutations determined by mapped RNA-seq reads at individual sites serves as the metric to measure RNA editing level. By integrative analysis, we found that ILF3 depletion can de-repress RNA editing levels at sites within Alu-derived ILF3 binding sites. Besides, we showed that ILF3 preferentially binds to Alu-Alu RNA duplex, which can serve as the substrate for RNA editing, further supporting the involvement of ILF3 with the regulation of RNA editing.

Other than RNA editing, we also investigate the effects of Alu-interactors on mRNA degradation. A well-studied pathway for RNA decaying is nonsense-mediated mRNA decay pathway (NMD), which eliminates faulty transcripts containing premature termination codons

(PTCs)[7-9]. UPF1, the key factor of NMD pathway[10,11], is among the identified Alu-interactors in Chapter 2. We found that most of Alu derived UPF1 binding sites are within 3'UTR region. Moreover, we observed that the expression levels of transcripts with Alu-derived UFP1 binding sites on 3'UTR are up-regulated after UPF1 depletion in K562 cells, which is consistent with the role of UPF1 in mRNA surveillance.

3.2 Results

3.2.1 ILF3 specifically suppresses RNA editing in inverted repeat Alu elements

The effect of Alu-interacting RBPs on the A-to-I RNA editing level [4,12,13] is surveyed on within-Alu editable sites from the REDIportal database. The REDIportal database is the most comprehensive RNA editing database to date [14,15]; 95% of its entries are A-to-I transitions located within Alu elements. Within Alu-derived RBP binding peaks, the number of sites whose RNA editing levels are changed upon RBP depletion are evaluated relative to the expected number (**Methods**). Upon ILF3 depletion, the number of sites with enhanced RNA editing level is significantly larger than the expected number (**Figure 3.1A)**; in contrast, the number of sites with decreased RNA editing level was not beyond the expected number (**Supplementary Figure 3.3**). These results indicate that ILF3 binding on Alu repressed RNA editing levels at its binding sites. Other Alu-interacting RBPs showed little effect on the regulation of RNA editing (**Figure 3.1A** and **Supplementary Figure 3.3**). RNA editing sites repressed by ILF3 binding in gene RPSA are shown in **Figure 3.1B**. The role ILF3 in repression of RNA editing were confirmed and validated by two recent independent studies [16,17].

ILF3 has a dual binding preference to both sense Alu and antisense Alu (**Figure 2.2A** and **Figure 2.2B**). This motivated us to further examine the characteristics of Alu elements bound by

ILF3. Intriguingly, Alu elements bound by ILF3 were closer to their most proximal predicted inverted repeat Alu (IR*Alu*) partner within the same gene, as compared to the background distribution of all intragenic Alu elements (**Figure 3.1C** and **Figure 3.1D**). A shorter distance between Alu and its putative IRAlu partner would confer a greater chance of *in vivo* formation of a double-stranded Alu-Alu RNA duplex. In accord with this, Alu elements bound by RBPs with unidirectional binding affinities (**Figure 2.2A** and **Figure 2.2B)** were farther away from their closest IRAlu partner (**Figure 3.1C** and **Figure 3.1D**). The specific targeting of ILF3 to the Alu-Alu duplex structure is implicated by the presence of two double-strand RNA (dsRNA) binding domains in ILF3 protein [18].

The capacity of binding both sense Alu and antisense Alu does not necessarily imply specific binding to Alu-Alu duplexes. In addition to ILF3, our eCLIP-seq analysis also showed that RBPs AKAP1 and UPF1 can bind both sense and antisense Alu (**Figure 2.2A** and **Figure 2.2B**). However, AKAP1 seems to bind to Alu-Alu duplexes only in K562 cells, whereas UPF1 does not have this binding property at all (**Figure 3.1C** and **Figure 3.1D**).

3.2.2 The regulation of RNA stability by Alu elements

Our eCLIP-seq analysis indicated that several RBPs, including UPF1, AKAP1, and FAM120A, preferably bind to Alu elements within 3'UTRs (**Figure 2.2A** and **Figure 2.2B**). UPF1 is one of critical RBPs regulating RNA stability in cells. Therefore, we next examined the impact of Alu elements located in 3'UTR on RNA stability, another important post-transcriptional regulation process in RNA metabolism.

