

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Illuminating Molecular Mechanisms for Sequence-Specific RNA-Binding and Endogenous Cellular RNA-Protein Interactions in situ

Permalink

<https://escholarship.org/uc/item/1sm7z1vx>

Author

Liang, Qishan

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

**Illuminating Molecular Mechanisms for Sequence-Specific RNA-Binding and Endogenous
Cellular RNA-Protein Interactions *in situ***

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Biochemistry and Molecular Biophysics

by

Qishan Liang

Committee in charge:

Professor Kevin Corbett, Chair
Professor Neal Devaraj, Co-Chair
Professor Gene Yeo, Co-Chair
Professor Mike Gilson
Professor Amy Pasquinelli
Professor Akif Tezcan
Professor Jin Zhang

2024

Copyright

Qishan Liang, 2024

All rights reserved.

The Dissertation of Qishan Liang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

EPIGRAPH

The way to live is to throw ourselves, not into faith, but into our lives, conducting them in affirmation of every moment, exactly as it is, without wishing that anything was different, and without harboring peevish resentment against others or against our fate.

- *At the Existentialists Café*, Sarah Bakewell

TABLE OF CONTENTS

DISSERTATION APPROVAL PAGE	iii
TABLE OF CONTENTS.....	v
LIST OF FIGURES	vii
LIST OF TABLES	ix
ACKNOWLEDGEMENTS	x
VITA.....	xii
PUBLICATIONS.....	xii
ABSTRACT OF THE DISSERTATION	xiii
INTRODUCTION	1
Chapter 1 Engineering a novel RNA-binding zinc-finger protein for RNA sequence specificity	6
Chapter 1.1 Introduction	6
Chapter 1.2 Results	8
RNA-bind-n-seq (RBNS) characterizes the RNA binding profile of ZnFs	8
Rational mutagenesis of ZRANB2 ZnF in search of novel RNA-binding motifs	14
All-atom MD simulations revealed ZnF-RNA recognition mechanism	18
Chapter 1.3 Conclusion.....	23
Chapter 1.4 Methods	24
Acknowledgements	28
REFERENCES	37
Chapter 2 INSCRIBE enables highly sensitive <i>in situ</i> capture of endogenous RNA-protein interactions in fixed tissues	40
Chapter 2.1 Introduction	40
Chapter 2.2 Results	44

An APOBEC1-nanobody fusion protein mediates targeted RNA editing <i>in situ</i>	44
Comparison of INSCRIBE and eCLIP reveals differences in resolution	50
INSCRIBE enables RNA isoform distinctions in RBP binding sites discovery	51
INSCRIBE identifies RBP binding sites with low sequencing depth and ultra-low input RNA	55
INSCRIBE identifies authentic TDP-43 binding sites through confident C-to-U edit clusters	60
INSCRIBE is compatible with multiple cell fixation methods	63
INSCRIBE identifies RBP-RNA interaction sites in fixed tissue samples	66
Chapter 2.3 Discussion	69
Chapter 2.4 Methods	72
Acknowledgements	80
REFERENCES	100
CONCLUSION	103
REFERENCES	106

LIST OF FIGURES

Figure 1.1 Design of ZRANB2 ZnF mutant and RNA bind-n-seq (RBNS) assay	11
Figure 1.2 Benchmarking of the modified RNA bind-n-seq assay	13
Figure 1.3 RNA bind-n-seq results of selective ZnF single mutants	16
Figure 2.1 Schematics of <i>in situ</i> RNA C-to-U editing mediated by recombinant APOBEC1-Nanobody	43
Figure 2.2 INSCRIBE identifies authentic RBFOX2 binding sites through confident C-to-U edit clusters.	47
Figure 2.3 Long-read sequenced INSCRIBE enables mRNA isoform distinctions	53
Figure 2.4 The performance of INSCRIBE at various sequencing depths and ultra-low RNA input	57
Figure 2.5 INSCRIBE identifies authentic TDP-43 binding sites through confident C-to-U edit clusters	61
Figure 2.6 INSCRIBE exhibited comparable performance in PFA-fixed cells	64
Figure 2.7 INSCRIBE unveiled RNA targets of RBFOX2 in mouse brain tissue	67
Supplemental Figure 1.1 Benchmarking of the modified RNA bind-n-seq (RBNS) assay	29
Supplemental Figure 1.2 Purification of ZnF-array	30
Supplemental Figure 1.3 Top 10 6-mers in RNA bind-n-seq of other ZnF single mutants	31
Supplemental Figure 1.4 Top 10 6-mers in RNA bind-n-seq of ZnF double mutants with compromised binding affinity	32
Supplemental Figure 1.5 RNA bind-n-seq results of selective ZnF double mutants	33
Supplemental Figure 1.6 Molecular dynamics simulation results of ZnF-RNA complexes	34
Supplemental Figure 1.7 Molecular dynamics simulation results of ZnF-RNA complexes (continued)	35
Supplemental Figure 1.8 Molecular dynamics simulation results of ZnF-RNA complexes (continued)	36
Supplemental Figure 2.1 Establishment and optimization of APO1-Nb mediated RNA editing <i>in situ</i>	89
Supplemental Figure 2.2 Establishment and optimization of APO1-Nb mediated RNA editing <i>in situ</i> (continued)	91
Supplemental Figure 2.3 Comparison of INSCRIBE to eCLIP for RBFOX2	92
Supplemental Figure 2.4 PacBio sequenced INSCRIBE enables RNA isoform distinctions	93

Supplemental Figure 2.5 The performance of INSCRIBE at various sequencing depths and ultra-low RNA input (continued)	95
Supplemental Figure 2.6 Additional IGV tracks of INSCRIBE controls with various sequencing depths and ultra-low RNA input	97
Supplemental Figure 2.7 INSCRIBE exhibited comparable performance in PFA-fixed cells (continued).	98
Supplemental Figure 2.8 Additional IGV tracks of RBFOX2-mouse brain RBFOX2-INSCRIBE	99

LIST OF TABLES

Table 1.1 List of single mutants of ZnF.....	12
Table 1.2 List of double mutants of ZnF	17
Table 1.3 RNA binding profile of RBNS of all double mutants	18
Supplemental Table 2.1 DNA, RNA and protein sequences used in the study	81
Supplemental Table 2.1 DNA, RNA and protein sequences used in the study (continued).....	82
Supplemental Table 2.1 DNA, RNA and protein sequences used in the study (continued).....	83
Supplemental Table 2.1 DNA, RNA and protein sequences used in the study (continued).....	84
Supplemental Table 2.1 DNA, RNA and protein sequences used in the study (continued).....	85
Supplemental Table 2.2 Number of editing clusters in individual replicates in the down-sampling experiment.....	86
Supplemental Table 2.3 Number of edit clusters prior and after overlapping individual replicates	87
Supplemental Table 2.4 Number of edit clusters in each RNA regions associated with sequencing depth (RBFOX2 INSCRIBE)	88

ACKNOWLEDGEMENTS

This Ph.D. thesis exists because of these people: my Ph.D. advisors Dr. Kevin Corbett and Dr. Gene Yeo, my family, my colleagues, and my dear friends here in San Diego as my local family.

My heartfelt gratitude goes to both of my thesis advisors. Dr. Kevin Corbett is one of the most understanding and supportive mentors I've ever had, whose feedback, advice and encouragement has been a source of inspiration and strength of mine to overcome challenges. Dr. Gene Yeo always sees the bigger picture and imagines an ambitious solution to the real-world challenges, which eventually led to my accomplishment of developing a novel technology ("INSCRIBE") that advances the RNA biology field. The joint mentorship has been tremendously beneficial to my scientific knowledge and communication skills and allowed me to fulfill my dream entering the Ph.D. program- conducting multidisciplinary research and bridging the gap between biology and physical sciences.

I would especially like to thank my parents, who supported my decision of flying 8,000 miles away from home for six years in the US. Even though they don't fully understand my research subjects, they value higher education and constantly support my academic endeavors. Thank you for your beliefs in my ability to get through hardships independently and shape my own adventures oversea. My thanks to my aunt, who has been courageously battling with cancer. I'm especially inspired by her trust in me and encouragement to advance modern medicine to benefit more people like her.

I am indebted to my "comfort zone" here in San Diego- Hoang Nguyen and An Hsieh, the extremely kind and compassionate besties I am so blessed to have. You accompanied me amongst all the frustrations and storms throughout the journey. The lows and highs we've been

through together are the core memories that I would treasure for life. The national parks we've travelled, the Highway No.1 we've driven, the beers we've had, the memes and reels we've shared... You are undoubtedly the ultimate comfort zone and lovable family for me.

My special thanks to my other dear family in San Diego, the Southern Sea Lion Dance Association. It's a warm, caring, and fun place that I was grateful to discover in 2018. All my lion dance friends - you are the connection to my hometown culture when I am away from home. You inspired me to be strong both mentally and physically. The exhilaration, enthusiasm, passion and love I absorbed from you are invaluable. A huge thank to all of you, for making San Diego the place I call "home".

Of course, I couldn't have appreciated more about my mentors, colleagues, and collaborators within or outside of lab. Many thanks to Dr. Joy Xiang and Dr. Amar Deep, my kind, clever, and reliable mentors, for guiding me in science as well as personal growth. There are too many other awesome and helpful colleagues to mention but include Charlene Her, Dr. Tao Yu, Dr. Qiaozhen Ye, and Dr. Sarah Ur.

Chapter 1, in full, is currently being prepared for submission for publication: Liang, Qishan; Xiang, Shengnan J.; Corbett, Kevin D.; Yeo, Gene W.. "*Engineering and characterization of the Sequence Specificity of an RNA-Binding Zinc Finger Protein*". The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is currently reviewed at *Nature Communications*: Liang, Qishan; Yu, Tao; Kofman, Eric; Jagannatha, Pratibha; Rhine, Kevin; Yee, Brian A.; Corbett, Kevin D.; Yeo, Gene W.. "*High-sensitivity in situ capture of endogenous RNA-protein interactions in fixed cells and primary tissues*". The dissertation author was the primary investigator and author of this paper.

VITA

- 2018 Bachelor of Science in Chemistry with honors, Zhejiang University
- 2020 Master of Science in Chemistry, University of California San Diego
- 2024 Doctor of Philosophy in Biochemistry and Molecular Biophysics, University of California San Diego

PUBLICATIONS

1. Yu T., Liang Q., Hou K., Corbett K.D., Yeo G.W., Mapping RBP binding site using recombinant base editors, *Nature Protocol*, *submitted* (2024)
2. Liang Q., Xiang J., Corbett K.D., Yeo G.W., Rational design yields RNA-binding zinc finger domains with altered sequence specificity, *RNA*, *in review* (2024)
3. Deep A., Liang Q., Enustun E., Pogliano J., Corbett K.D., The bacterial PARIS antiviral system is triggered by a phage anti-restriction protein and blocks protein translation, *Nature*, *accepted* (2024)
4. Liang Q., ..., Corbett K.D., Yeo G.W., High-sensitivity *in situ* capture of endogenous RNA-protein interactions in fixed cells and primary tissues, *Nature Communications* (2024)
5. Enustun E., ... , Liang Q., ..., Cleveland D., Ghassemian M., Yeo G.W., Pogliano J., Corbett K.D., A phage nucleus-associated RNA-binding protein is required for jumbo phage infection, *Nucleic Acids Research* (2024)
6. Correia, J.C., ..., Liang, Q., ..., Yeo, G.W., Ruas, J.L., Zfp697 is an RNA-binding protein that regulates skeletal muscle inflammation and regeneration, *PNAS* (2024)
7. Morelli K., Liang Q., Gosztyla M., ..., Yeo G.W., RNA-Targeting Zinc Finger Nuclease Reverses C9orf72-linked ALS Phenotypes, *Nature Neuroscience*, *in revision* (2023)
8. Liang Q., ..., Corbett K.D., Structure and activity of a bacterial defense-associated 3'-5' exonuclease, *Protein Science* (2022)
9. Morgan S.C., ..., Liang Q., ..., Knight R., Yeo G.W., Laurent L.C. Automated, miniaturized, and scalable screening of healthcare workers, first responders, and students for SARS-CoV-2 in San Diego County. *medRxiv* (2021)

ABSTRACT OF THE DISSERTATION

**Illuminating Molecular Mechanisms for Sequence-Specific RNA-Binding and Endogenous
Cellular RNA-Protein Interactions *in situ***

by

Qishan Liang

Doctor of Philosophy in Biochemistry and Molecular Biophysics

University of California San Diego, 2024

Professor Kevin Corbett, Chair
Professor Neal Devaraj, Co-Chair
Professor Gene Yeo, Co-Chair

RNA and RNA-binding-proteins (RBPs) play essential roles in diverse cellular functions and disease-relevant pathways. Currently, the sequence-specificity of RNA-RBP interactions and

the RNA binding of disease-related RBPs are under studied. Investigating RNA-RBP interaction mechanisms, both at a molecular level and the whole transcriptomic level, is thus crucial to the understanding of fundamental RNA biology and the development of potential RNA therapeutics.

The primary focus of this dissertation is on RNA-RBP interaction mechanisms and novel tool development for sequence-specific RNA-targeting and endogenous RNA-RBP interaction profiling. In Chapter 1, to understand the RNA-recognition rules and to develop a designer protein toolbox for RNA targeting, mutations of an RNA-binding zinc finger (ZnF) in human ZRANB2 was systematically introduced at the key RNA-recognizing residues. RNA-bind-n-seq (RBNS) assay identified mutant ZnFs with an altered sequence specificity, preferring to bind a GGG motif instead of the GGU preferred by wild-type ZRANB2 ZnF. Further, through all-atom molecular dynamics (MD) simulations, we characterized changes in the hydrogen-bond network between the protein and RNA that underly the observed sequence specificity changes. In Chapter 2, we developed INSCRIBE (I*n situ* Sensitive Capture of RNA-protein Interactions in Biological Environments), circumventing the current challenges in probing RBP-RNA interactions within endogenous biological contexts. Harnessing the *in situ* RNA labeling by the APOBEC1 enzyme directed by the nanobody-antibody recognition, this method enables highly specific RNA-binding site identification across a diverse range of fixed biological samples such as HEK293T cells and mouse brain tissue. Applicable to any RBP with available primary antibodies, INSCRIBE enables sensitive capture of RBP-RNA interactions from ultra-low input equivalent to ~5 cells.

Chapter 1 in this dissertation greatly deepens the understanding of RNA-ZnF molecular recognition, which is an important foundation for engineering RNA-binding ZnFs for programmable RNA targeting. The technology presented in Chapter 2, INSCRIBE, can be

broadly applied in RBP-RNA interaction profiling in a variety of biological samples, bridging the critical gap in unveiling RBP roles in RNA-related diseases within endogenous biological contexts such as developmental studies, animal disease models, and clinical samples. In summary, this dissertation advances the knowledge of RNA-ZnF molecular recognition and introduces technology for profiling RBP-RNA interactions *in situ*, addressing the challenges in understanding RNA-relevant diseases and therapeutics developments.

INTRODUCTION

RNA, the intermediate of genetic information, and protein, the executer of many cellular functions, are crucial components of the central dogma of biology. RNA-binding-proteins (RBPs) play critical roles in the life cycle of mRNAs and other functional RNAs. They are indispensable for processes such as RNA modification, splicing, stability, localization, translation, and various other regulatory events occurring after transcription^{1,2,3}. The aim of this thesis is investigating RNAs and RBPs that process RNAs, which is critical for advancing the fundamental knowledge of biology.

As an important technology in modern biological research as well as in therapeutics development, targeting and manipulating cellular RNAs endogenously comes in many flavors adopting various strategies. Existing sequence-specific RNA-targeting tools include anti-sense oligos (ASOs), designer RNA-binding proteins of the PUF (Pumilio and FBF) protein family, and CRISPR-Cas proteins^{4,5,6,7,8}. Each of these tools has limitations that restrict its use in laboratory and/or clinical settings. In Chapter 1, I aim to develop a modular RNA-targeting toolbox using an RNA-targeting zinc finger (ZnF) protein⁹ as a scaffold. The compactness and the minimal immunogenicity render it a potentially advantageous RNA-targeting tool over many others. Since very limited natural ZnFs exhibit defined target sequences, I engineered the single ZRANB2 ZnF domain¹⁰ by rational design to develop ZnF domain with altered RNA sequence specificity. Using a modified RNA bind-n-seq (RBNS) assay¹¹, I identified one single mutant that retained high RNA binding affinity but showed an altered preferred target sequence GGG, instead of the GGU motif recognized by the wild-type ZnF. Using molecular dynamics simulations, I found that the changed hydrogen bonding patterns and binding conformations of

the mutant ZnF underlying its altered sequence specificity. The engineering and characterization of ZRANB2 ZnF-RNA binding expands the knowledge of ZnF-RNA recognition principles in general, which provides a foundation for engineering compact, designer RNA-targeting proteins.

The dysregulation of RBP functions is associated with many diseases including neurodegenerative disorders and cancer^{12,13,14} arising from disrupted RBP-RNA interactions and RNA processing. Therefore, extensive efforts have been invested into mapping and profiling the binding sequences of RBPs among the RNA transcriptome. However, the “gold standard” approach CLIP-seq (Cross-Linking and Immunoprecipitation followed by sequencing)¹⁵ is laborious and requires substantial sample materials as an immunoprecipitation (IP)-based technology. The recent emerging methods that harness RNA modifying enzymes have overcome the drawback of the IP-based method, enabling RNA isoforms distinguishment with minimal input materials. However, due to the overexpression of the RBP-enzyme fusion protein in living cells or model organisms, such methods as STAMP (Surveying Targets by APOBEC-Mediated Profiling)¹⁶ and TRIBE (Targets of RNA-binding proteins Identified By Eediting)¹⁷ unavoidably lead to artifacts in the endogenous RBP's localization, functionality, and RNA-binding characteristics. In Chapter 2, I developed a novel technology INSCRIBE (I*n situ* Sensitive Capture of RNA-protein Interactions in Biological Environments) that preserves an authentic RBP-RNA interaction map in undisrupted biological contexts. INSCRIBE utilizes a purified recombinant protein APOBEC1-nanobody that is precisely directed to the RBP of interest through the nanobody-primary antibody recognition. We demonstrated the accurate and robust mapping of RBFOX2 and TDP-43 by INSCRIBE in a variety of fixated samples, including methanol- or formaldehyde-fixed HEK293T cells and mouse brain tissue. Remarkably, INSCRIBE can identify RNA isoform-specific RBP-RNA interactions and transcriptome-wide

endogenous RBP-RNA interactions with as few as ~5 cells. We envision INSCRIBE to be the approach to breakthrough numerous biological inquiries in disease development and therapeutic response for its applications in precious clinical samples where preserving the native biological context is crucial.

REFERENCES

1. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat Rev Genet.* 2014 Dec;15(12):829–45.
2. Ule J, Blencowe BJ. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Molecular Cell.* 2019 Oct 17;76(2):329–45.
3. Das S, Vera M, Gandin V, Singer RH, Tutucci E. Intracellular mRNA transport and localized translation. *Nat Rev Mol Cell Biol.* 2021 Jul;22(7):483–504.
4. Lauffer MC, van Roon-Mom W, Aartsma-Rus A. Possibilities and limitations of antisense oligonucleotide therapies for the treatment of monogenic disorders. *Commun Med.* 2024 Jan 5;4(1):1–11.
5. Zhao YY, Mao MW, Zhang WJ, Wang J, Li HT, Yang Y, Wang Z, Wu JW. Expanding RNA binding specificity and affinity of engineered PUF domains. *Nucleic Acids Research.* 2018;
6. Sugimoto M, Suda A, Futaki S, Imanishi M. Effective RNA Regulation by Combination of Multiple Programmable RNA-Binding Proteins. *Applied Sciences* 2020, Vol 10, Page 6803. 2020 Sep 28;10(19):6803.
7. Konermann S, Lotfy P, Brideau NJ, Oki J, Shokhirev MN, Hsu PD. Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. *Cell.* 2018 Apr 19;173(3):665-676.e14.
8. Lau CH, Suh Y. In vivo epigenome editing and transcriptional modulation using CRISPR technology. *Transgenic Res.* 2018 Dec;27(6):489–509.
9. Font J, Mackay JP. Beyond DNA: Zinc Finger Domains as RNA-Binding Modules. *Methods in Molecular Biology.* 2010;649:479–91.
10. Nguyen CD, Mansfield RE, Leung W, Vaz PM, Loughlin FE, Grant RP, Mackay JP. Characterization of a family of RanBP2-Type zinc fingers that can recognize single-stranded RNA. *Journal of Molecular Biology.* 2011;407(2):273–83.
11. Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB. RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Molecular Cell.* 2014 Jun 5;54(5):887–900.
12. Prashad S, Gopal PP. RNA-binding proteins in neurological development and disease. *RNA Biology.* 2021 Jul 3;18(7):972–87.
13. Pereira B, Billaud M, Almeida R. RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends in Cancer.* 2017 Jul 1;3(7):506–28.

14. Wolozin B, Ivanov P. Stress granules and neurodegeneration. *Nat Rev Neurosci*. 2019 Nov;20(11):649–66.
15. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, Stanton R, Rigo F, Guttman M, Yeo GW. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods*. 2016 Jun;13(6):508–14.
16. Brannan KW, Chaim IA, Marina RJ, Yee BA, Kofman ER, Lorenz DA, Jagannatha P, Dong KD, Madrigal AA, Underwood JG, Yeo GW. Robust single-cell discovery of RNA targets of RNA-binding proteins and ribosomes. *Nat Methods*. 2021 May;18(5):507–19.
17. McMahon AC, Rahman R, Jin H, Shen JL, Fieldsend A, Luo W, Rosbash M. TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. *Cell*. 2016 Apr 21;165(3):742–53.

Chapter 1 Engineering a novel RNA-binding zinc-finger protein for RNA sequence specificity

Chapter 1.1 Introduction

Proper regulation of RNA processing, localization, translation, and degradation is critical for the health of all cells, and defects in any of these processes can lead to disease¹. As such, manipulation of endogenous RNAs in a sequence-specific manner represents an important strategy for both fundamental research into RNA-related pathways, and for the eventual development of RNA-targeting therapeutics. Existing sequence-specific RNA targeting tools include antisense oligos (ASOs), designer RNA-binding proteins of the PUF (Pumilio and FBF) protein family, and CRISPR-Cas proteins. Each of these tools has drawbacks that limit its use in laboratory and/or clinical settings. For example, ASOs are difficult to deliver specifically to affected tissues and have to be frequently redosed²; PUF proteins show limited RNA sequence specificity and cannot efficiently target RNA sequences over 8 nucleotides in length^{3,4}; and CRISPR-Cas proteins are large and therefore challenging to deliver as transgenes. In addition, the bacterial origin of CRISPR-Cas proteins introduces concerns about their potential immunogenicity when used in a clinical setting^{5,6}.