We examined the impact of UPF1 depletion on expression of genes with UPF1 binding on their 3'UTRs (**Figure 3.2**). For genes with non-Alu-derived UPF1 binding sites within their 3'UTRs, the elevated gene expression was seen in both K562 and HepG2 cells; however, for genes with Alu-derived UPF1 binding sites in their 3'UTRs, the up-regulation of gene expression was seen in K562 cells, but not in HepG2 cells (**Figure 3.2**). The similar phenomenon was described for RBP HuR, where Alu-derived binding sites have even opposite effects on gene expression, relative to non-Alu-derived binding sites[19]. These observations suggest cellular context and TE context may affect the function of RBP-occupied transposable elements.

AKAP1 binding in the 3'UTR down-regulates gene expression levels in HepG2 cells, but not in K562 cells when AKAP1 was depleted (**Figure 3.2**). The cell type specific observation agrees with a previous analysis done in the same dataset by other researchers. FAM120A binding in the 3'UTRs does not seem to affect gene expression.

3.3 Discussions

By integrating RBP binding data with RNA-seq data, we further investigated the effects of RBP-TE interaction on two more post-transcriptional RNA processing steps- RNA editing and RNA degradation.

For RNA editing, we found that ILF3 specifically represses RNA editing at ILF3 binding sites in inverted repeat Alu elements (IRAlu) (**Figure 3.1A, Figure 3.1C** and **Figure 3.1D**). The preferred binding of ILF3 to IRAlu is reminiscent of the involvement of ILF3 in the biogenesis of circular RNAs[20]. We intended to identify circular RNAs from ENCODE RNA-seq datasets; however, the polyadenylated RNA enrichment strategy used in ENCODE shRNA-seq libraries excludes circular RNA transcripts. It is interesting in future study to investigate the impact of other Alu-interacting RBPs on the regulation of circular RNA biogenesis[21].

For RNA degradation, we have shown that Alu elements bound by UPF1 within 3'UTR are functional in carrying out RNA degradation (**Figure 3.2**). RNA degradation rate can be more precisely measured by sequential RNA-seq following the treatment of Actinomycin D, an anti-neoplastic agent that inhibits gene transcription[22,23]. We postulate that the precise measurement can help clarifying the effects of cellular context and/or TE context on mRNA decay regulation.

We found that RBP-interacting TEs can be involved in regulating alternative polyadenylation (APA) [24,25]. RBP CPSF6 is critical for 3' RNA cleavage and polyadenylation, and its depletion facilitates the use of proximal polyadenylation sites [26,27]. We identified the RBP CPSF6 as

interacting with sense L2 elements in Chapter 2 (**Figure 2.2A**). Consistent with its role in polyadenylation, we found that half of their interaction occurs in 3'UTRs (**Figure 2.2A)**. Therefore, it will be also interesting to further explore how CPSF6-interacting L2 elements regulate alternative polyadenylation.

3.4 Methods

3.4.1 Detection of inverted-repeat Alu elements

We determine inverted-repeat Alu elements that are reverse complementary to each other by the YASS similarity algorithm [28]. Pairwise alignment by YASS was applied to each intragenic Alu element and the other Alus in the same gene but in the opposite orientation. Two Alu elements are considered paired if at least 75% of the two sequences are overlapped in alignment. An Alu element and its putative IRAlu partner are identified if the partner is its nearest paired partner and the distance between them <= 10000nt.

3.4.2 Integrative analyses for effects of RBP binding on RNA editing

We obtained differential RNA-seq data from ENCODE for RBPs associated with Alu in eCLIP. RNA-seq raw reads of KD/KO and control were mapped against the human genome (hg19 assembly) using STAR in two-pass mode [29]. Each technical replicate is separately subjected to the removal of PCR duplicates based on reads coordinates. Subsequently, the read alignment files of two technical replicates are combined. A-to-I RNA editing level is surveyed on every editable site within Alu from the REDIportal database. The REDIportal database is the largest RNA editing resource for humans, in which 16 million A to I events have been collected from various sources. Only sites with read coverage >= 10 in both the collapsed depletion sample and

the collapsed control sample are retained as qualified sites to be used in the subsequent analysis. Qualified sites with differences in RNA editing level >=0.05 are designated sites with increased RNA editing level after KD/KO. Qualified sites with differences in RNA editing level <= -0.05 are designated sites with decreased RNA editing level after KD/KO.