One class of proteins with the potential to overcome the challenges of existing RNA-targeting tools is RNA-binding zinc finger proteins (ZnFs). Modular arrays of DNA-binding ZnFs have been used for programmable, sequence-specific recognition of DNA for over two decades^{7,8,9}. More recently, a class of RNA-binding ZnF proteins has been identified¹⁰, whose individual ZnF domains (each around 3 kDa) specifically bind 3-base single-stranded RNA (ssRNA) motifs and can be assembled into arrays to target longer sequences. For example, the

two ZnF domains of *H. sapiens* ZRANB2 (ZnF Ran Binding Domain-Containing Protein 2) each bind a core GGU motif with micromolar affinity¹⁰, and an artificial array of three ZRANB2 ZnFs binds 5'-GGUGGUGGU-3' with a dissociation constant (K_d) of around 10 nM¹¹. Further, a 6-ZnF array was shown to recognize a 20-nucleotide RNA sequence in a modular manner both *in vitro* and in living cells¹². These data suggest that modular arrays of RNA-binding ZnFs could represent an RNA-targeting tool with distinct advantages: (1) compactness: only four ZnFs (~100 amino acids total) may be required for site specificity in the human transcriptome (12 RNA bases); (2) minimal immunogenicity: ZnF arrays could be assembled from individual domains with human origins.

While RNA-binding ZnFs hold promise as programmable sequence-specific targeting reagents, the number of ZnFs with defined target sequences is extremely limited. Here, we undertook a rational design approach to develop mutants of a single ZRANB2 ZnF domain with altered RNA sequence specificity. Using a modified RNA bind-n-seq (RBNS) approach^{13,14}, we systematically tested a total of 38 single and double mutants of a ZRANB2 ZnF domain for RNA binding affinity and sequence specificity. We identified one single mutant that retained high RNA binding affinity but showed an altered sequence specificity, preferring GGG rather than the GGU targeted by the wild-type ZnF. Using a series of molecular dynamics simulations of the ZnF-RNA complex, we find that this mutant adopts a distinctive conformation and hydrogen bonding pattern that accounts for its altered sequence specificity. Overall, our engineering and characterization of ZRANB2 ZnF-RNA binding enhances the understanding of ZnF-RNA recognition principles and serves as a basis for the engineering of compact, designer RNA-targeting proteins.

Chapter 1.2 Results

RNA-bind-n-seq (RBNS) characterizes the RNA binding profile of ZnFs

The *H. sapiens* alternative splicing regulator ZRANB2 is a member of the broad SR protein family, which typically contain at least one RNA recognition motif (RRM) and a serine/arginine (SR) rich domain implicated in protein-protein interactions (**Figure 1.1A**)^{15,16}. ZRANB2 encodes two RanBP2-type ZnFs (here termed N' and C') at its N-terminus, which are 48% identical to one another at the amino acid level and both recognize single stranded RNA (ssRNA) with a 5'-AGGUAA-3' motif, where "GGU" is the core 3-mer motif (**Figure 1.1B**)¹⁵. To establish an experimental framework for rational design and testing of RNA-targeting ZnFs, we designed a modular 2-ZnF array with one copy of the ZRANB2 N'-ZnF (the "anchor" ZnF) followed by a short linker (amino acid sequence: GSGSG) and a second ZnF (the "test" ZnF; **Figure 1.1C**). In combination with a modified RBNS assay (described below), this experimental design enables rapid assembly and testing of diverse test ZnFs for RNA binding affinity and specificity. We generated an initial panel of 2-ZnF arrays with test ZnFs from ZRANB2 (N' and C', both of which recognize GGU) and a second RBP, RBM5, whose ZnF recognizes GGG^{17,18}. As a negative control, we also generated a mutant of the ZRANB2 C'-ZnF with a key tryptophan residue mutated to alanine (W79A); this residue interacts with the two guanine residues of the core GGU motif through pi-stacking¹⁵, and its mutation to alanine is expected to abolish RNA binding.

We designed a modified high-throughput RNA-bind-n-seq (RBNS) assay^{13,14} to assess the RNA binding characteristics of the recombinant 2-ZnF arrays (**Figure 1.1C**). In RBNS, a randomized pool of RNAs is affinity-purified by an immobilized RNA binding protein, then the

purified RNA pool is deep-sequenced and compared to the input RNA pool to calculate the fold-enrichment of each sequence motif. By performing RBNS at a range of protein concentrations, an estimate of binding affinity for each sequence motif can be determined, in addition to identification of preferred binding motifs¹³. We designed an RNA pool with the sequence 5'-AAAGGUNNNNNNAAA-3', where "N" represents any base (A/U/G/C) mixed at equal frequencies in the library. The leading GGU motif serves as a binding site for the anchor ZnF, allowing the test ZnF to bind preferred sequences within the 6-base random window. Native ZRANB2 has a 23-residue disordered linker between its N' and C' ZnFs, and the protein recognizes tandem GGU motifs with a spacing of 2-5 bases between motifs¹⁵; we anticipated that the shorter five-residue linker in our 2-ZnF arrays would enable recognition of more closely spaced motifs. The 6-base random window in our RNA pool enables the test ZnF to selectively bind RNAs with preferred motifs spaced between 0 and 3 bases from the anchor ZnF's GGU motif.

We performed RBNS with our initial panel of 2-ZnF arrays at four different protein concentrations (160, 320, 640, and 1280 nM), deep-sequenced the input and purified RNA pools, and performed two analyses: first, we calculated the fold-enrichment for each 6-base sequence by comparing input to purified RNA pools. Second, we used kpLogo¹⁹ to search for enriched 3-mer motifs at each position (+0, +1, +2, and +3) within the 6-base random window. In our initial panel of 2-ZnF arrays, we found that the ZRANB2 N'-ZnF, ZRANB2 C'-ZnF, and RBM5 ZnF all showed high selectivity (specificity) at the lowest tested protein concentration (160 nM), with the most-enriched sequences being 100 to 200-fold enriched over input levels (**Figure 1.2A-C**). Fold-enrichments for preferred sequences were lower at higher protein concentrations, indicating that at these concentrations the protein nonspecifically binds a wide range of sequences. The k-

means clustering analysis revealed distinct clusters among sequences, each characterized by similar patterns of the fold-enrichment values at four different protein concentrations (**Supplemental Figure 1.1A**). The most highly enriched sequences for the two ZRANB2 ZnFs showed a preponderance of sequences containing GGU motifs within the 6-base random window, while the RBM5 ZnF showed enrichment of sequences containing GGG motifs (**Supplemental Figure 1.1B**). In agreement with these findings, kpLogo analysis with both ZRANB2 ZnFs showed GGU as the most preferred motif at all positions (+0 to +3), while the RBM5 ZnF showed GGG as the preferred motif at most positions (**Figure 1.2E-F**). As expected, the ZRANB2 C'-ZnF W79A mutant showed essentially no specific enrichment, and kpLogo analysis did not identify consistently preferred motifs across different protein concentrations (**Figure 1.2D, 2F**). Overall, these data show that the 2-ZnF array design coupled to our modified RBNS assay can robustly identify preferred RNA binding motifs of individual ZnFs.

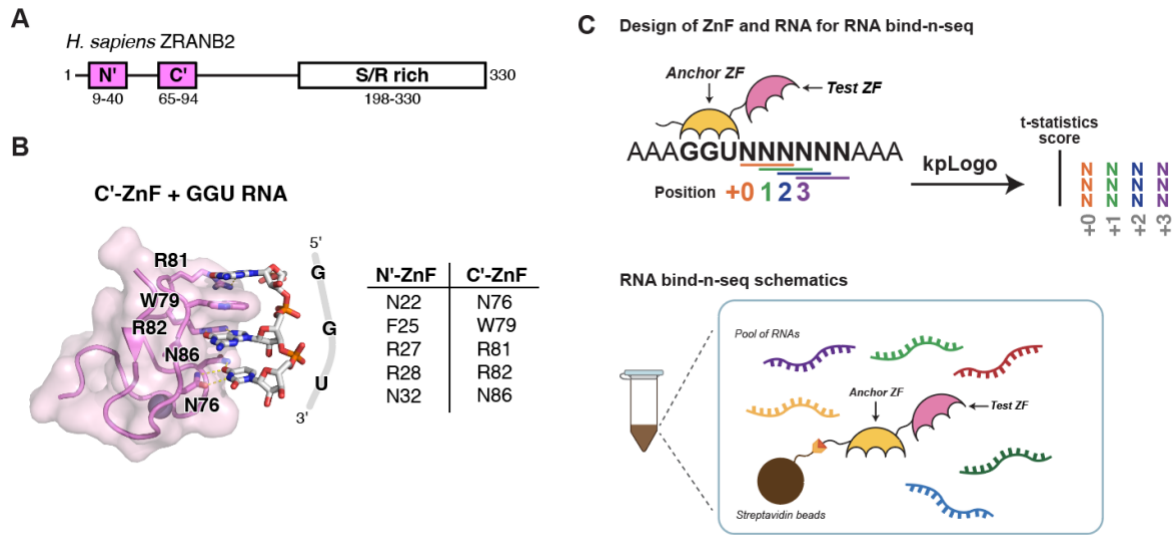


Figure 1.1 Design of ZRANB2 ZnF mutant and RNA bind-n-seq (RBNS) assay

(A) Domains of human ZRANB2 protein. (B) Structure of human ZRANB2 C'-ZnF in complex with GGU RNA (PDB: 3g9y) and the homology between N'-ZnF and C'-ZnF. (C) Schematics of the modified RBNS assay with the ZnF array.

Table 1.1 List of single mutants of ZnF

Mutant No.	N22	R27	R28	N32
1	R			
2	D			
3	E			
4	Q			
5	H			
6	S			
7		D		
8		E		
9		N		
10		Q		
11		H		
12		S		
13			D	
14			E	
15			N	
16			Q	
17			H	
18			S	
19				R
20				D
21				E
22				Q
23				H
24				S

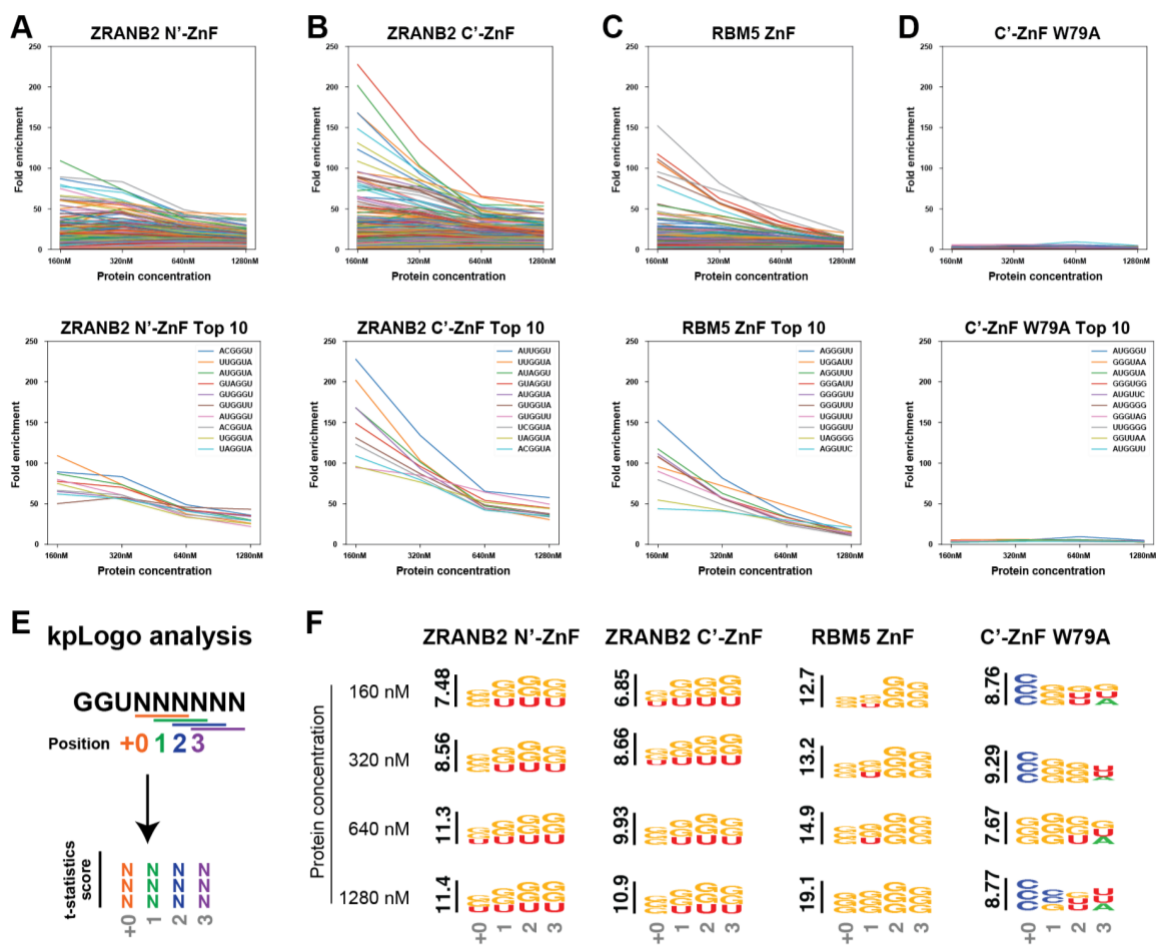


Figure 1.2 Benchmarking of the modified RNA bind-n-seq assay

(A-D) Fold enrichment score of all RNA 6-mers (top panel) or top 10 RNA 6-mers of ZRANB2 N'-ZnF, ZRANB2 C'-ZnF, RBM5 ZnF and ZRANB2 C'-ZnF W79A mutant. (E) KpLogo analysis schematics. (F) KpLogo results of ZRANB2 N'-ZnF, ZRANB2 C'-ZnF, RBM5 ZnF and ZRANB2 C'-ZnF W79A mutant.

Rational mutagenesis of ZRANB2 ZnF in search of novel RNA-binding motifs

Next, we attempted to rationally design ZRANB2 ZnF mutants with altered RNA sequence specificity. Based on a crystal structure of the ZRANB2 C'-ZnF bound to a 5'-AGGUAA-3' RNA¹⁵, we identified all protein side-chains within hydrogen bonding distance of the RNA bases in the core GGU motif (**Figure 1.1**). Because we had observed higher expression and better resistance to proteolysis during purification with the 2-ZnF array containing the ZRANB2 N'-ZnF compared to the C'-ZnF (**Figure 1.1A**), we chose to engineer the N'-ZnF. We designed a panel of 18 single mutants of the ZRANB2 N'-ZnF with one of three residues – R27, R28, or N32 – mutated to a different residue with altered hydrogen-bonding characteristics and/or side-chain length (D, E, N, R, Q, H, or S; **Table 1.1**). All mutants could be expressed and purified equivalently to the unmutated 2-ZnF array (**Supplemental Figure 2**), indicating that the mutations do not affect folding or stability of the protein. We then performed RBNS on each protein with the RNA pool containing the 6-base random window, at four protein concentrations.

Of the 18 single mutant ZnFs, six (R28E, R28Q, R28H, R28S, N32D, N32E) showed almost no enrichment of any RNA sequences at any tested protein concentration, indicating that these mutants completely lost the ability to bind RNA (**Supplemental Figure 1.3A-F**). Five of the 18 (R27D, R27E, R28D, R28N, N32S) showed limited enrichment only at the highest tested protein concentrations, indicating that these mutants retain only residual RNA binding activity (**Supplemental Figure 1.3G-K**). A further five mutants (R27N, R27Q, R27H, R27S, N32Q) showed some level of specificity as demonstrated by significant fold-enrichment of preferred motifs at 320 nM and 640 nM protein concentrations (**Figure 1.3A**). R27N and R27Q mutants showed enrichment values for preferred motifs at 1280 nM equivalent to that observed with the wild-type N'-ZnF at 160 nM, indicating that these mutants retain sequence-specific binding but

show overall lower RNA binding affinity than the wild-type ZnF. Finally, two mutants (N32R and N32H, **Figure 1.3B**) showed high fold-enrichment for preferred motifs at the lowest protein concentrations, and KpLogo analysis showed that these mutants showed an altered RNA motif preference of GGG, in contrast to the wild-type ZnF's preferred GGU motif. Altered preference for the third base in the core motif is consistent with the location of this residue within hydrogen bonding distance of the "U" base in the crystal structure of the ZRANB2 C'-ZnF bound to RNA (**Figure 1.1B**)¹⁵. Our analysis of single mutants showed that the N32R mutant shows an altered motif preference (GGG rather than GGU) while retaining overall high RNA binding affinity, as indicated by high fold-enrichment of preferred motifs at low protein concentrations.

We next designed double mutants of the N'-ZnF in search of further alternation of the motif preference, involving N32 and a nearby residue, N22, mutated to different combinations of D, E, R, Q, H, or S. Based on the location of these two residues near the 3rd RNA base in the core motif, we reasoned that double mutants of N32 and N22 might further alter the RNA sequence preference at the third base of the motif. Although N32D, N32E and N32S appeared to disrupt RNA binding and selectivity in the context of a single mutant, we could not rule out the possibility of a cooperative mutation pair at both N32 and N22 that might recover RNA binding. We designed and purified a panel of 20 double mutants in the 2-ZnF array (**Table 1.2**) and assayed each by RBNS. Using our modified RBNS assay, we found that 12 of 20 tested double mutants showed compromised RNA binding (**Supplemental Figure 1.4**). The other 8 mutants exhibited variable effects on RNA binding and motif, yet none of them further changed the RNA binding motif from GGG (**Supplemental Figure 1.5; Table 3**). All tested double mutants bearing N32R retained various levels of RNA binding affinity, yet most of them still preferred to bind GGG.

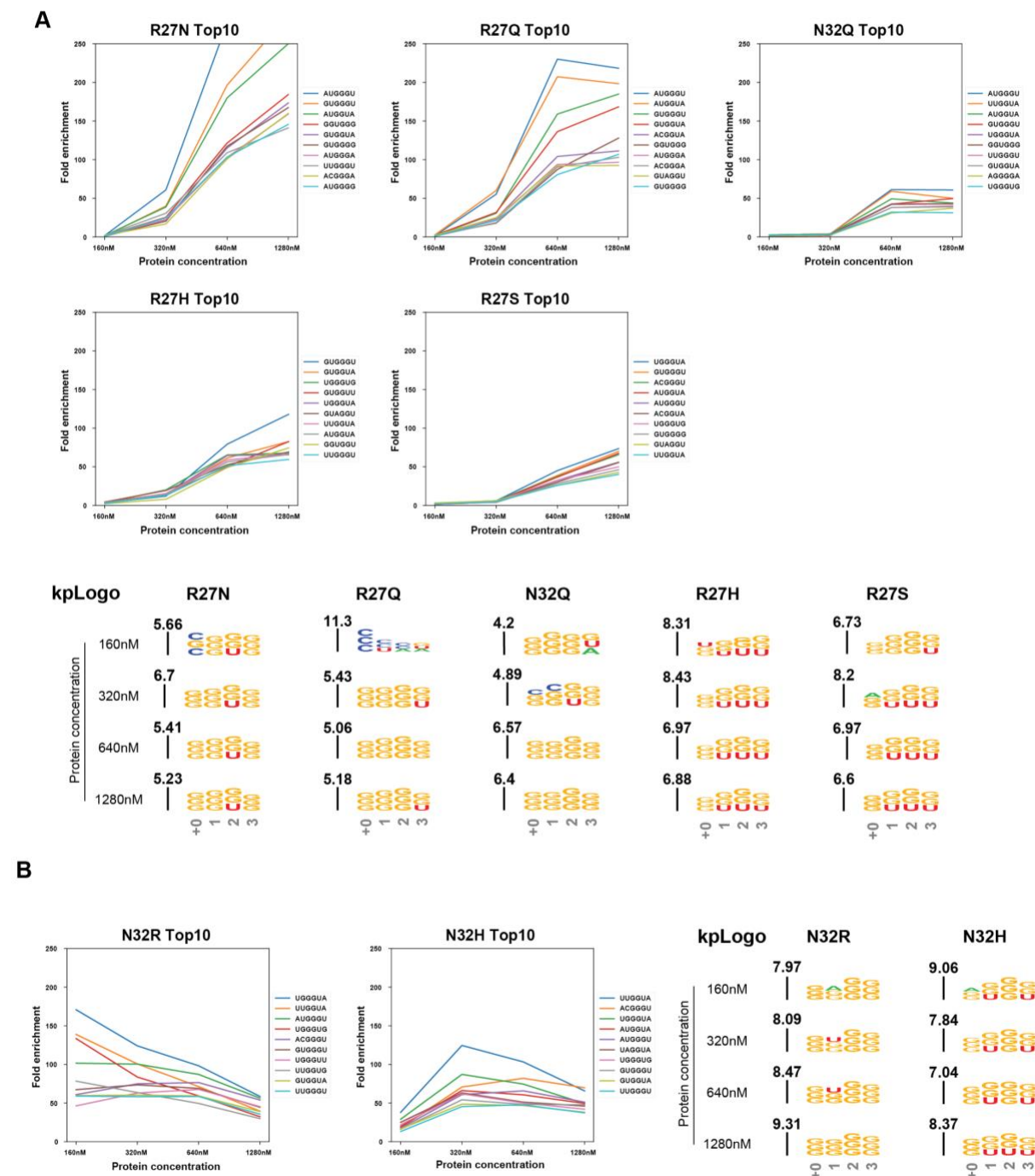


Figure 1.3 RNA bind-n-seq results of selective ZnF single mutants

(A-E) Top10 enriched RNA 6-mers of R27N, R27Q, N32Q, R27H, and R27S. This group of single mutants showed overall lower RNA binding affinity than the wild-type ZnF despite retaining some degree of specificity. (F) KpLogo results of R27N, R27Q, N32Q, R27H, and R27S. (G-H) Top10 enriched RNA 6-mers of N32R and N32H. This group of single mutants displayed RNA binding specificity and retained RNA binding affinity. (I) KpLogo results of N32R and N32H

Table 1.2 List of double mutants of ZnF

Double Mutant No.	N22	N32
1	R	R
2	R	H
3	D	R
4	D	H
5	E	R
6	E	D
7	E	E
8	E	Q
9	E	H
10	E	S
11	Q	R
12	Q	H
13	H	R
14	H	H
15	S	R
16	S	D
17	S	E
18	S	Q
19	S	H
20	S	S

Table 1.3 RNA binding profile of RBNS of all double mutants

ZnF double mutant		Mutant RNA binding profile	
N22	N32	Affinity	Preferred motif
R	R	similar	GGU
R	H	lower	GGG
D	R	lower	GGG
D	H	N/A	N/A
E	R	lower	GGG
E	D	N/A	N/A
E	E	N/A	N/A
E	Q	N/A	N/A
E	H	N/A	N/A
E	S	N/A	N/A
Q	R	lower	GGG
Q	H	N/A	N/A
H	R	similar	GGG
H	H	lower	GGG
S	R	lower	GGG
S	D	N/A	N/A
S	E	N/A	N/A
S	Q	N/A	N/A
S	H	N/A	N/A
S	S	N/A	N/A

All-atom MD simulations revealed ZnF-RNA recognition mechanism

Our RNA binding data shows that the single point mutation N32R alters the preferred RNA binding motif of the ZRANB2 N'-ZnF from GGU to GGG. To gain a molecular understanding of the change of RNA recognition pattern, we performed all-atom molecular dynamics (MD) simulations using AMBER14^{20,21,22}. Since the structure of RNA-bound ZRANB2 N'-ZnF bound to RNA has not been determined, we built a model homologous of this complex based on the known structure of the C'-ZnF (48% identical to the N'-ZnF) bound to

RNA (**Figure 1.1A**). The N32R mutation in the N'-ZnF corresponds to N86R in the C'-ZnF; to avoid confusion we refer to this mutation simply as “NRmut”. To investigate the differences in protein-RNA interaction for both WT and NRmut ZnFs, we designed four models (WT + GGU RNA, WT + GGG RNA, NRmut + GGU RNA, NRmut + GGG RNA) and performed 50-nanosecond (ns) simulations in triplicate for each model. Root mean squared distance (RMSD) analysis indicated that the models remained stable throughout the simulations in all replicates (**Supplemental Figure 1.6**).

We first investigated the effects of the ZnF mutation on the overall structural flexibility of the protein-RNA complex. We examined root mean squared fluctuation (RMSF) of the protein and RNA backbone for all four models (**Figure 1.4A**). The protein backbone dynamics were not affected by the N-to-R mutation and remained similar when the ZnF was in complex with either GGU or GGG RNA. However, the GGU RNA displayed a higher overall flexibility in complex with the NRmut, in contrast to the GGG RNA. This indicates the NRmut-GGU RNA binding is less stable compared to NRmut-GGG RNA binding. In addition, we observed the highest variation of RMSF in both the protein and RNA backbone across replicates in the NRmut-GGU RNA complex (**Supplemental Figure 1.7A**), corroborating our speculation.

We then sought to understand the hydrogen bonding interactions between the ZnF and the bound ssRNA, which play essential roles in specificity to intermolecular interactions²³. We calculated the hydrogen-bond relative intensity score based on the fraction of time a given protein-RNA hydrogen bonding interaction existed (determined and tracked by CPPTRAJ²⁴) over the course of each 50-ns simulation. The fraction of time for each hydrogen-bond donor and acceptor atom pair are summed for every combination of protein and RNA residues, defining the “hydrogen-bond relative intensity” of each interacting protein + RNA residue pair (**Figure 1.4B**;

Supplemental Figure 1.7B). In the WT + GGU RNA simulations, we observe long-lived hydrogen bonding interactions between RNA base G1 and protein residues Asp68 and Arg81; between RNA base G2 and protein residues Val77 (backbone), Trp79, and Arg82; and RNA base U3 and protein residues Asn76, Asn86. However, when the uridine (U3) was mutated to a guanosine (G3), the hydrogen bonds of all nucleobase-amino acid sidechain pairs were weakened by an average of 14.2% (**Supplemental Figure 1.8**), indicating the weaker interaction of WT ZnF with a GGG 3-mer compared to GGU. Intriguingly, the asparagine to arginine mutation in NRmut completely disrupted interaction of this residue with RNA base U3 in the NRmut + GGU model (**Figure 1.4B**, bottom left panel), indicating the loss of specific recognition of GGU motif by NRmut. In contrast, the NRmut + GGG RNA simulation showed a restored hydrogen-bond interaction between the newly-introduced arginine and RNA base G3, along with an additional interaction with RNA base G2. Besides the restored hydrogen bonding between the protein and the 3rd RNA base, the total hydrogen bond intensity was also restored to levels equivalent to those in the WT ZnF + GGU RNA simulations (**Supplemental Figure 1.8**), consistent with our experimental results that the NRmut prefers to bind the RNA motif GGG over GGU.

Finally, we investigated the molecular mechanism underlying the altered RNA binding preference in NRmut by examining the protein-RNA binding conformations in the MD simulation trajectories. Hierarchical clustering derived five representative conformations from each 50-ns simulation and the average model of the major conformation (>50% of the simulation time) was visualized (**Methods**). Structural overlays of the average model of the major conformation (>50% of the simulation time) in the ZnF WT and NRmut in complex with GGU RNA showed a high degree of overlap of the RNA-protein complex backbone (**Figure 1.4C**). As

expected, the closest N-O distance between the newly introduced Arginine residue in NRmut and the RNA base U3 (5.1 Å) is much larger than the median distance between amide C=O acceptors and N-H donors in a typical hydrogen bond (~ 2.9 Å)²⁵(**Figure 1.4C-D**), indicating loss of hydrogen bond interactions. Strikingly, structural overlays of representative structures of ZnF NRmut + GGU RNA and NRmut + GGG RNA revealed a significant displacement of the protein backbone (**Figure 1.4E**). NRmut seemed to be displaced away from the GGG RNA compared to the simulations with GGU RNA. The closest N-O distance between Arg86 and G3 were measured 2.9 Å, indicating the restored hydrogen bonding for NRmut with GGG RNA in contrast to GGU RNA. The results showed that the NRmut likely adapts to GGG RNA binding by the backbone displacement and rotation, despite the interaction pair changes from short amino acid side chain–small nucleobase (Asn–Uridine) to long amino acid side chain–bulky nucleobase (Arg–Guanine). In summary, the MD simulations rationalize how a single amino acid substitution (from asparagine to arginine) can alter the RNA motif recognition by an RNA-binding ZnF, thereby deepening our understanding of the mechanisms involved in protein-RNA recognition.

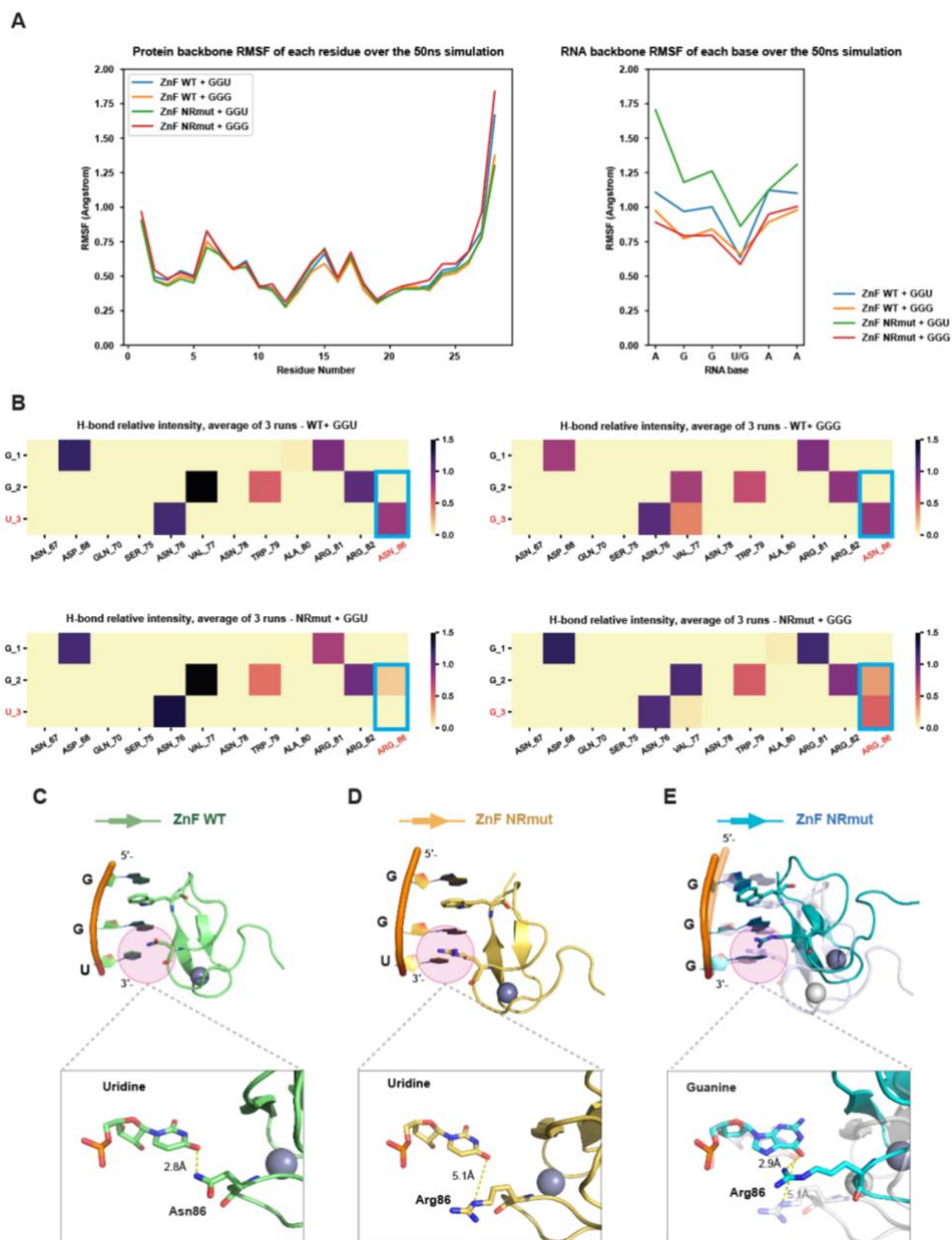


Figure 1.4 Molecular dynamics simulation of ZnF-RNA complexes

(A) Root mean squared fluctuation (RMSF) of each protein residue and RNA base during the simulation (average of 3 replicates of 50-ns simulations) of in 4 models (ZnF WT + GGU, ZnF WT + GGG, ZnF NRmut + GGU, ZnF NRmut + GGG). (B) Heat map of H-bond relative intensity between the three RNA bases and the RNA-contacting residues during the simulation. (C-E) Representative structures of ZnF-RNA complexes during the simulations. C, the representative structure of ZnF WT + GGU (green); D, the representative structure of ZnF NRmut + GGU (yellow); E, the representative structure of ZnF NRmut + GGG (cyan) overlaying with the representative structure of ZnF NRmut + GGU (white). The protein residues highlighted in the top panel are Arg86/Asn86 and Trp79.

Chapter 1.3 Conclusion

Here we show that rational design of single mutations in RNA-contacting residues can alter the sequence specificity for an RNA-binding ZnF without compromising RNA binding affinity, opening the door to further rational design or screening of mutants with further alterations in sequence specificity. Our modified RBNS assay robustly profiles both sequence specificity and overall RNA binding affinity, and represents a promising avenue for high-throughput screening of ZnF mutants. Our MD simulations show that the NRMut ZnF likely adapts to its preferred GGG motif through both main-chain and side-chain movement, revealing a high degree of flexibility in this family of RNA-binding proteins. Our work has defined a set of key interactions that define the specificity for the third RNA base in the ZRANB2 ZnF preferred motif; we anticipate that future work will identify interactions that define specificity for the first and second bases of this motif, enabling rational design of individual ZnFs and ZnF arrays for high-specificity RNA targeting.

Chapter 1.4 Methods

Purification of ZRANB2 ZnF and mutants

The DNA sequences of N' and C' ZnF in human ZRANB2 (NP_005446.2) were codon-optimized for expression in *E. coli* strain. A 2-ZnF array of ZnF_WT-ZnF* was cloned into a modified UC Berkeley Macrolab vector 2CT (Addgene #29706) to generate an N-terminal fusion to a His₆-MBP (Maltose-Binding Peptide) tag with a GSGSG-linker in-between the ZnFs. A Streptavidin-Binding Peptide (SBP)-tag was cloned at the N-terminal of the 2-ZnF array. The His-MBP-SBP-ZnF_WT-ZnF* protein was expressed in *E. coli* strain Rosetta2 pLysS (EMD Millipore) by growing cultures in 2xYT media to mid-log phase at 37°C, followed by induction with 0.25 mM IPTG at 18°C for 16 hours supplemented with 100 µM ZnCl₂. For protein purification, cells were harvested by centrifugation, suspended in resuspension buffer (20 mM Tris-HCl pH 7.5, 500 mM NaCl, 20 mM imidazole, 2 mM β-Mercaptoethanol and 10% glycerol) and lysed by sonication. Lysates were clarified by centrifugation (16,000 rpm 30 min), then the supernatant was loaded onto a 1mL Ni²⁺ affinity column (HisTrap HP, Cytiva) pre-equilibrated with the resuspension buffer. The column was washed with a buffer containing 20 mM imidazole and 500 mM NaCl for 10 column volume (CV), followed by loading 0.5mg MBP-TEV protease (Tobacco Etch Virus nuclear-inclusion-a endopeptidase) onto the HisTrap column for on-column His-MBP tag cleavage. The MBP-TEV protease was expressed and purified from pKR1043 (Addgene #8835) as described²⁷. After 4-hour TEV cleavage at room temperature, the His-MBP tag free protein was washed by 10 CV with TEV wash buffer (20mM Tris pH7.5, 150mM KCl, 2mM β-ME, 10% glycerol). The protein was then concentrated by ultrafiltration, aliquoted and frozen at -80°C for future use. For His-MBP tagged protein, no on-column TEV protease

cleavage was carried out and the protein was eluted with a buffer containing 250 mM imidazole and 150 mM KCl.

RNA-bind-n-seq (RBNS)

RBNS experiments with an RNA pool containing a randomized window were performed as indicated in literature^{13,14}. In brief, randomized RNA oligonucleotides (6 nt) containing a sequence of “AAAGGUNNNNNNAAA” flanked by adaptor sequences were synthesized through in vitro transcription from synthesized DNA libraries (Integrated DNA Technologies). The pool of RNA was incubated with the SBP-tagged recombinant 2-ZnF-array at five concentrations of protein (160, 320, 640, 1280 nM), while keeping RNA concentrations constant at 1 μ M. RNA-protein complexes were immobilized with streptavidin-conjugated affinity magnetic resin and were subjected to multiple washing steps. Then the RNA was eluted and prepared for deep sequencing. 0.2-0.5 million reads per RBP pulldown concentration were targeted to ensure more than 50-fold coverage for each unique RNA sequence (4096 sequences x 50 fold coverage = 204,800 reads minimum per sample). The original pool of RNA was also sequenced with a similar depth as the input library.

RBNS data processing analysis

The fold enrichment of each RNA 6-mer was calculated as the frequency of each 6-mer in the pulldown library reads divided by its frequency in the input library. Fold enrichments were ranked and plotted using Matplotlib in python. To identify the core 3-mer motif from the enriched 6-mers, kpLogo¹⁹ was used to report the 3-mer enrichment at each position using student's t-test. The command “kpLogo weighted_RBNS_file.txt -weighted -k 3 -plot s -fontsize 50” was used to generate kpLogo output, with 6-mer RNA sequences weighted by their fold enrichments as input. K-means clustering of the RNA 6-mers was performed using the Scikit-

learn package in python, with the fold enrichment of each sequence as the features. Number of clusters k was set to 4 and k cluster centroids were initialized using k-means++ and a maximum of 500 iterations were performed. K-mean clustering code: `KMeans(n_clusters=4, init="k-means++", _init=50, _iter=500, random_state=42)`

All-atom Molecular Dynamics (MD) simulation

The ZnF-RNA complexes were constructed based on crystal structure (PDB: 3g9y). All crystallographic waters were preserved. The NRmut and the RNA mutant (AGGUAA → AGGGAA) were prepared by inducing virtual mutations to the ZRANB2-C'-ZnF-RNA structure using the mutagenesis plugin available in PyMOL. Protein and RNA atoms were represented using the AMBER^{20,21} ff14SB force field²⁸ and the OL3 parameters²⁹, respectively. Zinc chelation was represented using Zinc AMBER Force Field (ZAFF)²². The protein and RNA are solvated in cubic TIP3P water box with 10 Å distance from the solute to the boundaries of the simulation box. Hydrogen atoms were added through LeaP program in AMBER. Varying number of Na⁺ ions were added to neutralize the system. The simulation cell was then replicated infinitely in three dimensions to impose periodic boundary conditions. All MD simulations were performed under periodic boundary conditions using the CUDA accelerated version of PMEMD implemented in Amber14^{20,21} suite of programs on TSCC²⁶. The structures were first relaxed using a combination of steepest descent (10000 steps, 2 fs each) and conjugate gradient minimization (10000 steps, 2 fs each). The lengths of bonds involving hydrogen atoms were constrained by SHAKE. The particle mesh Ewald method (PME) was employed to treat the long-range electrostatics. The non-bonded cutoff for the long-range interactions was set to 10 Å. Heating was conducted from 0 K to 300 K in 800 ps, using Langevin dynamics with a collision frequency of 2 ps⁻¹ as temperature control, followed by pre-equilibrium with 800-1000 ps at

same condition of heating. Another pre-equilibrium was performed in an NPT (constant number of atoms, pressure, and temperature) ensemble at 300K and 1 atm, using Langevin dynamics for temperature regulation and a Berendsen barostat for pressure control. No restraints were applied since this stage. A following 10 ns equilibrium steps with similar conditions as pre-equilibrium except that Monte Carlo barostat was used for pressure control. A 5-ns of post-equilibrium run was conducted before the production simulations and excluded from the analysis. Subsequently, three replicates of 50-ns unbiased MD simulations were carried out for the ZnF WT + GGU RNA, ZnF WT + GGG RNA, ZnF NRmut + GGU RNA, ZnF NRmut + GGG RNA complexes.

MD simulation data analysis

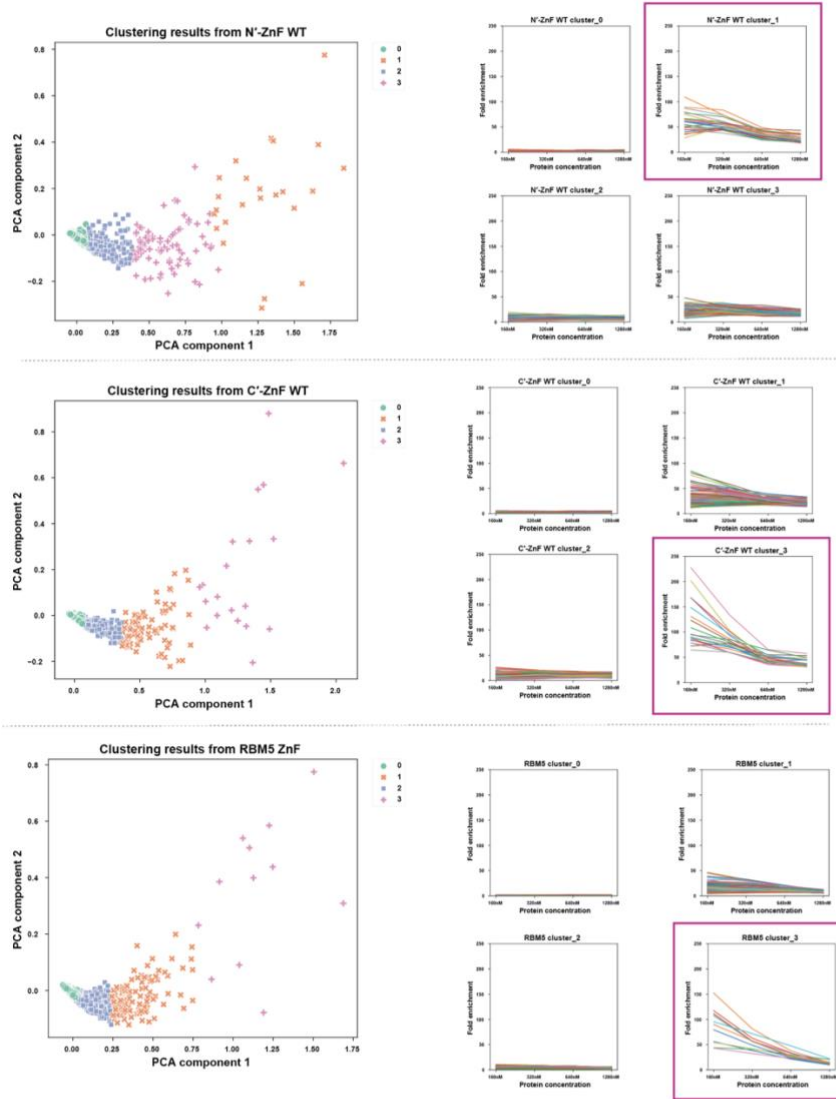
The CPPTRAJ²⁴ module was used to analyze all the MD trajectories. The root mean squared distance (RMSD) and the root mean squared fluctuation (RMSF) of the ZnF-RNA complexes from each MD trajectory were calculated, with respect to the C α atoms of the protein backbone and the C3' and C4' atoms of the RNA backbone in reference of the final structure of the post-equilibrium run. Hydrogen bonds between residues were defined and measured by the CPPTRAJ package. The visualization of the MD trajectories was rendered using Visual Molecular Dynamics (VMD)³⁰, and data were plotted using Matplotlib. Hierarchical clustering was performed after stripping water and ions and based on the protein and RNA backbone positions excluding Hydrogen atoms (command: cluster c1 hieragglo epsilon 3.0 clusters 5 averagelinkage rms :1-36@C,N,O,CA,P,O3',O5',C3',C4',C5'&!.@H=). The average model of the top #1 representative cluster was used for overlay. PyMOL was used for structure visualization and overlay.

Acknowledgements

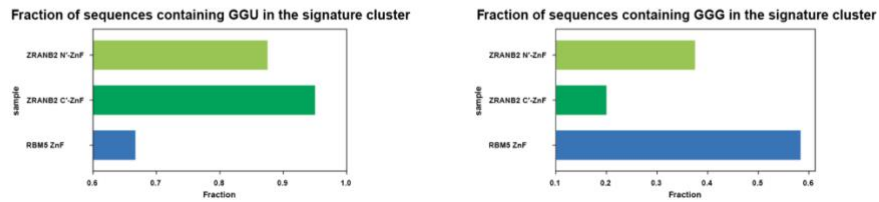
Chapter 1, in full, is currently being prepared for submission for publication: Liang, Qishan; Xiang, Shengnan J.; Corbett, Kevin D.; Yeo, Gene W.. “*Engineering and characterization of the Sequence Specificity of an RNA-Binding Zinc Finger Protein*”. The dissertation author was the primary investigator and author of this paper.

The authors thank Dr. Kartik L. Rallapalli for the inspiration and guidance in the Molecular Dynamics simulation set-up. The authors acknowledge support from the National Institutes of Health (R35 GM144121 to K.D.C., R01 HG004569, HG009889, HG011864, NS103172 and MH131907 to G.W.Y.). Q.L. was supported by an individual predoctoral fellowship from the American Heart Association. This publication includes data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from a National Institutes of Health SIG grant (#S10 OD026929) and data generated at the Sequencing Core Facility, La Jolla Institute utilizing an Illumina NovaSeq 6000 that was acquired through the Shared Instrumentation Grant (SIG) Program (#S10OD025052). Computational resources were provided by the Department of Defense High Performance Computing Modernization Program (HPCMP) and the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center (SDSC)²⁶. We would also like to thank the Stem Cell Genomics Core at the Sanford Stem Cell Institute for providing sequencing services.