We integrate eCLIP-seq with differential RNA editing to identify putative regulators of RNA editing. In each KD/KO dataset, the foreground set is defined as a collection of qualified sites located within Alu-derived binding sites of the corresponding RBP. We match each site in the foreground set with a randomly selected site from Alu elements not occupied by the corresponding RBP that has identical RNA editing level in control condition. The group of randomly selected sites are considered a background set. 100 background sets are independently produced. The significance of the number of sites with increased/decreased editing levels in the foreground set is evaluated by a one-tailed binomial test parameterized with the expected fraction, which is the averaged fraction across the 100 background sets.

3.4.3 Integrative analyses for the effect of RBP binding on RNA stability

We downloaded transcript expression in KD/KO and control datasets of UPF1 (K562 and HepG2), AKAP1(K562 and HepG2) and FAM120A (K562) from ENCODE. TPM expression values for each transcript were averaged over replicates and a pseudo-count of 1 was added to each combined TPM value before log2 transformation. The 3'UTR region for each transcript was obtained from GENCODE release 40[30]. For a certain RBP, transcripts with RBP binding sites in 3'UTR are grouped into two categories, based on whether the RBP binding sites in 3'UTR overlap Alu elements or not. The same number of transcripts without RBP binding in 3'UTR are

randomly selected to give rise to the background transcript set. Expression levels under control condition of the selected background transcripts are required to match to those of transcripts with RBP binding sites in 3'UTR. A Wilcox test was utilized to assess if RBP binding in 3'UTR causes more shift in expression level upon RBP depletion, as compared to background.

3.5 References

1. Li, S. & Mason, C.E. The pivotal regulatory landscape of RNA modifications. *Annu Rev Genomics Hum Genet* **15**, 127-50 (2014).

2. Song, C.X., Yi, C. & He, C. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nat Biotechnol* **30**, 1107-16 (2012).

3. Nishikura, K. A-to-I editing of coding and non-coding RNAs by ADARs. *Nat Rev Mol Cell Biol* **17**, 83-96 (2016).

4. Savva, Y.A., Rieder, L.E. & Reenan, R.A. The ADAR protein family. *Genome Biol* **13**, 252 (2012).

5. Bazak, L. *et al.* A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res* **24**, 365-76 (2014).

6. Athanasiadis, A., Rich, A. & Maas, S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol* **2**, e391 (2004).

7. Lykke-Andersen, S. & Jensen, T.H. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol* **16**, 665-77 (2015).

8. Kurosaki, T., Popp, M.W. & Maquat, L.E. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. *Nat Rev Mol Cell Biol* **20**, 406-420 (2019).

9. Lejeune, F., Li, X. & Maquat, L.E. Nonsense-mediated mRNA decay in mammalian cells involves decapping, deadenylating, and exonucleolytic activities. *Mol Cell* **12**, 675-87 (2003).

10. Kim, Y.K. & Maquat, L.E. UPFront and center in RNA decay: UPF1 in nonsense-mediated mRNA decay and beyond. *RNA* **25**, 407-422 (2019).

11. Perlick, H.A., Medghalchi, S.M., Spencer, F.A., Kendzior, R.J., Jr. & Dietz, H.C. Mammalian orthologues of a yeast regulator of nonsense transcript stability. *Proc Natl Acad Sci U S A* **93**, 10928-32 (1996).

12. Bass, B.L. RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem* **71**, 817-46 (2002).

13. Nishikura, K. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* **79**, 321-49 (2010).

14. Schaffer, A.A. *et al.* The cell line A-to-I RNA editing catalogue. *Nucleic Acids Res* **48**, 5849-5858 (2020).

15. Lo Giudice, C., Tangaro, M.A., Pesole, G. & Picardi, E. Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal. *Nat Protoc* **15**, 1098-1131 (2020).

16. Freund, E.C. *et al.* Unbiased Identification of trans Regulators of ADAR and A-to-I RNA Editing. *Cell Rep* **31**, 107656 (2020).

17. Chan, T.W. *et al.* RNA editing in cancer impacts mRNA abundance in immune response pathways. *Genome Biol* **21**, 268 (2020).