A



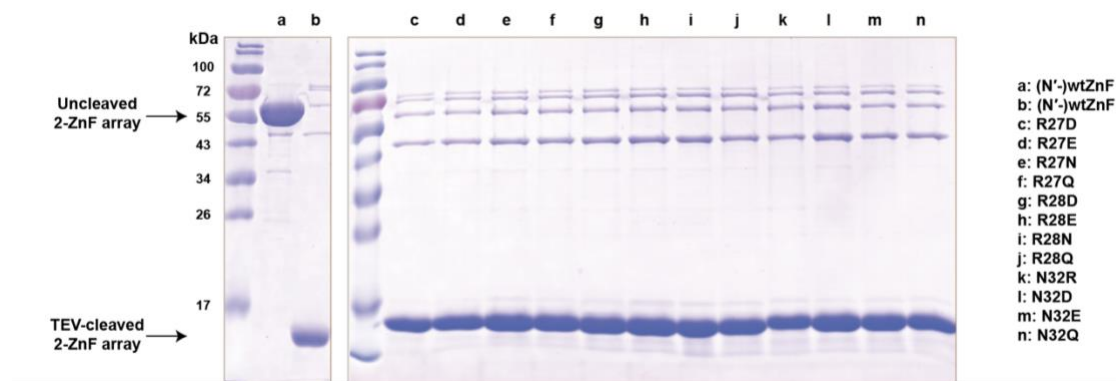
B



Supplemental Figure 1.1 Benchmarking of the modified RNA bind-n-seq (RBNS) assay

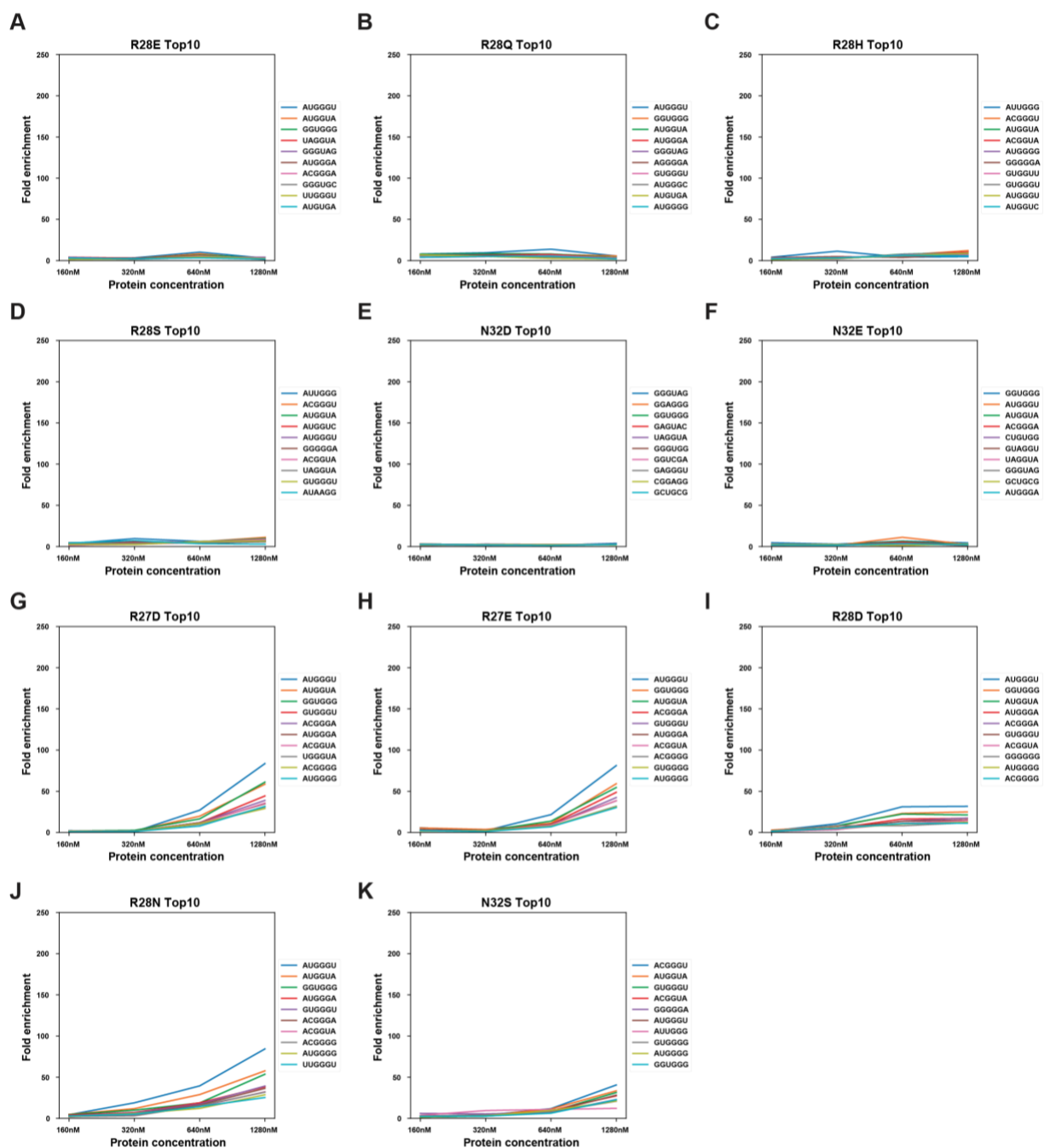
(A) K-mean clustering of RNA 6-mers based on their fold enrichment in the benchmarking RBNS experiments. Red box: signature cluster, defined as the RNA 6-mers with distinct binding affinities to the protein at low protein concentration.

(B) Fraction of sequences containing GGU or GGG in the signature cluster in the benchmarking RBNS experiments.

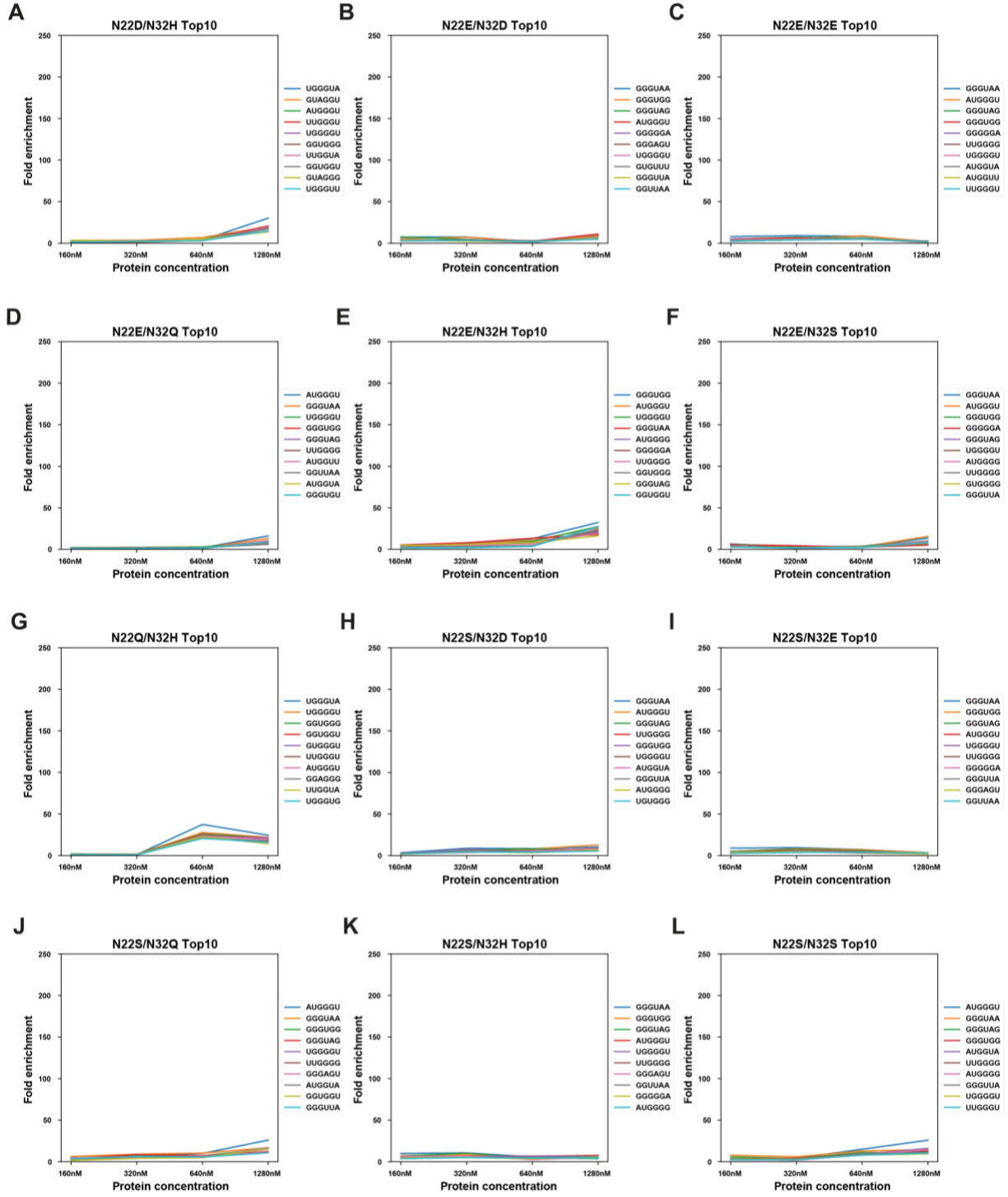


Supplemental Figure 1.2 Purification of ZnF-array

Representative SDS-PAGE gel image of the high-throughput purification of ZnF arrays.

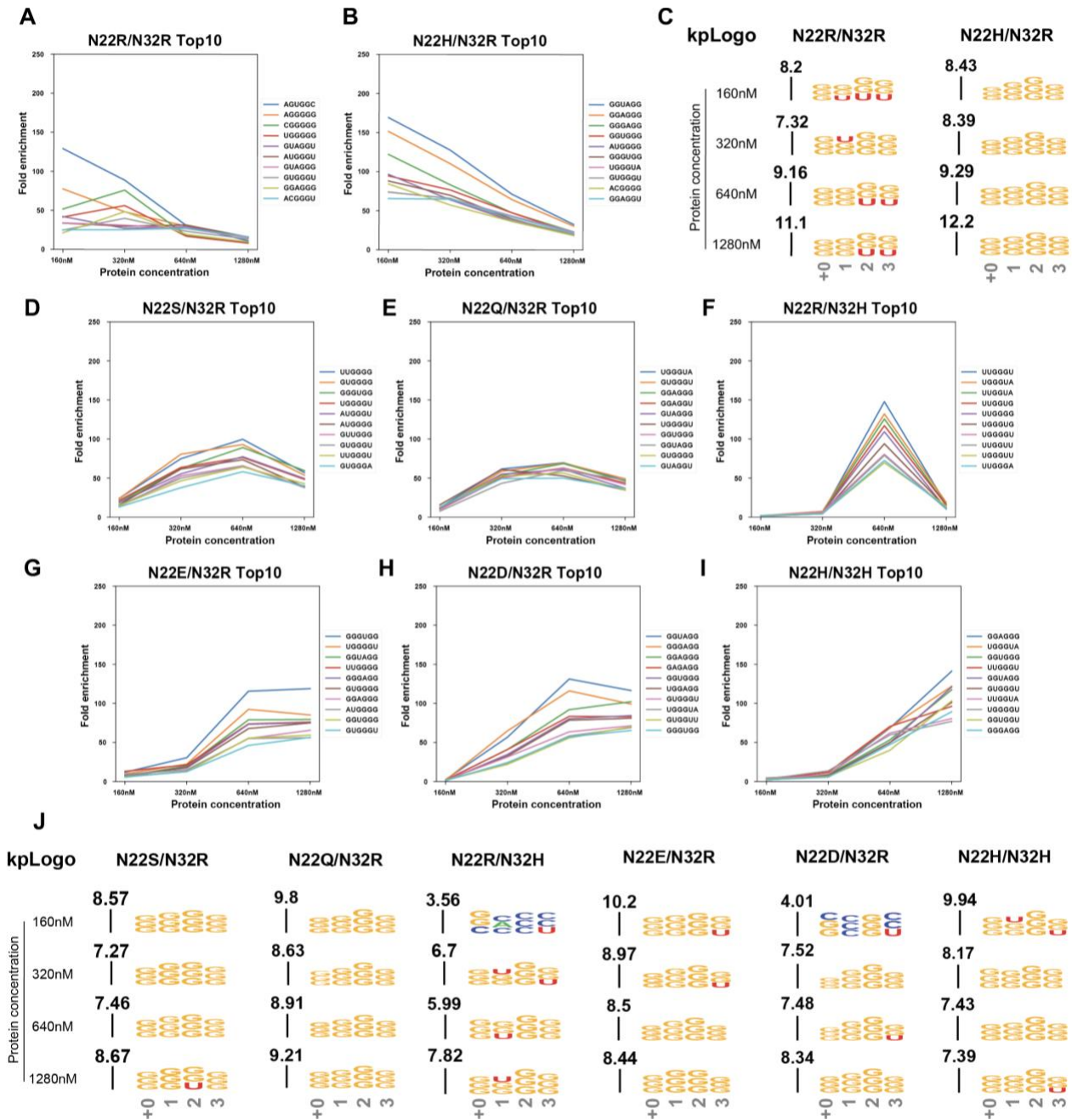


Supplemental Figure 1.3 Top 10 6-mers in RNA bind-n-seq of other ZnF single mutants (A-K) Top10 enriched RNA 6-mers of ZnF single mutant R28E, R28Q, R28H, R28S, N32D, N32E, R27D, R27E, R28D, R28N, and N32S



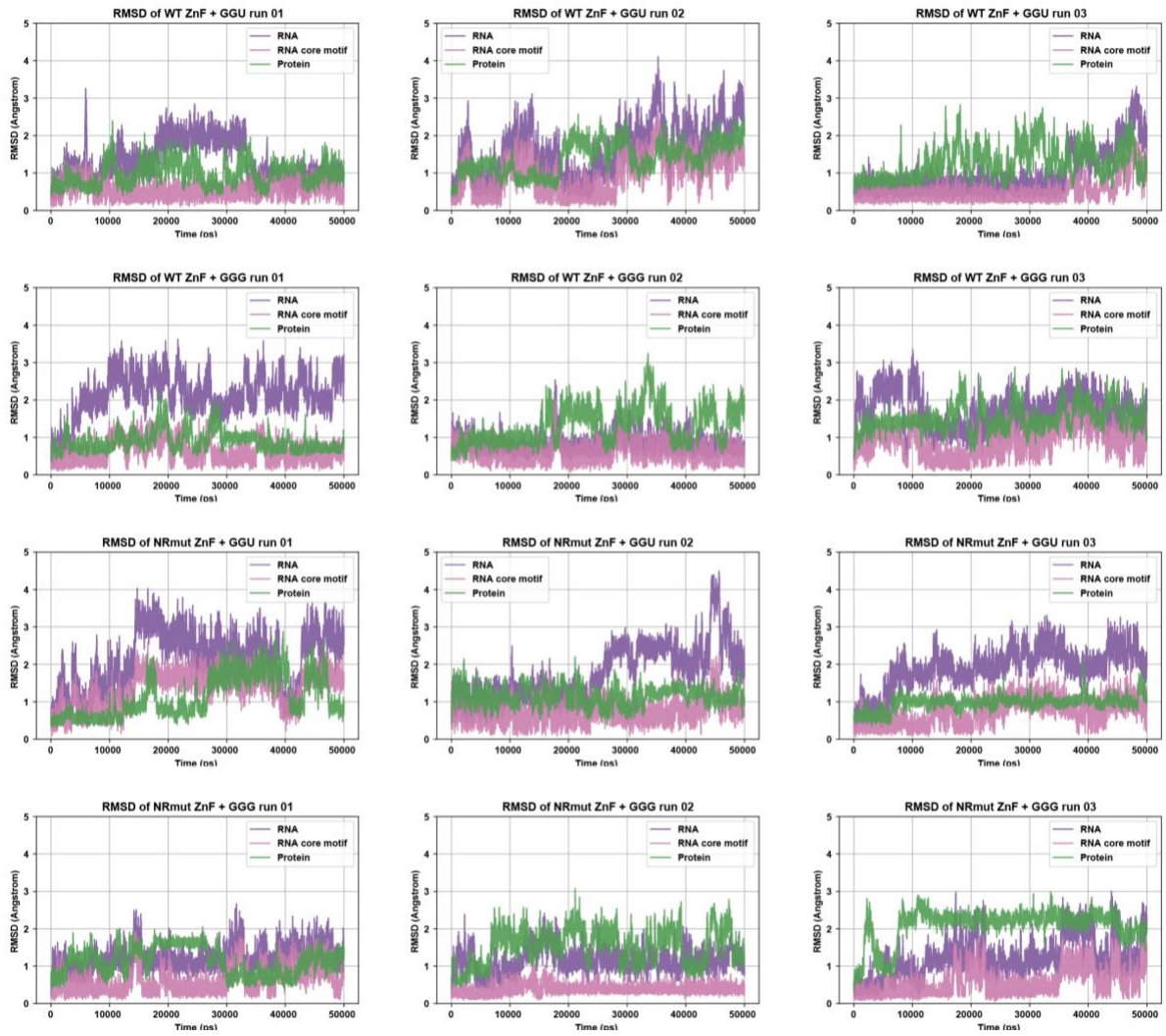
Supplemental Figure 1.4 Top 10 6-mers in RNA bind-n-seq of ZnF double mutants with compromised binding affinity

(A-L) Top10 enriched RNA 6-mers of ZnF double mutant N22D/N32H, N22E/N32D, N22E/N32E, N22E/N32Q, N22E/N32H, N22E/N32S, N22Q/N32H, N22S/N32D, N22S/N32E, N2S/N32Q, N22S/N32H, and N22S/N32S.



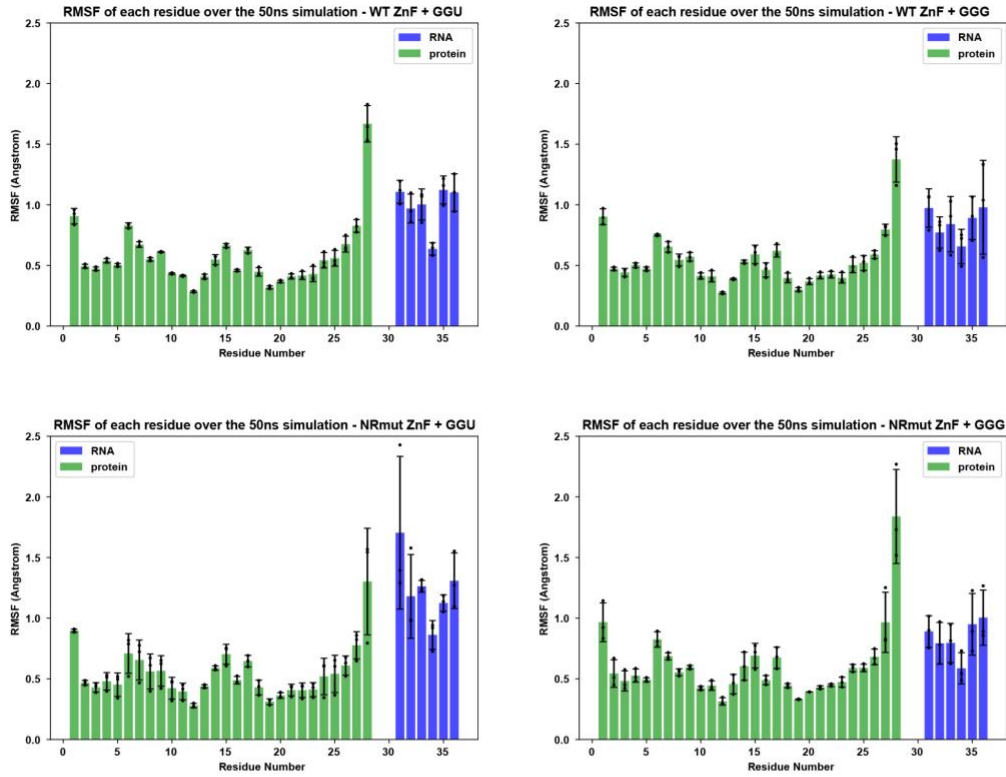
Supplemental Figure 1.5 RNA bind-n-seq results of selective ZnF double mutants

(A-B) Top10 enriched RNA 6-mers of N22R/N32R and N22H/N32R. This group of double mutants displayed RNA binding specificity to GGU/GGG and retained RNA binding affinity. (C) KpLogo results of N22R/N32R and N22H/N32R. (D-I) Top10 enriched RNA 6-mers of N22S/N32R, N22Q/N32R, N22R/N32H, N22E/N32R, N22D/N32R, and N22H/N32H. This group of double mutants displayed RNA binding specificity to GGG with lower overall RNA binding affinity compared to the wild-type ZnF. (J) KpLogo results of N22S/N32R, N22Q/N32R, N22R/N32H, N22E/N32R, N22D/N32R, and N22H/N32H.

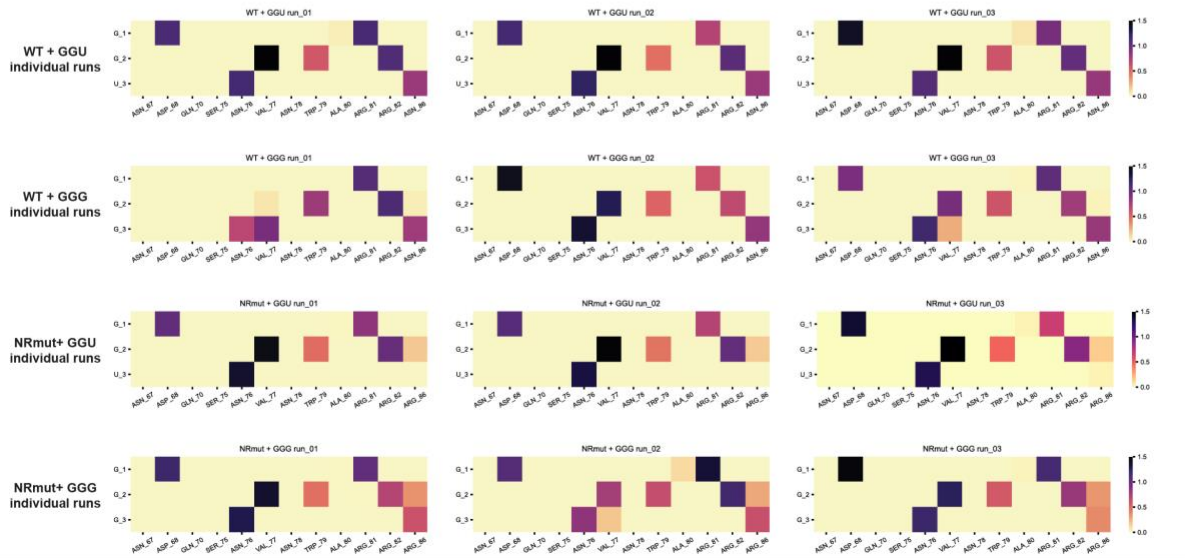


Supplemental Figure 1.6 Molecular dynamics simulation results of ZnF-RNA complexes
Root mean squared distance (RMSD) of the protein or RNA backbone throughout the simulations, in reference to the equilibrated structure.

A

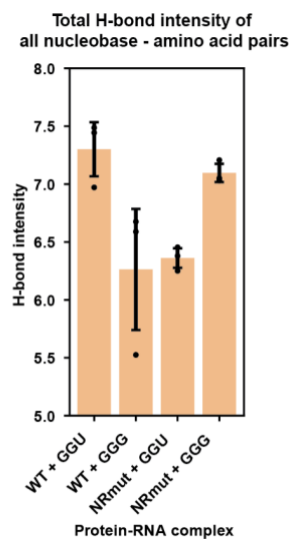


B



Supplemental Figure 1.7 Molecular dynamics simulation results of ZnF-RNA complexes (continued)

(A) Root mean squared fluctuation (RMSF) of each protein residue and RNA base during the simulation of in 4 models (ZnF WT + GGU, ZnF WT + GGG, ZnF NRmut + GGU, ZnF NRmut + GGG). Individual replicates are shown, the error bars indicate standard deviations, n=3. (B) Heat map of H-bond intensity between the three RNA bases and the RNA-contacting residues during the simulation. Individual replicates of 50-ns simulations are shown.



Supplemental Figure 1.8 Molecular dynamics simulation results of ZnF-RNA complexes (continued)

Total H-bond intensity of all nucleobase-amino acid pairs in 4 models: ZnF WT + GGU, ZnF WT + GGG, ZnF NRmut + GGU, ZnF NRmut + GGG.

REFERENCES

1. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat Rev Genet.* 2014 Dec;15(12):829–45.
2. Lauffer MC, van Roon-Mom W, Aartsma-Rus A. Possibilities and limitations of antisense oligonucleotide therapies for the treatment of monogenic disorders. *Commun Med.* 2024 Jan 5;4(1):1–11.
3. Zhao YY, Mao MW, Zhang WJ, Wang J, Li HT, Yang Y, Wang Z, Wu JW. Expanding RNA binding specificity and affinity of engineered PUF domains. *Nucleic Acids Res.* 2018 May 18;46(9):4771–82.
4. Sugimoto M, Suda A, Futaki S, Imanishi M. Effective RNA Regulation by Combination of Multiple Programmable RNA-Binding Proteins. *Applied Sciences* 2020, Vol 10, Page 6803. 2020 Sep 28;10(19):6803.
5. Konermann S, Lotfy P, Brideau NJ, Oki J, Shokhirev MN, Hsu PD. Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. *Cell.* 2018 Apr 19;173(3):665-676.e14.
6. Lau CH, Suh Y. In vivo epigenome editing and transcriptional modulation using CRISPR technology. *Transgenic Res.* 2018 Dec;27(6):489–509.
7. Pabo CO, Peisach E, Grant RA. Design and Selection of Novel Cys2His2 Zinc Finger Proteins. *Annual Review of Biochemistry.* 2001 Jul 1;70(Volume 70, 2001):313–40.
8. Beerli RR, Barbas CF. Engineering polydactyl zinc-finger transcription factors. *Nat Biotechnol.* 2002 Feb;20(2):135–41.
9. Isalan M, Choo Y. Engineering Nucleic Acid-Binding Proteins by Phage Display. In: Moss T, editor. *DNA-Protein Interactions: Principles and Protocols.* Totowa, NJ: Humana Press; 200. p. 417–29. Available from: <https://doi.org/10.1385/1-59259-208-2:417>
10. Nguyen CD, Mansfield RE, Leung W, Vaz PM, Loughlin FE, Grant RP, Mackay JP. Characterization of a family of RanBP2-Type zinc fingers that can recognize single-stranded RNA. *Journal of Molecular Biology.* 2011;407(2):273–83.
11. O'Connell MR, Vandevenne M, Nguyen CD, Matthews JM, Gamsjaeger R, Segal DJ, Mackay JP. Modular assembly of RanBP2-Type zinc finger domains to target single-stranded RNA. *Angewandte Chemie - International Edition.* 2012;51(22):5371–5.
12. De Franco S, Vandenameele J, Brans A, Verlaine O, Bendak K, Damblon C, Matagne A, Segal DJ, Galleni M, Mackay JP, Vandevenne M. Exploring the suitability of RanBP2-type Zinc Fingers for RNA-binding protein design. *Scientific Reports.* 2019;9(1):1–13.

13. Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB. RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Molecular Cell*. 2014 Jun 5;54(5):887–900.
14. Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, Bazile C, Lambert NJ, Van Nostrand EL, Pratt GA, Yeo GW, Graveley BR, Burge CB. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular Cell*. 2018 Jun 7;70(5):854-867.e9.
15. Loughlin FE, Mansfield RE, Vaz PM, McGrath AP, Setiyaputra S, Gamsjaeger R, Chen ES, Morris BJ, Guss JM, Mackay JP. The zinc fingers of the SR-like protein ZRANB2 are single-stranded RNA-binding domains that recognize 5' splice site-like sequences. *Proceedings of the National Academy of Sciences*. 2009 Apr 7;106(14):5581–6.
16. Shepard PJ, Hertel KJ. The SR protein family. *Genome Biology*. 2009 Oct 27;10(10):242.
17. Nguyen CD, Mansfield RE, Leung W, Vaz PM, Loughlin FE, Grant RP, Mackay JP. Characterization of a Family of RanBP2-Type Zinc Fingers that Can Recognize Single-Stranded RNA. *Journal of Molecular Biology*. 2011 Mar 25;407(2):273–83.
18. Soni K, Jagtap PKA, Martínez-Lumbreras S, Bonnal S, Geerlof A, Stehle R, Simon B, Valcárcel J, Sattler M. Structural basis for specific RNA recognition by the alternative splicing factor RBM5. *Nat Commun*. 2023 Jul 15;14(1):4233.
20. Case DA, Aktulga HM, Belfon K, Ben-Shalom IY, Berryman JT, Brozell SR, Cerutti DS, Cheatham TE III, Cisneros GA, Cruzeiro VWD, Darden TA, Forouzesh N, Giambasu G, Giese TJ, Gilson MK, Gohlke H, Goetz AW, Harris J, Izadi S, Izmailov SA, Kasavajhala K, Kaymak MC, King E, Kovalenko A, Kurtzman T, Lee T, Li P, Lin C, Liu J, Luchko T, Luo R, Machado M, Man V, Manathunga M, Merz KM, Miao Y, Mikhailovskii O, Monard G, Nguyen H, O'Hearn KA, Onufriev A, Pan F, Pantano S, Qi R, Rahnamoun A, Roe DR, Roitberg A, Sagui C, Schott-Verdugo S, Shajan A, Shen J, Simmerling CL, Skrynnikov NR, Smith J, Swails J, Walker RC, Wang J, Wang J, Wei H, Wu X, Xiong Y, Xue Y, York DM, Zhao S, Zhu Q, Kollman PA. *Amber 2023*. University of California, San Francisco; 2023. 1016 p.
21. Case DA, Aktulga HM, Belfon K, Cerutti DS, Cisneros GA, Cruzeiro VWD, Forouzesh N, Giese TJ, Götz AW, Gohlke H, Izadi S, Kasavajhala K, Kaymak MC, King E, Kurtzman T, Lee TS, Li P, Liu J, Luchko T, Luo R, Manathunga M, Machado MR, Nguyen HM, O'Hearn KA, Onufriev AV, Pan F, Pantano S, Qi R, Rahnamoun A, Risheh A, Schott-Verdugo S, Shajan A, Swails J, Wang J, Wei H, Wu X, Wu Y, Zhang S, Zhao S, Zhu Q, Cheatham TE III, Roe DR, Roitberg A, Simmerling C, York DM, Nagan MC, Merz KM Jr. *AmberTools*. *J Chem Inf Model*. 2023 Oct 23;63(20):6183–91.
22. Peters MB, Yang Y, Wang B, Füsti-Molnár L, Weaver MN, Merz KM Jr. Structural Survey of Zinc-Containing Proteins and Development of the Zinc AMBER Force Field (ZAFF). *J Chem Theory Comput*. 2010 Sep 14;6(9):2935–47.

23. Hubbard RE, Kamran Haider M. Hydrogen Bonds in Proteins: Role and Strength. In: eLS. John Wiley & Sons, Ltd; 2010. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0003011.pub2>
24. Roe DR, Cheatham TEI. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput.* 2013 Jul 9;9(7):3084–95.
25. Bissantz C, Kuhn B, Stahl M. A Medicinal Chemist's Guide to Molecular Interactions. *J Med Chem.* 2010 Jul 22;53(14):5061–84.
26. San Diego Supercomputer Center (2022): Triton Shared Computing Cluster. University of California, San Diego. Service. 2022; Available from: <https://tritoncluster.sdsc.edu>
27. Kapust RB, Tözsér J, Fox JD, Anderson DE, Cherry S, Copeland TD, Waugh DS. Tobacco etch virus protease: mechanism of autolysis and rational design of stable mutants with wild-type catalytic proficiency. *Protein Eng.* 2001 Dec;14(12):993–1000.
28. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput.* 2015 Aug 11;11(8):3696–713.
29. Cheatham III TE, Case DA. Twenty-five years of nucleic acid simulations. *Biopolymers.* 2013;99(12):969–77.
30. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics.* 1996 Feb 1;14(1):33–8.

Chapter 2 INSCRIBE enables highly sensitive *in situ* capture of endogenous RNA-protein interactions in fixed tissues

Chapter 2.1 Introduction

As pivotal players in the life cycle of mRNAs and other functional RNAs, RNA-binding proteins (RBPs) are essential for RNA modification, splicing, stability, localization, translation, and many other post-transcriptional regulatory events¹⁻³. Dysregulation of RBPs can give rise to many diseases including neurodegenerative disorders and cancer⁴⁻⁶ due to disrupted RBP-RNA interactions and consequent improper RNA processing. Therefore, mapping RBP-RNA interaction profiles across the transcriptome is crucial to unravel RBP functions and understand RNA-related diseases.

Extensive efforts have been undertaken to characterize RBP-RNA interactions at the molecular and transcriptomic levels. The most widely-used RBP-RNA interaction profiling approaches are immunoprecipitation (IP)-based technologies such as CLIP-seq (Cross-Linking and Immunoprecipitation followed by sequencing)^{7,8,9,10}, which have significantly advanced our understanding of RBP-RNA interactions in various biological contexts. However, these methods are laborious and require substantial starting materials due to the loss of RNA during immunoprecipitation and low cross-linking efficiency¹¹. These drawbacks preclude the capture of RBP-RNA interactions in precious primary tissue samples and clinical samples. Additionally, the RNA digestion in CLIP-seq protocols obfuscates the distinction between RNA isoforms, which may have different regulatory functions in RNA metabolism.

Recently developed tools such as STAMP (Surveying Targets by APOBEC-Mediated Profilin¹²) and TRIBE (Targets of RNA-binding proteins Identified By Editin¹³) address the limitations of IP-based profiling methods by using RNA modifying enzymes to label RNA bases in proximity to an RBP of interest in living cells. Without IP enrichment or RNA digestion steps, these tools can identify RNA isoforms with minimal input materials, including at the single-cell level¹². However, both methods rely on overexpression of an exogenous RBP-enzyme fusion protein in living cells, which could lead to artifacts in the RBP's localization, functionality, and RNA-binding characteristics. While informative, the resulting RBP-RNA interaction profiles may not accurately reflect true RBP-RNA interaction profiles in endogenous biological environments. In addition, non-engineerable tissues such as clinical biopsy samples are not amenable to interrogation via these methods.

Here we present INSCRIBE (I*n situ* Sensitive Capture of RNA-protein Interactions in Biological Environments), a unique and versatile solution that combines the advantages of CLIP and STAMP while circumventing their respective disadvantages, preserving an authentic RBP-RNA interaction map in undisrupted biological contexts. This novel technology harnesses a purified recombinant protein APOBEC1-nanobody that is precisely directed to the RBP of interest through the nanobody-primary antibody recognition. Through *in situ* RNA C-to-U labeling by the cytosine deaminase APOBEC1 with a simple workflow, INSCRIBE identifies transcriptomic RBP-RNA interactions in unmodified cells and primary tissues fixed by either methanol or formaldehyde, without requiring any prior plasmid constructions and transformations into cells (**Figure 2.1A**). Coupled with an established computational pipeline that maps C-to-U conversion in the transcriptome (**Figure 2.1B**), INSCRIBE enables robust identification of endogenous RBP binding sites in fixed HEK293T cells and mouse brain tissue

slices. INSCRIBE is highly specific as it readily enriches the consensus binding motifs of two RBPs (RBFOX2 and TDP-43) and agrees with orthogonal methods such as eCLIP. The technique is compatible with cDNA PacBio long-read sequencing, enabling the distinction between different RNA isoform substrates. One of the advantages of INSCRIBE is its versatility: With a single recombinant enzyme-nanobody fusion, INSCRIBE can profile diverse RBPs with commercially available primary antibodies bearing the same IgG for nanobody recognition. Another great advantage of INSCRIBE lies in its high sensitivity and low-input material requirement. While routine INSCRIBE protocol uses ~150,000 cells or a slice of primary tissue, it can capture transcriptome-wide endogenous RBP-RNA interactions with as little as 100 pg of RNA, equivalent to ~5 cells, compared to the 5-10 million cells used in a typical eCLIP experiment. Additionally, the workflow is much less laborious and time-consuming, as the INSCRIBE experiment only takes 20-24 hours prior to RNA-seq library preparation, compared to ~4.5 days in the eCLIP workflow. The robust, versatile, sensitive, and convenient INSCRIBE method is particularly beneficial when handling preserved valuable samples, such as clinical samples and patient-derived stem cells. We envision INSCRIBE's broad application to bridge the critical gap in unveiling RBP roles in RNA-linked diseases within endogenous biological contexts such as developmental studies, animal disease models, clinical samples, and other biological contexts.

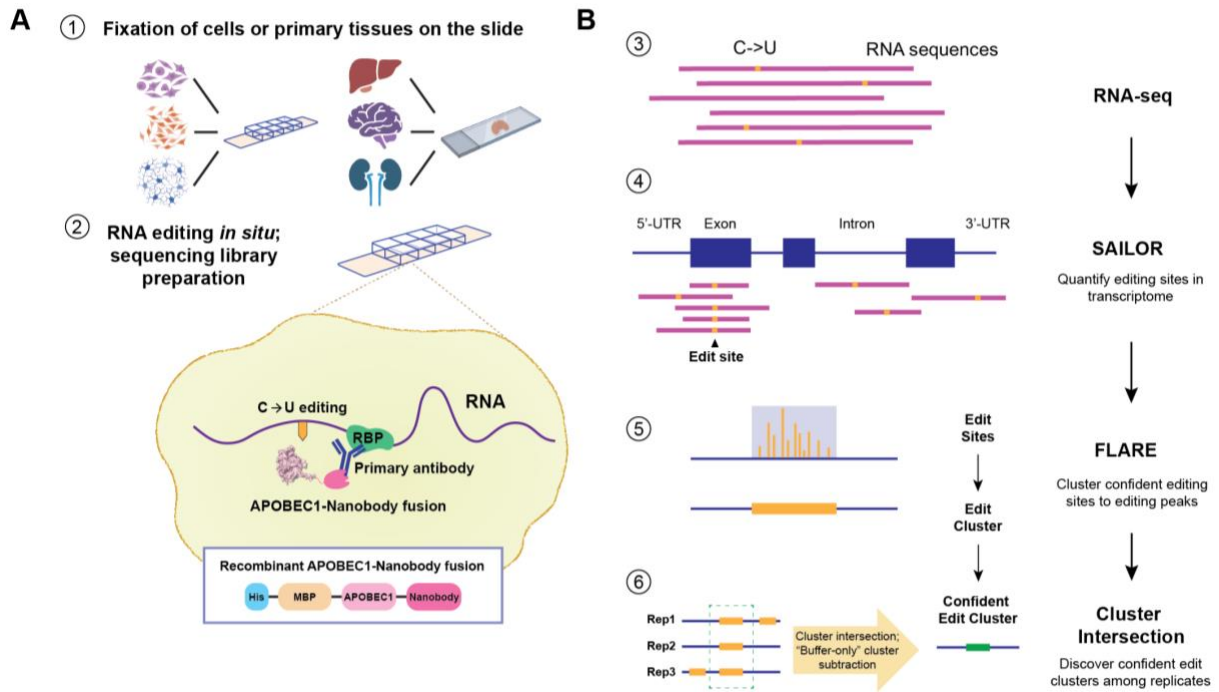


Figure 2.1 Schematics of *in situ* RNA C-to-U editing mediated by recombinant APOBEC1-Nanobody.

(A) Experimental procedure of INSCRIBE. After cell or tissue fixation, the RBP-targeting antibody (Ab) and the APO1-Nb are incubated sequentially, followed by RNA editing *in situ* at 37°C overnight. RNAs are harvested for RNA-seq library preparation (Partially created with BioRender.com).

(B) Workflow describing data analysis for INSCRIBE. The RNA sequences are aligned to the transcriptome, followed by C-to-U edit quantification by SAILOR and edit clusters identification by FLARE. Subsequently, confident edit clusters are determined by intersecting clusters from three replicates and subtracting the Buffer-only control clusters. UTR: untranslated region.

Chapter 2.2 Results

An APOBEC1-nanobody fusion protein mediates targeted RNA editing *in situ*

In the STAMP method, a chimeric fusion protein of APOBEC1 and an RBP-of-interest is overexpressed in cells, resulting in location-specific cytosine-to-uracil (C-to-U) RNA editing, enabling identification of authentic RBP binding sites by high-throughput sequencing from cellular RNAs¹². We sought a method to circumvent the need to engineer individual RBP-APOBEC1 fusions in cells of interest. Since APOBEC1 is also active *in vitro*¹⁴, we designed a fusion of APOBEC1 to a llama-derived single-domain antibody (nanobody) against rabbit IgG (Apo1-Nb; **Figure 2.1A**)¹⁵ and expressed the fusion protein in *E. coli*. A multi-step purification protocol yielded a >95% pure and homogeneous species with a molecular weight consistent with dimeric Apo1-Nb (**Supplementary Figure 2.1A-B**). We verified the deaminase activity of Apo1-Nb by incubating the protein with an *in vitro*-transcribed 300 nt RNA derived from the 3'-UTR of the human *APP* gene (**Supplemental Table 2.1**), resulting in ~25% C-to-U editing at multiple sites as assayed by Sanger sequencing (**Figure 2.2A**). Notably, we observed no editing in high-salt buffers (75 mM or 300 mM NaCl) or at low temperature (25°C), permitting precise control of the editing reaction and minimization of non-specific RNA editing (**Supplementary Figure 2.1C**). Immunostaining showed that Apo1-Nb was recruited to the nucleus of methanol-fixed HEK293T cells in the presence of a rabbit-derived antibody against RBFOX2 (**Supplementary Figure 2.1D-F**). Through immunostaining, we systematically varied the concentrations of both the anti-RBFOX2 primary antibody and Apo1-Nb to optimize the targeting specificity, as measured by the nuclear fraction of RBFOX2 localization and the co-localization of the anti-RBFOX2 antibody and Apo1-Nb. We chose a 1:200 dilution of anti-

RBFOX2 Ab and a 1:1000 dilution of APO1-Nb to maximize signal intensity (Global Intensity) while maintaining specificity (Nuclear Fraction). Thus, Apo1-Nb is an active RNA base editor that can be specifically recruited to sites of interest in fixed cells by primary antibodies.

We next carried out *in situ* RNA editing in methanol(MeOH)-fixed HEK293T cells, since methanol fixation largely maintains RNA integrity¹⁶. HEK293T cells were fixed on a slide and incubated with a rabbit-derived anti-RBFOX2 antibody. After washing, APO1-Nb was immobilized to the antibody on ice in phosphate buffered saline + Tween-20 (PBS-T) buffer, where the low temperature and high salt concentration (~175 mM) prevented non-specific RNA editing. Slides were then incubated overnight at 37°C in a buffer containing 20 mM NaCl to allow RNA editing, followed by RNA extraction and library preparation for RNA-seq. Multiple control samples were incorporated, including “Enzyme only” which lacked the anti-RBFOX2 primary antibody, and “Buffer only” which lacked both the anti- RBFOX2 antibody and Apo1-Nb.

We adapted the STAMP computational pipeline to identify clusters of C-to-U editing sites representing RBP-RNA interaction sites in the transcriptome from INSCRIBE samples. In brief, C-to-U editing events were identified by transcriptomic mapping and quantification through a modified SAILOR pipeline^{12,17}, followed by edit cluster identification through FLARE (FLagging Areas of RNA-editing Enrichment)¹⁸ (**Figure 2.1B**). FLARE builds on the edit site outputs of the SAILOR algorithm to identify regions that are statistically enriched for RNA editing. Based on the enzyme efficiency that we observed in the *in vitro* RNA editing assay, we filtered out editing sites with a SAILOR score <0.5 (P-value >0.5) or an editing fraction >80% (which likely represent cell line specific SNPs) before proceeding to FLARE, which retained 75%-80% SAILOR C-to-U edit sites. With ~60 million reads per replicate, the pipeline typically

identified around 200,000 edit clusters across the transcriptome in each replicate for RBFOX2-INSCRIBE, while roughly 120,000 edit clusters were determined in controls (**Supplemental Table 2.2**). Sites called in buffer-only controls represent C-to-U conversions introduced by RNA damage, reverse transcription errors, or PCR errors during the experimental protocol and library preparation. Nonetheless, all three replicates of RBFOX2-INSCRIBE showed clustered C-to-U editing sites at the 3'-untranslated region (3'-UTR) of the known RBFOX2 target mRNA *APP*, whereas enzyme-only controls showed few or no editing sites in this region (**Figure 2.2B**). The canonical RBFOX2 binding motif UGCAUG was enriched in *de novo* motif discovery¹⁹ in all replicates of RBFOX2-INSCRIBE (**Supplementary Figure 2.1G**), demonstrating the specificity of this method for discovering authentic RBP binding sites. Edit fractions across replicates in transcriptomic regions of RBFOX2-INSCRIBE edit clusters exhibited a strong correlation (Pearson R=0.83 and 0.81, **Supplementary Figure 2.2A**), indicating that INSCRIBE has strong reproducibility.