18. Nakamura, N. *et al.* Interleukin enhancer-binding factor 3/NF110 is a target of YM155, a suppressant of survivin. *Mol Cell Proteomics* **11**, M111 013243 (2012).

19. Kelley, D.R., Hendrickson, D.G., Tenen, D. & Rinn, J.L. Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biol* **15**, 537 (2014).

20. Li, X. *et al.* Coordinated circRNA Biogenesis and Function with NF90/NF110 in Viral Infection. *Mol Cell* **67**, 214-227 e7 (2017).

21. Dong, X. *et al.* circRIP: an accurate tool for identifying circRNA-RBP interactions. *Brief Bioinform* **23**(2022).

22. Ratnadiwakara, M. & Anko, M.L. mRNA Stability Assay Using transcription inhibition by Actinomycin D in Mouse Pluripotent Stem Cells. *Bio Protoc* **8**, e3072 (2018).

23. Wang, X. *et al.* N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**, 117-20 (2014).

24. Tian, B. & Manley, J.L. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**, 18-30 (2017).

25. Gruber, A.J. & Zavolan, M. Alternative cleavage and polyadenylation in health and disease. *Nat Rev Genet* **20**, 599-614 (2019).

26. Martin, G., Gruber, A.R., Keller, W. & Zavolan, M. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep* **1**, 753-63 (2012).

27. Gruber, A.R., Martin, G., Keller, W. & Zavolan, M. Cleavage factor Im is a key regulator of 3' UTR length. *RNA Biol* **9**, 1405-12 (2012).