We further investigated the effect of APO1-Nb concentration on the specificity of INSCRIBE by measuring enrichment of the canonical RBFOX2 motif (UGCAUG) in edit clusters (**Supplementary Figure 2.2B**). The best enrichment of proximal UGCAUG in edit clusters was observed in 1:1000 APO1-Nb dilution ratio (corresponding to a final concentration of 1 µg/mL). Higher or lower concentrations of APO1-Nb led to reduced RBP probing specificity, due to either insufficient Apo1-Nb and sub-optimal editing efficiency, or oversaturated APO1-Nb and increased non-specific editing. This result agrees with our immunofluorescent staining observations (**Supplementary Figure 2.1D-F**), and we therefore concluded that 1µg/mL APO1-Nb is optimal for RBFOX2-INSCRIBE.

Figure 2.2 INSCRIBE identifies authentic RBFOX2 binding sites through confident C-to-U edit clusters.

(A) Sanger sequencing result of the in vitro RNA editing assay validated the C-to-U editing enzymatic activity of the recombinant APOBEC1-nanobody (APO1-Nb). UTR: untranslated region.

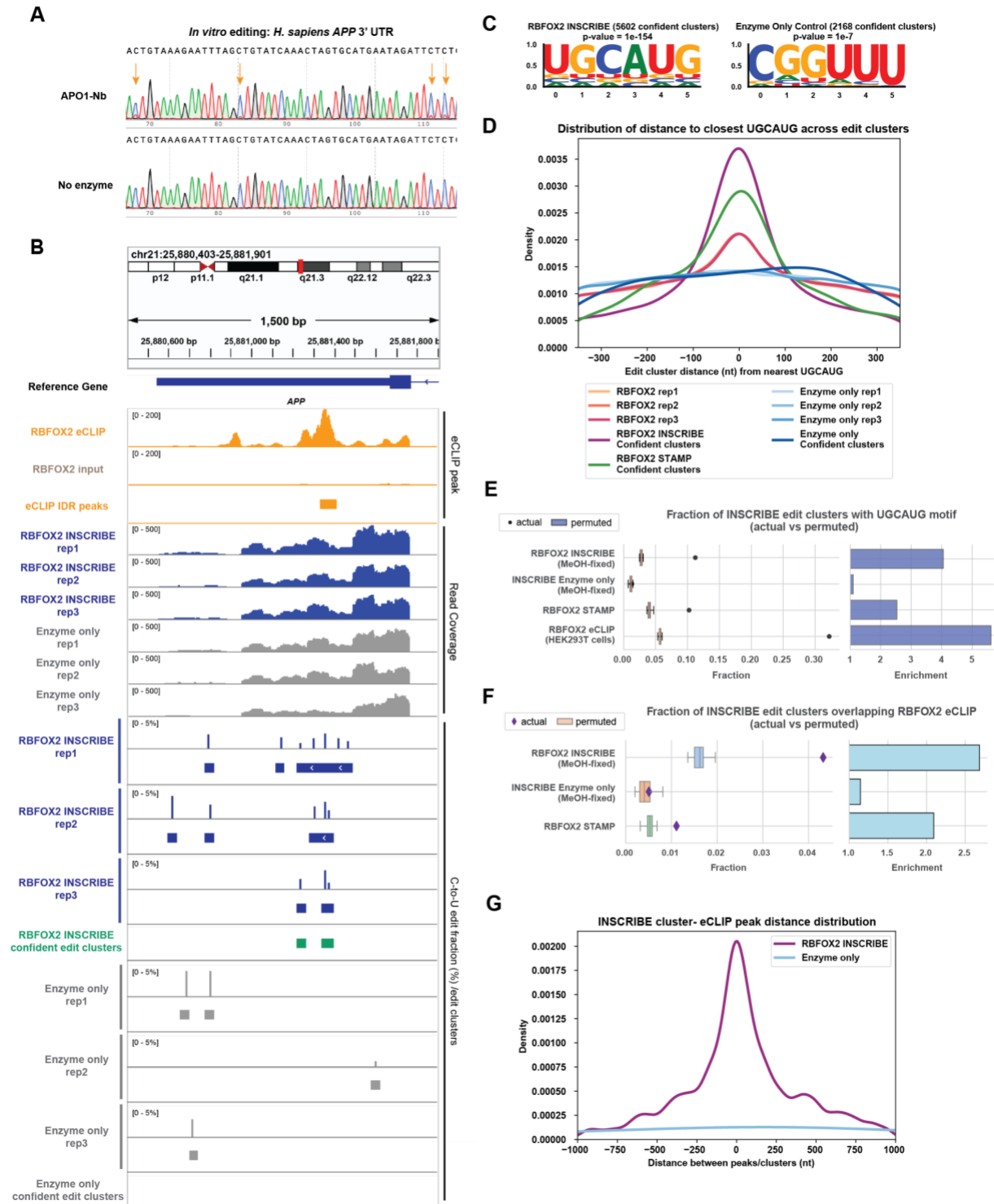
(B) Integrative genome viewer (IGV) tracks showing RBFOX2 eCLIP peaks (HEK293T) on the target gene APP expanding a 1500bp window, compared with the read coverage, SAILOR-quantified edit fraction and FLARE-determined edit clusters of three replicates of RBFOX2-INSCRIBE, along with the enzyme-only negative controls. The confident edit clusters derived from three replicates were also shown.

(C) HOMER de novo motif discovery identified the canonical RBFOX2 motif UGCAUG as the top motif in all RBFOX2-INSCRIBE replicates, using a cumulative hypergeometric distribution for p-values.

(D) Distribution of the distance between the closest UGCAUG and confident edit clusters for RBFOX2-INSCRIBE replicates (orange) and confident edit clusters (violet), enzyme-only control replicates (light blue) and confident edit clusters (dark blue), and RBFOX2-STAMP (green). A peak at 0nt distance indicated the enrichment of UGCAUG in proximity to the edit cluster center, demonstrating the RBFOX2-binding driven editing. The RBFOX2-INSCRIBE confident edit clusters presented much stronger enrichment of adjacent UGCAUG motif than the individual replicates.

(E, F) E, The actual fraction of UGCAUG-containing RBFOX2-INSCRIBE edit clusters/RBFOX2 eCLIP IDR peaks (dots) compared to the 20 permuted clusters (box plot, $n = 20$). Z-scores: RBFOX2-INSCRIBE, 49.93; Enzyme-only control, 0.43; RBFOX2-STAMP, 23.08; RBFOX2-eCLIP, 117.83. F, the actual fraction of eCLIP-overlapping RBFOX2-INSCRIBE edit clusters (diamonds) compared to the permuted clusters (box plot, $n = 20$). Z-scores: RBFOX2-INSCRIBE, 18.84; Enzyme-only control, 0.39; RBFOX2-STAMP, 5.45. In E-F, the box shows the quartiles while the whiskers extend to show the rest of the distribution; the median is represented by the center line. Enrichment is defined as the ratio of actual fraction to the mean of the permuted-derived fraction. The actual fraction was plotted using the confident edit clusters derived from 3 INSCRIBE technical replicates.

(G) Distribution of the distance between the closest eCLIP peak and confident edit clusters for RBFOX2-INSCRIBE (violet) and enzyme-only control (light blue).



Since the number of edit clusters identified in both buffer-only and enzyme-only negative controls were nontrivial, we applied further background noise elimination strategies (**Figure 2.1B**). First, only edit clusters that appeared in all three experimental replicates were retained. Second, to reduce experimental noise introduced by endogenous C-to-U editing in cells, edit clusters identified in buffer-only controls were removed. These two steps removed 97-98% of edit clusters in each individual replicate, and the remaining edit clusters were termed “confident edit clusters” and used for further analysis (**Supplemental Table 2.3**). FLARE identified 5,602 confident edit clusters in RBFOX2-INSCRIBE in methanol-fixed HEK293T cells, while 2,168 confident edit clusters were determined in enzyme-only controls. The canonical RBFOX2 motif was enriched in RBFOX2-INSCRIBE confident edit clusters ($p = 1e-154$), while it was not enriched in both enzyme-only and buffer-only controls (**Figure 2.2C**). Confident edit clusters after these noise-reduction strategies showed higher enrichment of UGCAUG compared to edit clusters identified in individual replicates, demonstrating enhanced specificity of RBFOX2 target RNA identification (**Figure 2.2D**).

Further examination of the confident edit clusters showcased the accuracy of RBP binding sites discovered by INSCRIBE. We found that around 10% of confident RBFOX2-INSCRIBE edit clusters contained the UGCAUG motif, 4-fold enriched over the randomly shuffled (permuted) size-matched clusters within the same regions (**Figure 2.2E**), which was comparable to RBFOX2-STAMP¹⁸. In contrast, confident edit clusters in enzyme-only controls showed no enrichment of the UGCAUG motif over permuted clusters, indicating the randomness of editing in the absence of the anti-RBFOX2 primary antibody. Empirically defined RBFOX2 sites were defined as reproducible peaks from irreproducible discovery rate (IDR) analysis of HEK293T RBFOX2 eCLIP data and were compared to RBFOX2-INSCRIBE. We observed that

the fraction of RBFOX2-INSCRIBE confident edit clusters overlapping with RBFOX2-eCLIP IDR peaks was enriched (>2.5 fold) compared to permuted clusters (**Figure 2.2F**) and that INSCRIBE clusters were enriched near eCLIP IDR peaks (**Figure 2.2G**), indicating that INSCRIBE identifies true RBFOX2 binding sites. Examination of individual edit clusters of INSCRIBE revealed their proximity and concurrence with eCLIP peaks (**Supplementary Figure 2.2C**). In summary, we established an optimal working protocol for INSCRIBE by APO1-Nb RNA editing in MeOH-fixed HEK293T cells, which enabled the successful identification of RBFOX2 binding sites *in situ*.

Comparison of INSCRIBE and eCLIP reveals differences in resolution

Next, we focused on how RBFOX2-INSCRIBE confident clusters compares to the binding sites identified by conventional gold-standard method eCLIP. Despite the low overlap (**Figure 2.2F**) of INSCRIBE edit clusters with eCLIP binding sites, we observed, expectedly, that RBFOX2-INSCRIBE clusters that overlap with RBFOX2-eCLIP peaks had enriched UGCAUG motifs (**Supplementary Figure 2.3A**). Interestingly, INSCRIBE clusters that did not have an overlapping eCLIP peak also exhibited statistically significant ($p\text{-value} < 10^{-111}$) UGCAUG motifs (**Supplementary Figure 2.3B**). When we studied the types of genic regions found to contain overlapping or non-overlapping clusters, we observed distinct preferences. (**Supplementary Figure 2.3C**). Compared to eCLIP, INSCRIBE discovered a lower fraction of edit clusters within introns and a higher fraction within 3'UTRs, likely due to the fundamental difference in selection and read coverage for the different regions within pre-mRNAs. Specifically, introns were not as deeply covered by reads and without a positive antibody selection as in eCLIP, intronic regions were less represented in the INSCRIBE data. Nevertheless, despite the lower-than-anticipated overlap between INSCRIBE edits and eCLIP

binding sites due to differences in regional representation, our results indicate that authentic binding sites are being recovered by INSCRIBE.

Another difference between eCLIP and INSCRIBE is the resolution of identified binding sites, as the APOBEC1 enzyme recognizes the RBP through an antibody-nanobody linkage. We expected that INSCRIBE edit clusters might reside at a distance from the actual nucleotides that interacted with the RBP. As RBFOX2 interacts with the UGCAUG motif, we plotted the cumulative distribution of INSCRIBE clusters or enzyme-only clusters with increased distance to UGCAUG (**Supplementary Figure 2.3D**). The slope of the enzyme-only control curve reflected the variation of APOBEC1 editing and the natural occurrence of UGCAUG in the transcriptome. We observed RBFOX2-INSCRIBE converges with enzyme-only at ~200nt from UGCAUG (**Supplementary Figure 2.3E**), which indicates that the INSCRIBE edit radius is at most 200nt from the true binding site. This likely also contributed to a lower overlap with eCLIP binding sites than expected if the edit cluster was computationally defined outside of the ~50-100 nt eCLIP-defined windows. Therefore, we observed a slightly lower enrichment of UGCAUG motifs within RBFOX2-INSCRIBE clusters compared to eCLIP-defined sites. Nevertheless, in summary, expanding the RBFOX2 motif to GCAYG (Y=C or U) and including secondary motifs GCUUG, GAAUG, GUUUG, GUAUG, GUGUG and GCCUG as previously described²⁰, around 54% of RBFOX2-INSCRIBE confident edit clusters have either eCLIP or motif-based evidence for RBFOX2 interactions.

INSCRIBE enables RNA isoform distinctions in RBP binding sites discovery

To examine whether INSCRIBE is compatible with long-read sequencing, which enables the detection of RBP binding at mRNA isoform resolution, we generated and sequenced libraries from RBFOX2-INSCRIBE and the enzyme-only control using the PacBio sequencing platform.

We chose the PacBio system as it currently has a lower base-calling error rate than the Oxford Nanopore Technologies platform^{12,21}. After removing SNPs (Methods), *de novo* motif discovery within edited regions confirmed that both replicates of long-read RBFOX2-INSCRIBE exhibited statistically significant enriched (p-value = 1e-98 and 1e-102 for replicate1 and replicate2) UGCAUG motifs (**Figure 2.3A**). Additionally, the edits detected at the mRNA isoform level showed a high degree of correlation between replicates as measured using editsC (ratio of number of edited Cs relative to the number of Cs across exons and UTRs of isoforms), with a Pearson correlation R=0.95 (**Supplementary Figure 2.4A**), consistent with the high reproducibility of INSCRIBE experiments.

RBFOX2-INSCRIBE displayed general higher C-to-U editing rates than enzyme-only control (**Figure 2.3B**). We conducted principal component analysis (PCA) using isoform-level editsC and reads per kilobase of transcript per million reads mapped (RPKM) independently to ensure that signal identified in RBFOX2-INSCRIBE was not due to differences in isoform expression. PCA showed a clear separation between RBFOX2-INSCRIBE and enzyme-only control with editsC, but not RPKM, confirming that signal differences were due to editing. (**Supplementary Figure 2.4B**). To evaluate isoform-specific RBP binding, we first filtered out RNA isoforms with read coverage <20 across replicates and edit fraction <0.02 to improve RBFOX2-INSCRIBE signal detection over background, as represented by the enzyme-only control (**Supplementary Figure 2.4C**). We constructed a linear model using the average isoform-level editsC calculated from RBFOX2-INSCRIBE and enzyme-only samples and selected high confidence isoforms in RBFOX2-INSCRIBE samples (>1.5 std.) for subsequent analysis (Orange in **Supplementary Figure 2.4D**, n = 204).

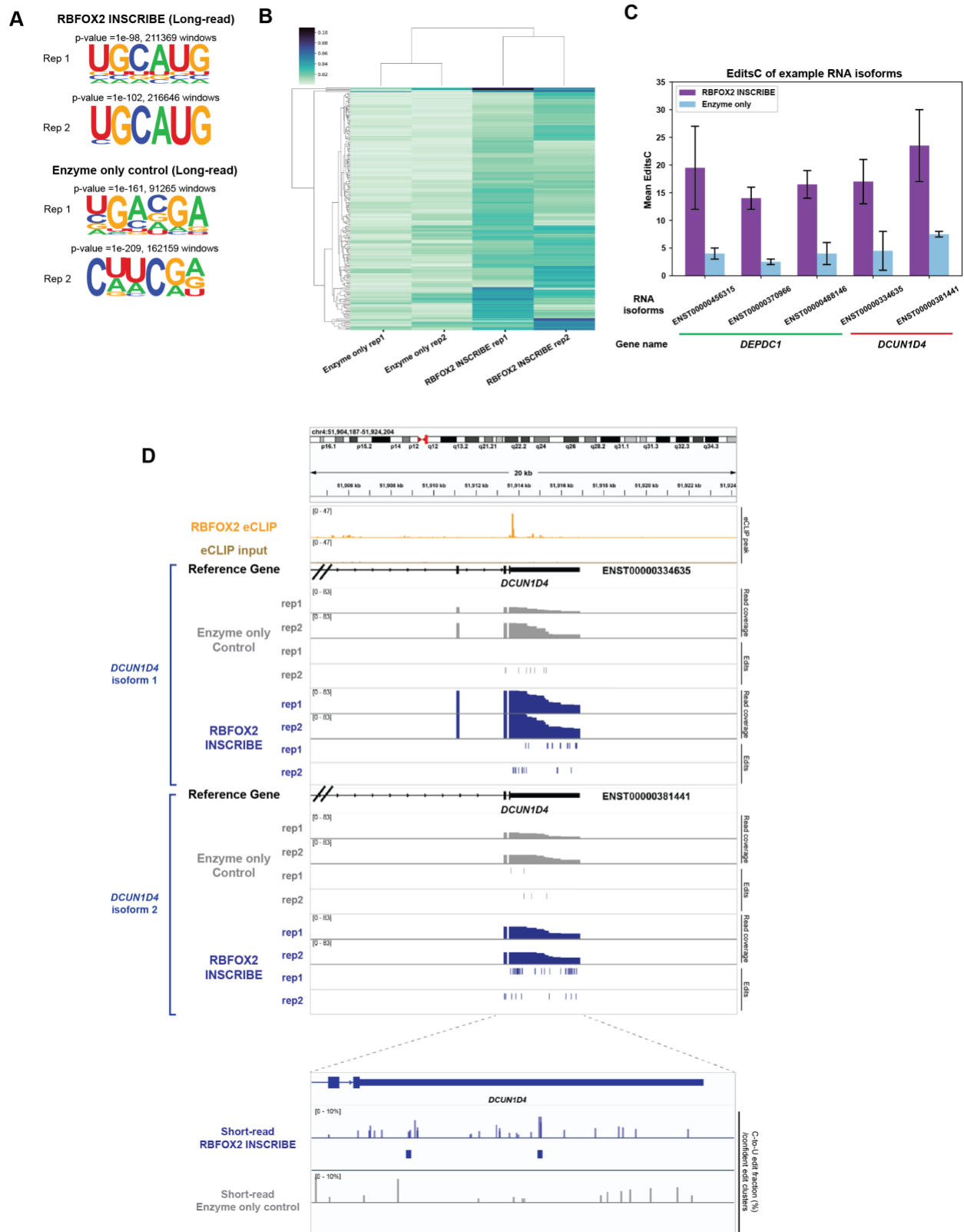
Figure 2.3 Long-read sequenced INSCRIBE enables mRNA isoform distinctions

(A) HOMER de novo motif discovery of PacBio sequenced INSCRIBE replicates identified the canonical RBFOX2 motif (UGCAUG), using a cumulative hypergeometric distribution for p-values.

(B) Heatmap of editsC (ratio of number of edited Cs relative to the number of Cs across exons and UTRs of isoforms) of individual RNA isoforms from both replicates of RBFOX2-INSCRIBE and enzyme-only controls.

(C) EditsC of example RNA isoforms in gene DEPDC1 and DCUN1D4 from both replicates of RBFOX2-INSCRIBE and enzyme-only control.

(D) Integrative genome viewer (IGV) tracks showing PacBio sequenced RBFOX2-INSCRIBE vs enzyme-only control edits on two isoforms of DCUN1D4 in a 20kb window, alongside with INSCCRIBE read coverage, eCLIP peaks (HEK293T). The short-read sequenced RBFOX2-INSCRIBE vs enzyme-only control edit fractions and the confident edit clusters are shown at the bottom panel for comparison.



Examples of differential isoform editing signatures are illustrated by *DCUN1D4* and *DEPDC1* genes (**Figure 2.3C-D** and **Supplementary Figure 2.4E**). These differential isoform edits agree with both RBFOX2 eCLIP and short-read RBFOX2-INSCRIBE edit clusters. We conclude that INSCRIBE enables isoform-sensitive identification of RBP-RNA interactions when coupled with long-read sequencing.

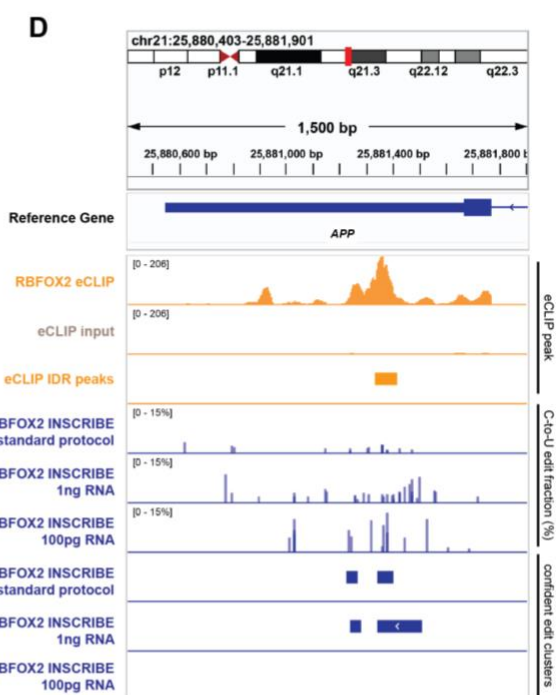
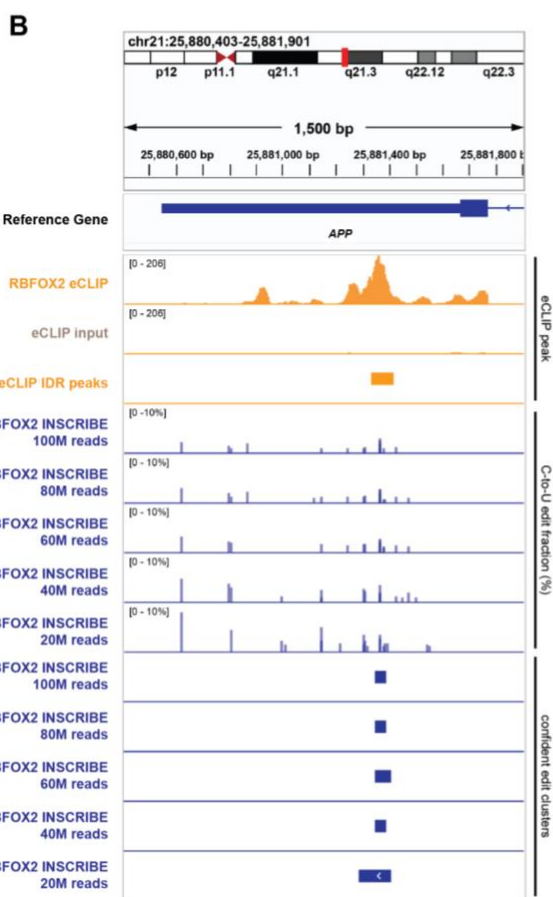
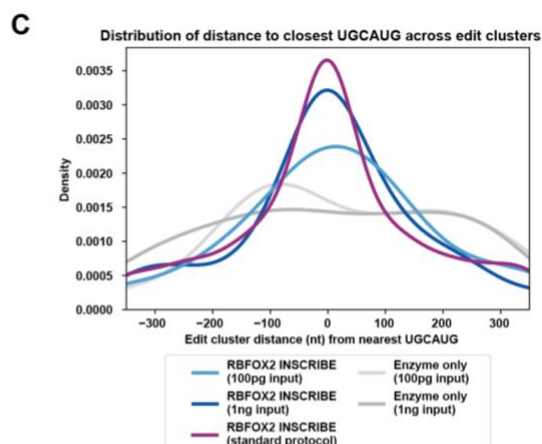
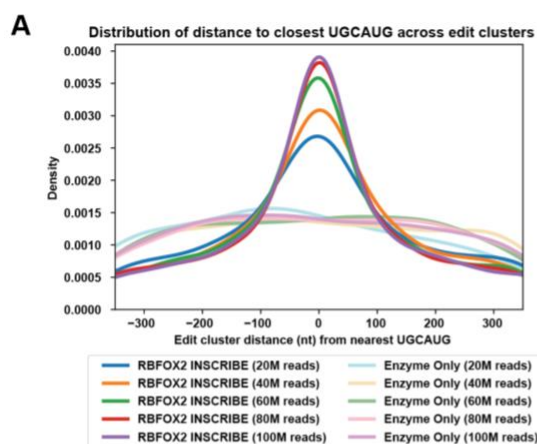
INSCRIBE identifies RBP binding sites with low sequencing depth and ultra-low input RNA

As the accuracy of C-to-U edit cluster calls depends on the sequencing read depth of the RNA in question, it is potentially challenging to identify RBP-RNA interactions on lowly expressed transcripts or in intronic regions with low read coverage. To evaluate the sequencing depth required by INSCRIBE for reliable edit cluster identification, we sequenced short-read libraries from RBFOX2-INSCRIBE and controls from methanol-fixed HEK293T cells to roughly 100 million (100M) reads per replicate, then randomly down-sampled each dataset to 80, 60, 40, and 20 million reads. The number of C-to-U edit clusters determined by FLARE was proportionally associated with sequencing depth in each sample (**Supplementary Figure 2.5A, Supplemental Table 2.4**). A greater number of confident edit clusters were identified with higher sequencing depth, but the canonical RBFOX2 motif was identified in all samples regardless of sequencing depth through *de novo* motif analysis (**Supplementary Figure 2.5B**). Enrichment of the UGCAUG motif was generally higher in deeper-sequenced data, although the difference between 100M reads and 80M reads was negligible (**Figure 2.4A and Supplementary Figure 2.5G**). As expected, increased sequencing depth resulted in an increased count and proportion of identified intronic edit clusters (**Supplementary Figure 2.5D**). With 100M reads, 61.5% (5,938 counts) of the confident edit clusters fell into intronic regions,

compared to 21.5% (285 counts) with 20M reads (**Supplemental Table 2.4**). At the level of individual sites, deeper sequencing depth effectively reduced some inflated C-to-U edit fractions and thereby reduced the noise arising from low read coverage (**Figure 2.4B**). Nonetheless, both the overall SAILOR-quantified edit site patterns and the FLARE confident edit clusters remained similar across all samples from 20M to 100M reads (**Figure 2.4B**), as opposed to the enzyme-only controls (**Supplementary Figure 2.6A**). Irrespective of sequencing depth, INSCRIBE showed a similar level of concordance with eCLIP IDR peaks (**Supplementary Figure 2.5H**). In conclusion, we observed that INSCRIBE demonstrates effective identification of RBP binding sites with sequencing depth as low as 20M reads. For targets such as intronic sites or low-level transcripts, deeper sequencing depths is helpful, but it is unnecessary to exceed 100M reads per sample.

Figure 2.4 The performance of INSCRIBE at various sequencing depths and ultra-low RNA input

- (A) The enrichment of UGCAUG motif at confident edit clusters center with different sequencing depths of RBFOX2-INSCRIBE and enzyme-only control. Increasing sequencing depth from 20 to 40, 60, 80, 100 million (M) reads enhanced the enrichment of UGCAUG, while the improvement from 80M to 100M was marginal.
- (B) Integrative genome viewer (IGV) tracks of representative RBFOX2 target gene (APP) examining the effect of sequencing depth on SAILOR-quantified C-to-U conversion pattern and FLARE confident edit clusters. The tracks include eCLIP peaks (orange), 3-replicate overlay of SAILOR-quantified edit fraction (blue, upper tracks) and FLARE confident edit clusters (blue, lower tracks) of RBFOX2-INSCRIBE.
- (C) The enrichment of UGCAUG motif at confident edit clusters center with libraries of different input total RNA of RBFOX2-INSCRIBE and enzyme-only controls. Both 1 ng (dark blue) and 100 pg (light blue) input RNA showed evident enrichment of UGCAUG motif as compared to standard input (>1000ng RNA) (purple) in RBFOX2-INSCRIBE, while lower amount of input material decreased the enrichment.
- (D) IGV tracks of representative RBFOX2 target gene (APP) examining the effect of input starting total RNA on SAILOR-quantified C-to-U conversion pattern and FLARE confident edit clusters. The tracks include eCLIP peaks (orange), 3-replicate overlay of SAILOR-quantified edit fraction (blue, upper tracks) and FLARE confident edit clusters (blue, lower tracks) of RBFOX2-INSCRIBE.
- (E) Continued IGV tracks showing the FLARE-identified edit clusters in individual replicates for the 100 pg input RNA libraries.



Next, we sought to determine the detection limit of INSCRIBE, as the input material requirement has been the major hurdle of profiling RBP-RNA interactions in precious tissues such as clinical samples or stem cell-derived models. In the standard RBFOX2-INSCRIBE experiment in HEK293T cells, more than 1000 ng of total RNA is typically extracted from ~150,000 cells. Aware of the cost per experiment for valuable samples, we scaled down the amount of total RNA as input over 1000-fold, to either 1 ng or 100 pg, and prepared RNA-seq libraries. Notably, with ~50M sequencing depth, both 1 ng and 100 pg input showed promising enrichment of the RBFOX2 canonical motif UGCAUG (**Figure 2.4C and Supplementary Figure 2.5E**). The ultra-low input RBFOX2-INSCRIBE also showed robust performance in recovering the UGCAUG motif from only 776 edit clusters from 1 ng total RNA and 214 clusters from 100 pg total RNA (**Supplementary Figure 2.5C**). Like the INSCRIBE library with lower sequencing depth, ultra-low input INSCRIBE unavoidably presented inflated C-to-U edit fractions when mapped onto the transcriptome (**Figure 2.4D**). Nonetheless, the SAILOR-quantified edits still cluster around eCLIP peaks in contrast to the enzyme-only controls (**Supplementary Figure 2.6B**). For 100 pg input RNA, FLARE successfully discovered this binding site in all three individual replicates (**Figure 2.4E**). We recognized that some RNA targets might be dropped out with 100 pg of input RNA, such as the known binding site within the 3'UTR in the *APP* gene, which is a common caveat of low-input or single cell sequencing technologies, but INSCRIBE can still identify accurate RNA targets of RBFOX2. Both ultra-low input libraries showed significant enrichment of overlap with eCLIP peaks in INSCRIBE-discovered edit clusters (**Supplementary Figure 2.5F**). Therefore, we demonstrate high sensitivity of INSCRIBE and showcase the transcriptomic discovery of endogenous RBP-RNA

interactions from as little as 100 pg of input RNA, which is equivalent to five cells as the starting material.

INSCRIBE identifies authentic TDP-43 binding sites through confident C-to-U edit clusters

To demonstrate the versatility of the APO1-Nb design and INSCRIBE's performance in profiling other RBPs, we optimized the dilution factor for a rabbit-derived anti-TDP-43 antibody and successfully profiled TDP-43, a nuclear RNA/DNA-binding protein implicated in many neurodegenerative diseases²². FLARE identified 5,751 confident edit clusters and *de novo* motif discovery on confident edit clusters of TDP-43-INSCRIBE showed enrichment of the canonical TDP-43 binding motif UGUGUG ($p = 1e-20$) (**Figure 2.5A**). The fraction of confident TDP-43-INSCRIBE edit clusters exhibiting a UGUGUG motif is notably higher than in shuffled size-matched peaks (**Figure 2.5B**). The motif was also enriched in proximity of TDP-43-INSCRIBE edit clusters relative to enzyme-only controls (**Figure 2.5C**). Moreover, we noted a more than 2.5-fold enrichment in the proportion of TDP-43-INSCRIBE peaks overlapping with TDP-43 eCLIP, compared to enzyme-only controls (**Figure 2.5D**). Known TDP-43 binding sites on the *TARDBP*, *NUCKS1*, *XPO1* and *FUS* genes were clearly captured by INSCRIBE C-to-U edit clusters and showed agreement with eCLIP peaks (**Figure 2.5E**). In conclusion, INSCRIBE demonstrates its versatility in robust RNA target identification of RBPs with a universally adaptable protocol.

Figure 2.5 INSCRIBE identifies authentic TDP-43 binding sites through confident C-to-U edit clusters

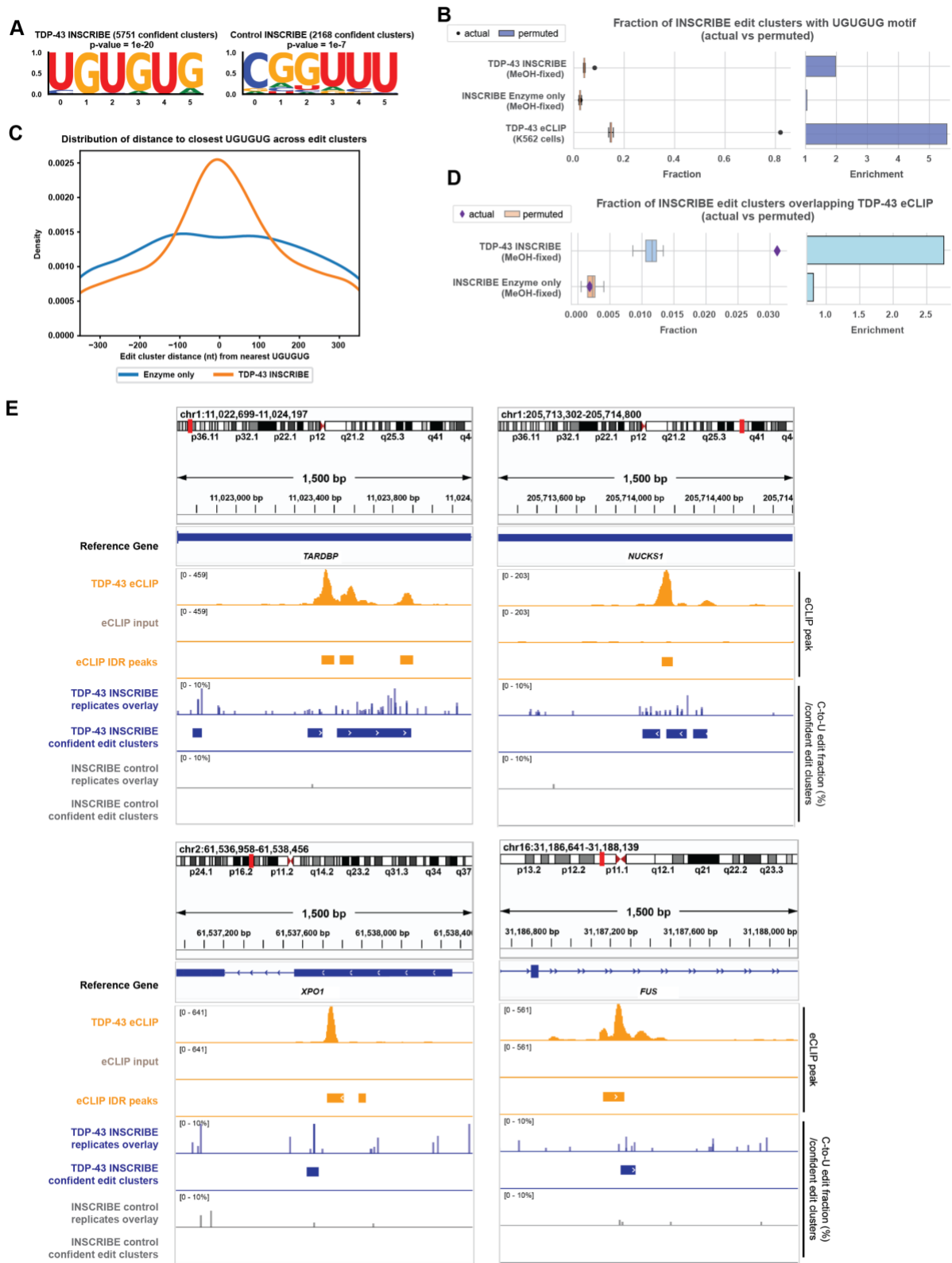
(A) HOMER de novo motif discovery of TDP-43 INSCRIBE replicates identified the canonical TDP-43 motif (UGUGUG), using a cumulative hypergeometric distribution for p-values.

(B) The actual fraction of UGUGUG-containing TDP-43-INSCRIBE edit clusters/TDP-43 eCLIP IDR peaks (dots) compared with that of the 20 permuted clusters (box plot, $n = 20$). The actual fraction was plotted using the confident edit clusters derived from 3 INSCRIBE technical replicates. The box shows the quartiles while the whiskers extend to show the rest of the distribution; the median is represented by the center line. Z-scores: TDP-43-INSCRIBE, 22.22; Enzyme-only control, 0.33; K562 TDP-43 eCLIP, 146.11.

(C) The distribution of the nearest UGUGUG distance to clusters of confident edits reveals a peak at 0nt for TDP-43-INSCRIBE (orange), indicating an abundance of UGUGUG sequences in close proximity to the center of the edit cluster, which demonstrates TDP43-binding driven C-to-U editing.

(D) The actual fraction of eCLIP-overlapping TDP-43-INSCRIBE edit clusters (diamonds) compared to that of the 20 permuted clusters (box plot, $n = 20$). The actual fraction was plotted using the confident edit clusters derived from 3 INSCRIBE technical replicates. The box shows the quartiles while the whiskers extend to show the rest of the distribution; the median is represented by the center line. Z-scores: TDP-43-INSCRIBE, 15.61; Enzyme-only control, -0.46.

(E) Integrative genome viewer (IGV) tracks of example RBFOX2 target genes (TARDBP, NUCKS1, XPO1, FUS) with 1500bp windows showing RBFOX2 eCLIP peaks (orange), RBFOX2-INSCRIBE (blue) and the enzyme-only control (gray). The 3-replicate overlay of SAILOR-quantified edit fraction of INSCRIBE is shown along with the FLARE-determined confident edit clusters.



INSCRIBE is compatible with multiple cell fixation methods

As formalin or formaldehyde fixation is more prevalent than methanol fixation in clinical sample collection workflows²³, we next aimed to implement INSCRIBE after formaldehyde fixation, which is known to damage RNAs and induce RNA modifications including nonspecific cytosine deamination²⁴. We applied a mild de-crosslinking protocol that effectively preserved RNA integrity from paraformaldehyde (PFA) fixation (**Supplementary Figure 2.7A**). We observed comparable numbers of confident edit clusters in PFA-fixed HEK293T cells as in methanol-fixed cells for both RBFOX2 (2,970 clusters) and TDP-43 (2,540 clusters). *De novo* motif discovery identified both the canonical RBFOX2 motif ($p=1e-78$) and TDP-43 motif ($p=1e-5$; 2nd top motif). These motifs were enriched in regions adjacent to confident edit cluster centers comparable to methanol-fixed INSCRIBE (**Figure 2.6A, Figure 2.6B**). We also observed a similar enrichment of the canonical motif and of the eCLIP IDR peak overlap in confident edit clusters in PFA-fixed samples as in MeOH-fixed samples (**Figure 2.6C, Figure 2.6D**). In PFA-fixed INSCRIBE, the C-to-U editing and the confident edit clusters at representative target genes also uncovered their striking proximity and alignment with eCLIP peaks (**Figure 2.6E, Supplementary Figure 2.7B**). In summary, INSCRIBE is compatible with PFA-fixed cells, enhancing its utility in analysis of clinical samples and extending its adaptability to a broader range of experimental workflows.

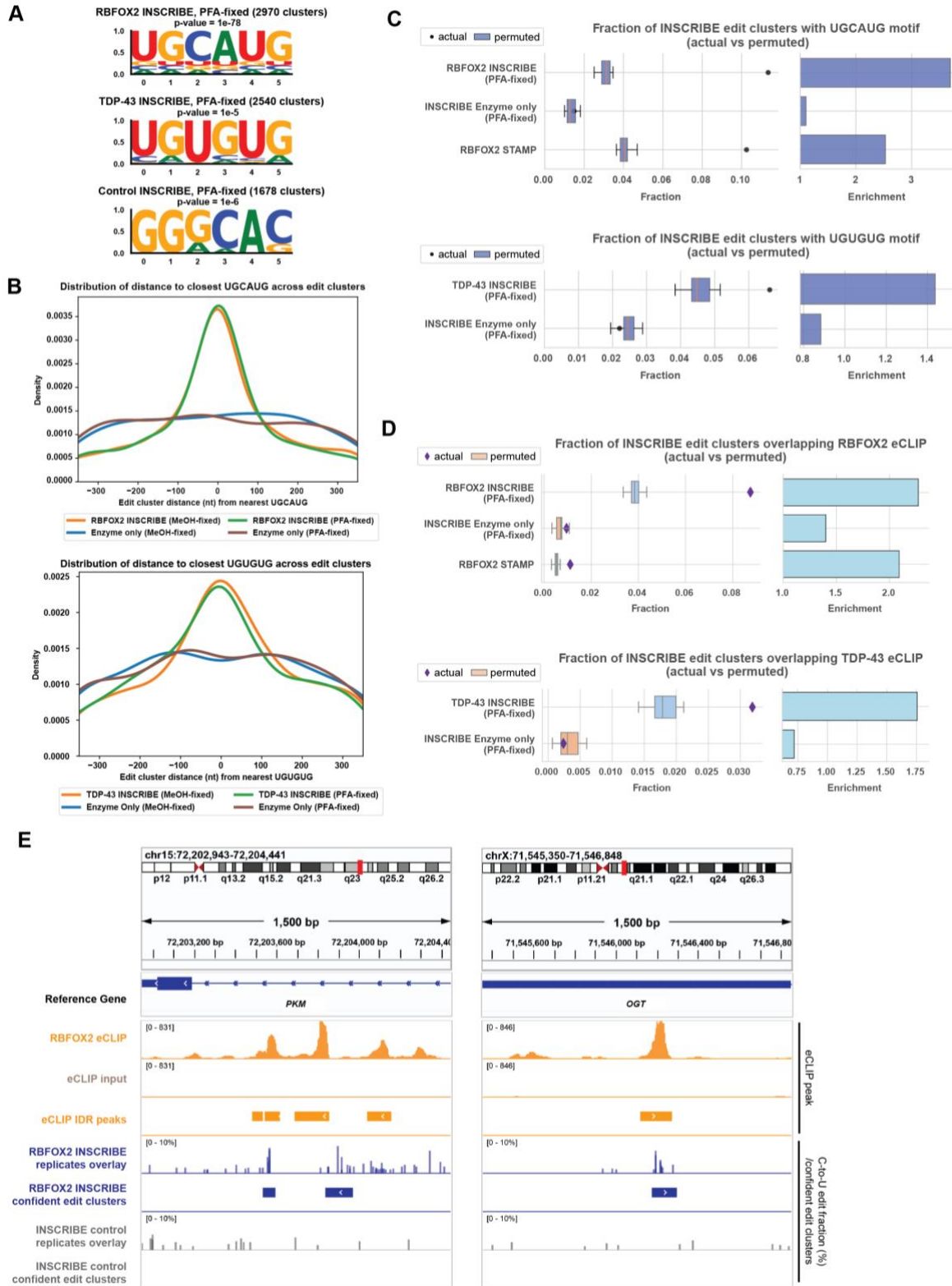
Figure 2.6 INSCRIBE exhibited comparable performance in PFA-fixed cells

(A) HOMER de novo motif analysis of confident edit clusters revealed the canonical RBFOX2 motif UGCAUG (top motif) or TDP-43 motif UGUGUG (2nd top motif), for RBFOX2-INSCRIBE or TDP-43-INSCRIBE in paraformaldehyde (PFA)-fixed HEK293T cells respectively, using a cumulative hypergeometric distribution for p-values.

(B) A density plot of the distance between edit clusters to closest UGCAUG (for RBFOX2-INSCRIBE, upper panel) or UGUGUG (for TDP-43-INSCRIBE, lower panel). INSCRIBE in MeOH-fixed and PFA-fixed cells presented no obvious distinction in canonical motif enrichment at the INSCRIBE confident edit cluster center.

(C-D) C, Actual fraction of motif-containing INSCRIBE edit clusters (dots) compared with that of the 20 permuted clusters (box plot, $n = 20$). Top: RBFOX2-INSCRIBE in PFA-fixed HEK293T cells. Z-scores: RBFOX2-INSCRIBE (PFA-fixed), 29.27; Enzyme-only control (PFA-fixed), 0.54; RBFOX2-STAMP: 23.08; Bottom: TDP-43-INSCRIBE in PFA-fixed HEK293T cells. Z-scores: TDP-43-INSCRIBE (PFA-fixed), 4.61; Enzyme-only control (PFA-fixed), -1.02. D, the actual fraction of eCLIP-overlapping INSCRIBE edit clusters (diamonds) was significantly higher than that of the 20 permuted clusters (box plot, $n = 20$). Top: RBFOX2-INSCRIBE in PFA-fixed cells. Z-scores: RBFOX2-INSCRIBE (PFA-fixed), 18.67; Enzyme-only control (PFA-fixed), 1.36; RBFOX2-STAMP: 5.45; Bottom: TDP-43-INSCRIBE in PFA-fixed cells. Z-scores: TDP-43-INSCRIBE (PFA-fixed), 7.10; Enzyme-only control (PFA-fixed), -0.62. The actual fraction was plotted using the confident edit clusters derived from 3 INSCRIBE technical replicates. The box shows the quartiles while the whiskers extend to show the rest of the distribution; the median is represented by the center line.

(E) Integrative genome viewer (IGV) tracks of example RBFOX2 target gene (PKM, OGT) with 1500bp windows showing eCLIP peaks (orange), INSCRIBE in PFA-fixed cells (blue) and the enzyme-only control (gray). The overlay displays the edit fraction quantified by SAILOR in 3 replicates of INSCRIBE, alongside with the confident edit clusters identified by FLARE.