28. Noe, L. & Kucherov, G. YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res* **33**, W540-3 (2005).

29. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

30. Frankish, A. *et al.* Gencode 2021. *Nucleic Acids Res* **49**, D916-D923 (2021).

3.6 Figures

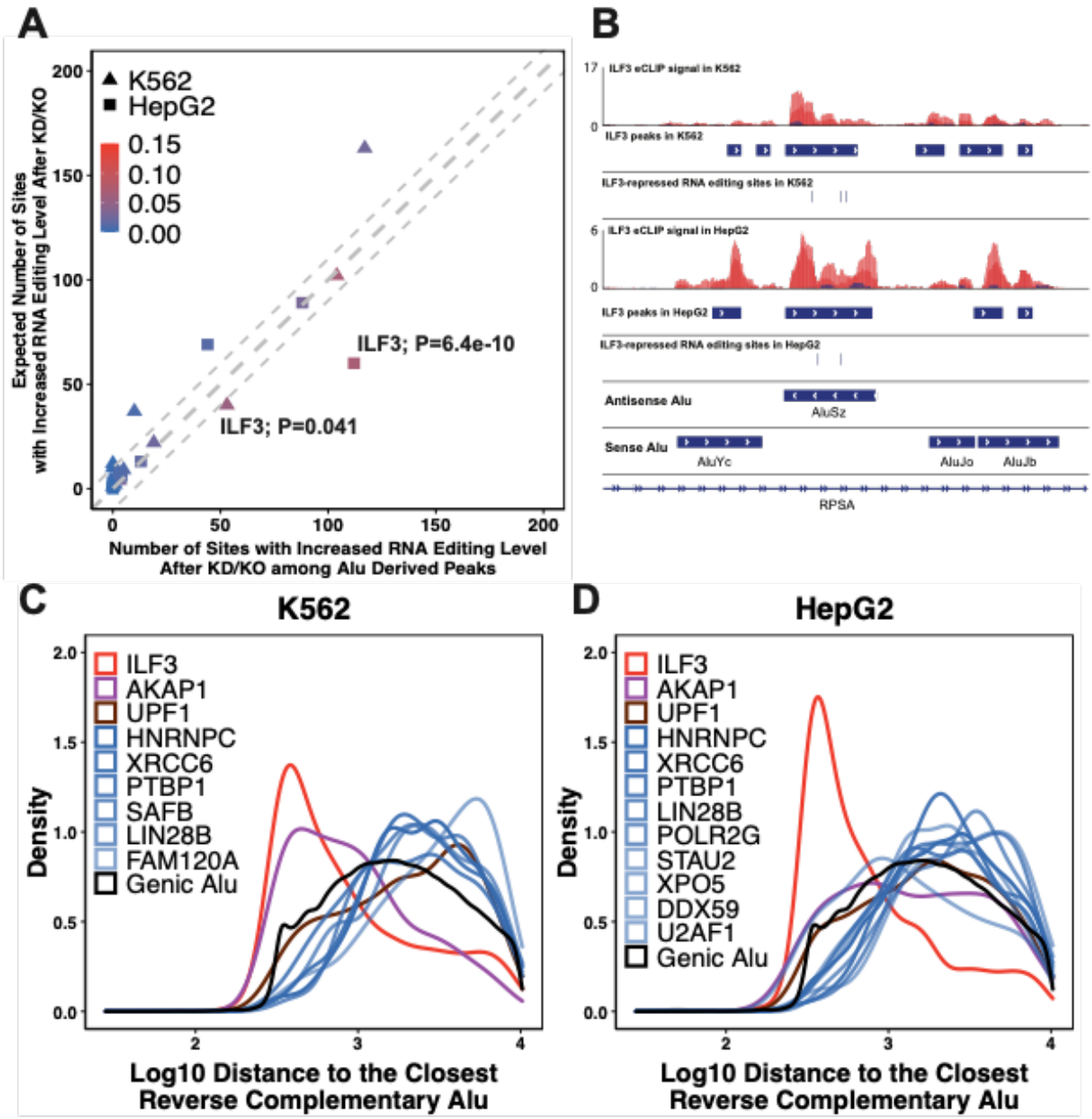


Figure 3. 1 ILF3 specifically suppresses RNA editing levels of sites on inverted repeat Alu elements

A Within Alu-derived RBP binding peaks, the number of sites whose RNA editing levels are increased upon individual RBP depletion is compared to the expected number (**Methods**). Each

dot denotes individual RBP depletion. The expected number (Y axis), based on unbound Alu, is calculated as described in **Methods**. The shading of red color indicates the fraction of sites with increased RNA editing level among editable sites within Alu-derived peaks. The dashed grey lines indicate the diagonal line, the line with slope=1 and intercept=10, and the line with slope=1 and intercept=-10. P-values indicate that the binding of ILF3 on Alu can significantly repress RNA editing levels in both K562 and HepG2 cells.

B. A few ILF3-repressed RNA editing sites within Alu-derived ILF3 binding sites in RPSA in K562 and HepG2 cells.

C and D. For each Alu-binding RBP, the density of distance from RBP-bound Alus to the closest IRAlu is portrayed. The Alu-IRAlu distance of all intragenic Alus is provided as background. The distance is in log10 scale.

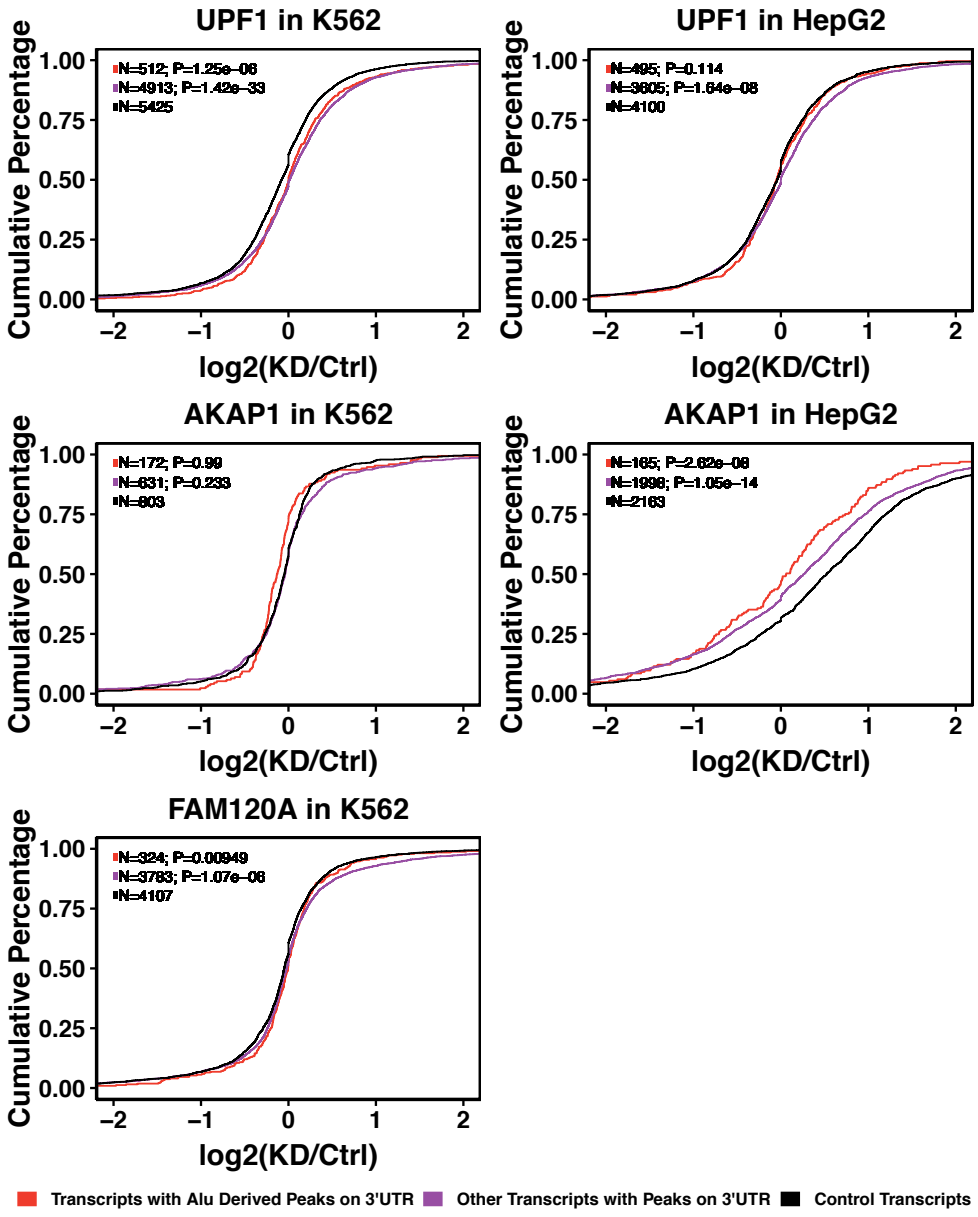
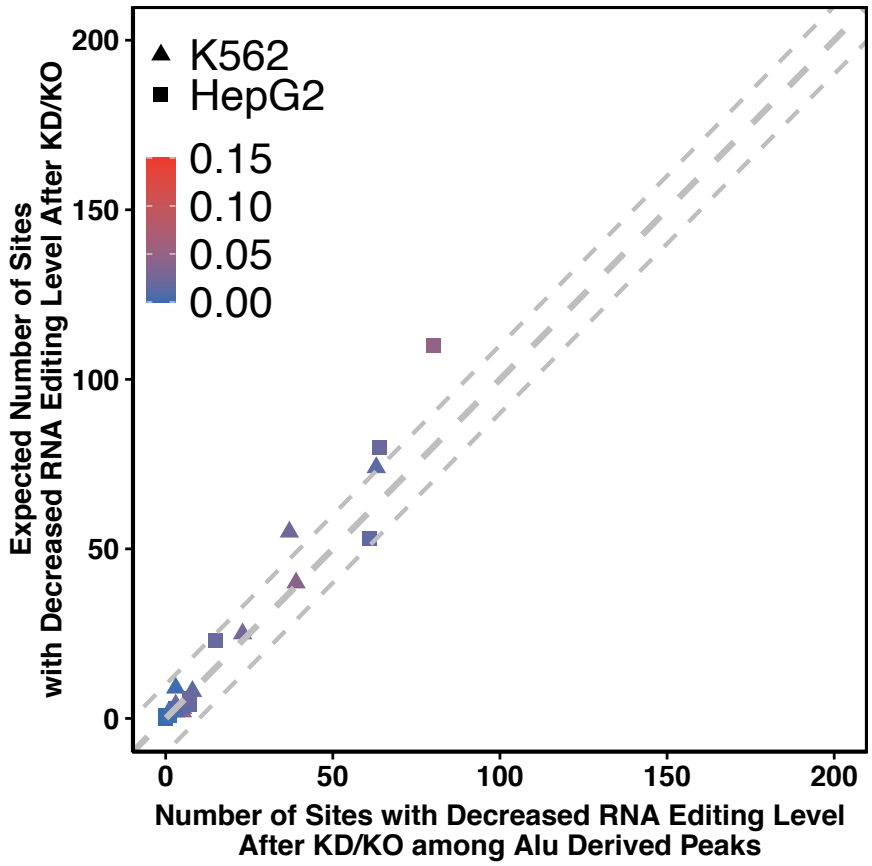


Figure 3. 2 Expression changes of transcripts with RBP binding sites on 3'UTR upon depletion of the same RBP

Transcripts with binding sites on 3'UTR are further dichotomized into transcripts whose 3'UTR peaks overlap Alu elements and those whose 3'UTR peaks do not overlap Alu elements. The expression-matching control transcripts are described in **Methods**. The evaluated significance between the logarithm of TPM ratio of the specified transcript set and that of the control set is

indicated next to the colored square symbol (One-tailed Wilcoxon test). Transcript numbers are placed next to P-values.



Supplementary Figure 3. 3 The number of sites with decreased RNA editing was not beyond the expected number

The same strategy as described in **Figure 3.1A** is implemented for sites with decreased RNA editing levels upon RBP depletion.

Chapter 4 Concluding Remarks

The advance of high-throughput sequencing has tremendously transformed the paradigm of the study in biology and medicine. Exponentially increased data generated by sequencing poses a great challenge to the community. Numerous dedicated computational tools have been developed to process data from individual experiments. In the meantime, efficient integration across multi-omics data adds on to demystify the complicated biological system. Nowadays, data-driven methodology becomes essential to provide new insights and make biological discoveries.

The thesis is a showcase of data-driven study in post-transcriptional regulation. Various post-transcriptional events, including but not limited to RNA capping, RNA splicing, polyadenylation, RNA editing and RNA decaying, enable regulation of gene expression at the RNA level. These events are carefully orchestrated by a multitude of RNA-binding proteins. On the other hand, transposable elements are the single largest component of the genetic material of most eukaryotes. They are inevitably incorporated into and becomes the building blocks of primary transcripts when located in transcribed regions. In the thesis, we attempt to bridge transposable elements to post-transcriptional regulation by means of RNA-binding proteins. We ask whether RNA-binding proteins can pervasively interact with transcribed transposable elements. If this is the case, what is the functional implications of these interplay in post-transcriptional regulation?

In Chapter 2, we proposed a dedicated computational framework to identify RBP binding sites on transposable elements. The use of multi-aligned reads enables recovering as many RBP binding sites as possible in transposable elements. By applying this framework into 223 eCLIP-

seq datasets from ENCODE, we found large-scale interactions between RBPs and TEs, most of which are supported by motif enrichment. Furthermore, we integrate eCLIP-seq data with RNA-seq data to investigate the effects of TE-interactors on TE exonization, a process of intronic TEs being incorporated into mature RNA. Two novel repressors of TE exonization are discovered-HNRNPM for repressing exonization of antisense L1 elements and XRCC6(Ku70) for repressing exonization of antisense Alu. The enhanced binding of these two proteins over included TE-derived exons upon depletion of the corresponding RBP further consolidates our finding. The following analysis examines the way in which XRCC6(Ku70) and HNRNPC (the global repressor for Alu exonization) repress Alu exons. We claimed that XRCC6(Ku70) can provide additional repressiveness for antisense Alu exons with shorter continuous U-tract upstream of Alu-encoded 3'splice sites, on which HNRNPC's effects are compromised.

In Chapter 3, we examine the effects of RBP-TE interaction on RNA editing and RNA stability. We found that ILF3 can suppress RNA editing at sites in inverted repeat Alu elements; the central NMD (Nonsense-mediated mRNA decay) factor UPF1 can enable RNA-degradation by binding to Alu elements on 3'UTR of the target genes.

Collectively, our integrative paradigm expands the prior knowledge regarding the interplay between RBPs and TEs. There are over 1500 RNA-binding proteins and millions of TE elements in human. We anticipate that the functional significance of RBP-TE interaction will draw the attention of more and more researchers. In the long run, the rapid development of single cell sequencing and spatial omics will bring in novel perspectives at single cell resolution, which

undoubtedly allows for a more elaborate investigation on post-transcriptional regulation in the future.