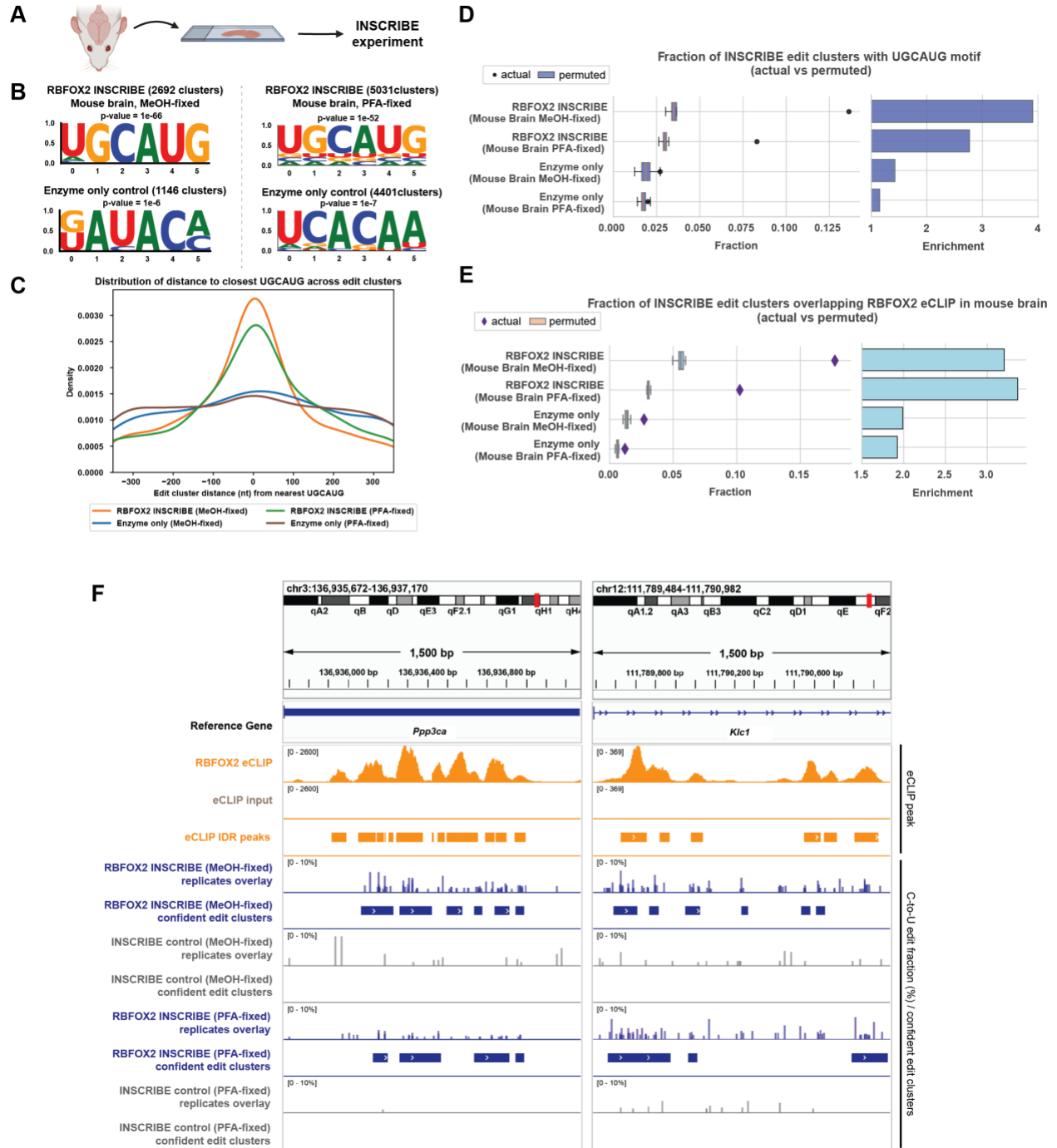


INSCRIBE identifies RBP-RNA interaction sites in fixed tissue samples

Finally, we sought to determine whether INSCRIBE could capture authentic RBP-RNA interactions in primary tissue samples, in which elements like the extracellular matrix could inhibit the binding of a primary antibody to its target RBP, the recognition of the primary antibody by Apo1-Nb, or the editing reaction. To test INSCRIBE's compatibility with primary tissues, we conducted RBFOX2-INSCRIBE in both methanol-fixed and PFA-fixed 20 μ m-thick mouse brain slices (**Figure 2.7A**). Remarkably, INSCRIBE displayed exceptional performance when applied to mouse brain tissue. The canonical RBFOX2 motif UGCAUG was found in both the methanol-fixed and PFA-fixed INSCRIBE confident edit clusters through *de novo* motif discovery (**Figure 2.7B**). The RBFOX2 motif was also enriched across all confident edit clusters compared to enzyme-only controls, with around 17% of edit clusters bearing a UGCAUG motif in the methanol-fixed samples and around 10% in the PFA-fixed samples (**Figure 2.7C, 2.7D**). Furthermore, approximately 13% or 8% of confident edit clusters overlapped RBFOX2 eCLIP peaks in methanol-fixed or PFA-fixed mouse whole brain tissue, with a >3-fold enrichment compared to permuted clusters in both samples (**Figure 2.7E**). Additionally, upon closer examination at individual RBFOX2 target genes, a notable correlation was observed between INSCRIBE confident edit clusters and eCLIP IDR peaks, indicating a strong alignment between the two (**Figure 2.7F and Supplementary Figure 2.8A**). In conclusion, INSCRIBE can capture RBP-RNA interactions in primary tissues, opening the possibility of its application to diverse precious clinical samples.

Figure 2.7 INSCRIBE unveiled RNA targets of RBFOX2 in mouse brain tissue

(A) Schematics of RBFOX2-INSCRIBE with mouse brain tissue (created with BioRender.com). (B) HOMER de novo motif discovery identified the canonical RBFOX2 motif UGCAUG as the top motif in MeOH-fixed and PFA-fixed mouse brain INSCRIBE confident edit clusters, but not the enzyme-only control, using a cumulative hypergeometric distribution for p-values. (C) A density plot of the distribution of the closest UGCAUG distance to confident edit clusters for INSCRIBE in MeOH-fixed mouse brain tissue (orange), PFA-fixed mouse brain tissue (green) and both enzyme-only controls (blue and purple). A peak at 0nt distance indicated the enrichment of UGCAUG in proximity to the edit cluster center. (D-E) D, the actual fraction of UGCAUG-containing mouse brain RBFOX2-INSCRIBE edit clusters (dots) was significantly higher than that of the 20 permuted clusters (box plot, $n = 20$). Z-scores: RBFOX2-INSCRIBE (Mouse Brain MeOH-fixed), 30.81; RBFOX2-INSCRIBE (Mouse Brain PFA-fixed), 28.05; Enzyme-only control (Mouse Brain MeOH-fixed), 1.95; Enzyme-only control (Mouse Brain PFA-fixed), 1.30. E, the actual fraction of mouse brain eCLIP-overlapping mouse brain RBFOX2-INSCRIBE edit clusters (diamonds) was significantly higher than that of the 20 permuted clusters (box plot, $n = 20$) but not the enzyme-only control, demonstrating edit clusters specifically captured mouse RBFOX2 targeted RNA sites. The eCLIP experiment was performed in mouse whole brain and the IDR peaks were used for the comparison. Z-scores: RBFOX2-INSCRIBE (Mouse Brain MeOH-fixed), 37.75; RBFOX2-INSCRIBE (Mouse Brain PFA-fixed), 40.89; Enzyme-only control (Mouse Brain MeOH-fixed), 7.65; Enzyme-only control (Mouse Brain PFA-fixed), 6.06. In D-E, the actual fraction was plotted using the confident edit clusters derived from 3 INSCRIBE technical replicates. Enrichment was defined as the ratio of actual fraction to the mean of the permuted-derived fraction. The box shows the quartiles while the whiskers extend to show the rest of the distribution; the median is represented by the center line. (F) Integrative genome viewer (IGV) tracks of example mouse RBFOX2 target genes (*Ppp3ca*, *Klc1*) with 1500 bp windows showing mouse RBFOX2 eCLIP peaks (orange), MeOH-fixed and PFA-fixed mouse brain RBFOX2-INSCRIBE (blue) and the respective enzyme-only controls (gray). The overlay displays the edit fraction quantified by SAILOR in 3 replicates of INSCRIBE, alongside with the confident edit clusters identified by FLARE.



Chapter 2.3 Discussion

Here we describe INSCRIBE, a novel approach to transcriptomic RBP-RNA interaction profiling with low-input material requirement and a versatile workflow. By harnessing a recombinant APOBEC1-nanobody fusion protein, *in situ* RNA labeling through APOBEC1-mediated cytosine deamination is directed towards RNAs bound by the RBP of interest. INSCRIBE is optimized for fixed cells or primary tissues, without any prior requirement for plasmid construction, transfection, or exogenous overexpression in cells or tissues. This strategy allows for the mapping of endogenous binding sites for RBFOX2 and TDP-43 transcriptome-wide in both methanol- and PFA-fixed HEK293T cells, as well as intact primary tissue samples – mouse brain tissue, with high accuracy and specificity.

Unlike conventional immunoprecipitation-based methods, we show that INSCRIBE seamlessly integrates with long-read sequencing, enabling discrimination of isoform-specific RBP-RNA interactions. Since a single APOBEC1-nanobody fusion recognizes the IgG element of all primary antibodies derived from a single animal system (rabbit in the current implementation), the method is compatible with many commercially available anti-RBP primary antibodies. The method's versatility and user-friendliness render it conveniently applicable to complex experimental systems, such as disease-related animal models, organoids, and clinical samples. It readily bridges the critical gap in RBP-RNA profiling in RNA-related diseases under authentic endogenous biological environments, facilitating insights into a wide spectrum of biological inquiries and disease mechanisms.

Remarkably, the high sensitivity of INSCRIBE enables the transcriptome-wide capture of RBP-RNA interaction with low sequencing depth (as low as 20 million reads) or ultra-low input

material (as low as 100 pg total RNA, equivalent to ~5 mammalian cells). With a typical INSCRIBE workflow, one could therefore investigate the spatial heterogeneity of RBP-RNA interactions in a tissue cost-effectively. For example, INSCRIBE experiments with tissue slices of different coordinates across the brain for RBM5, an alternative splicing regulator that binds RNA differentially in Huntington's disease mouse brain²⁵, could potentially reveal more detailed roles RBM5 plays in distinct brain regions, structures and cortical layers in Huntington's disease. Spatial profiling of RBP-RNA interactions with INSCRIBE would enable the scrutinization of the RNA target changes occurring in tissue micro-environments during disease development and therapeutic response. Currently, we are keen on advancing the technology for single-cell applications and spatial transcriptomics, empowering simultaneous RBP-RNA interaction and RNA localization analysis in primary tissues.

One limitation of the current INSCRIBE protocol is its reliance on highly specific anti-RBP antibodies, which are dependent on animals to produce. We anticipate that the advancement of AI(artificial intelligence)-designed nanobodies could alleviate this challenge²⁶. In the meantime, a broad range of disease-relevant RBPs, including TDP-43, can already be effectively targeted with commercially available high-quality antibodies. Although current INSCRIBE utilizes rabbit-derived primary antibodies, the adaptable nanobody segment can be tailored to target antibodies from various animals, broadening its scope.

Additional challenges faced by INSCRIBE are the inherent sequence and structural preferences in RNA-modifying enzymes and background noise. The APOBEC1 enzyme more readily edits less structured RNA sequences and appears to prefer specific local sequence contexts (5'-A/U of the edited C)^{27,28}. Future endeavors will focus on engineering alternative RNA-modifying enzymes with reduced substrate context preferences. We reason that a

significant contributor to the background noise arises from enzymes that are not immobilized to the antibody during the RNA editing reaction. To mitigate background noise from unbound enzymes, it is essential to carefully choose antibodies with superior performance in immunofluorescence contexts, and to optimize the dilution factor for each antibody. Based on the relatively high edit cluster counts in the enzyme-only controls, we conclude that meticulous optimization of both the anti-RBP-of-interest Ab and APO1-Nb concentrations will be required to minimize background noise. This noise could be particularly problematic for intronic regions or low-level transcripts, where the sequencing read coverage is generally low. Increasing sequencing depth could be helpful if the RBP-of-interest targets intronic or low-level transcripts. We also acknowledge the challenge associated with achieving an optimal signal-to-noise ratio when profiling low-abundance RBPs. In such scenarios, it will be crucial to maintain low concentrations of both the primary antibody and APO1-Nb to minimize nonspecific editing. Future advancements in this technology will prioritize enhancing the binding specificity and affinity of the antibody-IgG-recognizing nanobody.

Chapter 2.4 Methods

Plasmid construction and recombinant protein purification

The amino acid sequence of rat APOBEC1 (NP_037039.1) was codon-optimized for expression in *E. coli* strain and cloned into a modified UC Berkeley Macrolab vector 2CT (Addgene #29706) to generate an N-terminal fusion to a His₆-MBP (Maltose-Binding Peptide) tag. The sequence of an engineered anti-rabbit-IgG nanobody (TP897, Addgene plasmid # 104163)¹⁵ was cloned at the C-terminal of the APOBEC1 with an 18-aa linker (GSGTSGAGSATAGSGAGG) in-between. Plasmid pTP1183 was a gift from Dirk Görlich (Addgene plasmid # 104163; RRID: Addgene_104163). The His-MBP-APOBEC1-Nb (APO1-Nb) protein was expressed in *E. coli* strain Rosetta2 pLysS (EMD Millipore) by growing cultures in 2xYT media to mid-log phase at 37°C, followed by induction with 0.25 mM IPTG at 18°C for 16 hours supplemented with 100 µM ZnCl₂. For protein purification, cells were harvested by centrifugation, suspended in resuspension buffer (25 mM Tris-HCl pH 8.5, 1 M NaCl, 20 mM imidazole, 2 mM β-Mercaptoethanol and 10% glycerol) with 0.1 mg/mL RNase A and lysed by sonication. Lysates were clarified by centrifugation (16,000 rpm 30 min), then the supernatant was loaded onto a Ni²⁺ affinity column (HisTrap HP, Cytiva) pre-equilibrated with the resuspension buffer. The column was washed with a buffer containing 20 mM imidazole and 500 mM NaCl, and eluted with a buffer containing 250 mM imidazole and 250 mM NaCl. The elution was buffer-exchanged into the low salt buffer (20 mM Tris pH8.5, 50mM NaCl, 1mM DTT and 10% glycerol) and loaded onto an anion-exchange column (HiTrap Q HP, Cytiva) following washing with a gradient of 50 mM-1 M NaCl. The eluted fractions containing the protein (evaluated with SDS-PAGE) were collected and loaded onto a size exclusion column

(Superdex 200 Increase 10/300 GL, Cytiva) equilibrated with the size exclusion buffer (20 mM Tris pH7.5, 300 mM NaCl, 1 mM DTT and 10% glycerol). The peak containing the dimerized His-MBP-APOBEC1-Nb protein was collected concentrated by ultrafiltration, aliquoted and frozen at -80°C for future use.

***In vitro* RNA editing assay**

A 300nt RNA from the 3'-UTR of the human *APP* gene was *in vitro*-transcribed (MEGAscript T7 Transcription Kit, Invitrogen) and purified (Zymo RNA Clean & Concentrator). For a typical RNA editing assay, 0.1µM RNA was incubated with 1.2 µM APO1-Nb (or other Enzyme-Nb) at 37°C for 16 hours in RNA editing buffer (20mM Tris-HCl at desired pH, 5 mM DTT, 2U/µL RNaseOUT, desired final concentration of NaCl). The reaction was quenched at 95°C for 10min. The edited RNAs were reverse-transcribed to cDNA (High-Capacity cDNA Reverse Transcription Kit, Thermo) followed by 20-cycles of PCR amplification and Sanger sequencing (Azenta).

Cell culture

HEK293T cells were cultured in DMEM medium (high glucose, Gibco) with 10% (v/v) heat inactivated fetal bovine serum FBS (Gibco) and 5 U/mL Penicillin-Streptomycin (Gibco) at 37°C, 5% CO₂. All cell lines were maintained at 37°C and 5% CO₂ and routinely tested for the presence of mycoplasma.

Animals

Two 2-month-old female C57BL/6J wildtype mice were used in the study. Sex or age was not selected on purpose. The mouse strain was obtained from The Jackson Laboratory and bred and maintained in our laboratory. All mice were housed under specific pathogen-free conditions with a reversed 12-h light–dark cycle maintained at 23°C with 30–70% humidity and

provided with food and water *ad libitum*. The study was carried out in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health and approved by the Animal Care and Use Committee at University of California San Diego.

***in situ* RNA editing and RNA-seq library preparation**

Approximately 150K HEK293T cells were seeded on poly D-lysine (PDL, Sigma)-coated chamber of ibitreat 8-well slide (ibidi). Cells were cultured overnight to settle and attach to the slide. After removing the medium and washing with PBS, the cells were fixed with 250 μ L pre-chilled MeOH or 4% paraformaldehyde (PFA) for 15min on ice. For the PFA-fixed samples, cells were permeabilized with 0.3% Triton X-100 in PBS for 20mins following fixation. All the washings and incubations were done on ice unless otherwise stated. The Washing Buffer consists of PBS with 2mM DTT, 0.05% Tween-20. Following 3 cycles of washing in 300 μ L Washing Buffer, cells were incubated with 100 μ L blocking buffer (5% BSA in Washing Buffer with 0.2U/ μ L RNaseOUT) for 1 hour. After blocking, the cells were incubated in 150 μ L diluted primary antibodies (rabbit anti-RBFOX2, Bethyl, A300-864A; rabbit anti-TDP-43, Bethyl, A303-223A) with optimized concentration in Antibody Dilution Buffer (PBS with 1% BSA, 2mM DTT and 0.2U/ μ L RNaseOUT) for 1 hour. After 3 cycles of washing in PBS-T, 100 μ L 1 μ g/mL APO1-Nb in dilution buffer was added for 1-hour incubation followed by 6 times of washing in PBS-T. The slides were then washed twice in 200 μ L RNA editing buffer (20mM Tris pH7.5, 20mM NaCl, 5mM DTT and 0.2U/ μ L RNaseOUT) on ice. Finally, the slides were incubated with 100 μ L RNA editing buffer at 37°C for overnight (~16 hours). For the “Enzyme only” control, no antibody was added. For the “Buffer only” control, no antibody or enzyme was

added. Three replicates were performed in three separate slides for INSCRIBE experiments, enzyme-only controls and buffer-only controls.

On the next day, cells were harvested in 200 μ L Trizol. Total RNA was extracted (Zymo Direct-zol RNA Purification Kit), followed by DNA depletion (DNase I treatment and Zymo RNA Clean and Concentrator Kit) and ribosomal RNA depletion (Illumina Ribo-Zero Plus rRNA Depletion Kit). For the PFA-fixed samples, cells were reverse-crosslinked prior to RNA extraction with 72U/mL Proteinase K (NEB) in De-crosslinking Buffer (10mM Tris pH8.0, 200mM NaCl, 50mM EDTA, 2% SDS) for 2 hours at 55°C. The lysate post-decrosslinking were collected in Trizol LS. The RNA-seq library was prepared following the instruction of Illumina TruSeq Stranded Total RNA Library Preparation Kit. For the library preparation from ultra-low RNA input, 1 ng or 100 pg total RNA from each sample were used as starting materials in NEBNext® Single Cell/Low Input RNA Library Prep Kit (NEB) following the manufacturers' instructions.

For INSCRIBE in mouse brain tissue, the mouse brain was collected and snap frozen. The frozen brain was then cryosectioned into 20 μ m slices and fixed in pre-chilled 100% methanol for 15 min. It was then kept in methanol at -20°C until INSCRIBE experiment. Each individual replicate of either experiment or control were prepared from one single slice of the brain (coronal section). After 3 cycles of washing in PBS-T, INSCRIBE was performed similarly as described above as fixed cells.

Immunofluorescent imaging

Immunofluorescence (IF) was carried out similarly to the INSCRIBE protocol. All washings were done 3 times with PBS-T and the Antibody Dilution Buffer was used for all dilution series. For the primary antibody dilution series, after primary antibody (rabbit anti-

RBFOX2 or rabbit anti-TDP-43) incubation and washing, the cells were incubated with 1:1000 goat anti-rabbit IgG Alexa 488 (Invitrogen) for 1 hour, followed by DAPI stain and washing before being subjected to confocal imaging. For the APO1-Nb dilution series, after incubation with a fixed optimized concentration of primary antibody, cells were incubated with a series of APO1-Nb dilutions and then 1:1000 mouse anti-MBP antibodies (Proteintech). The cells were then washed and incubated with 1:1000 goat anti-rabbit IgG Alexa 488 and 1:1000 goat anti-mouse IgG Alexa 555 (Invitrogen). Subsequently, immunofluorescence samples were imaged on an LSM-880 fluorescence confocal microscope (Zeiss) equipped with 405 nm, 488 nm, and 561 nm lasers. Images were acquired with a 20x objective for traditional confocal microscopy and 63x for Airyscan superresolution imaging.

Analysis of Immunofluorescence Images

Confocal images were processed and analyzed with Fiji (NIH). The “Threshold” function was used to determine the average intensity of antibody or nanobody staining for the corresponding fluorescence channel; the threshold value was set to capture all cellular intensity of the antibody or nanobody. To determine the fraction of nuclear signal in each image, we created masks of the whole cell using the antibody or nanobody channel and masks of the nucleus using the DAPI channel. The DAPI mask was subtracted from the whole-cell mask to create a cytoplasm-only mask of each image. The masks were subtracted from the original image to remove extraneous intensity values, and the “Threshold” function was used to determine the intensity of the remaining pixels. The total intensity (i.e. area x intensity) of the nuclear antibody signal was divided by the total intensity of the whole cell to calculate the nuclear fraction. At least 400 cells were analyzed per condition. Colocalization of Airyscan images was calculated using the “Colocalization” analysis package in Fiji; the Pearson’s correlation coefficient was

determined for the nanobody and antibody channels of each image. Student's t-test was performed on the Pearson's correlation coefficient of 5 images to determine the p-value.

Illumina sequencing and RNA mapping

RNA-seq libraries were sequenced with single-end reads (100 nucleotides) with a typical read depth of 60M. Reads were then trimmed and filtered for repeat elements using sequences obtained from RepBase (v18.05) with STAR (2.4.0i). Reads that did not map to repeats were then mapped to the hg38 assembly with STAR (2.4.0i) for the HEK293T samples, or mapped to the mm10 assembly with STAR (2.5.2) for the mouse samples. The reads were then sorted with samtools (v1.9) and annotated against Gencode (v29).

SAILOR quantification of RNA edits and FLARE identification of edit clusters

The resulting BAM files were each used as inputs to SAILOR (v1.1.0) to determine C > U edit sites across the hg38 assembly¹². Briefly, SAILOR filters potential artifacts and known SNPs (dbSNP, v151) and returns a set of candidate edit sites and outputs the number of C > U conversions found among aligned reads. The C-to-U editing sites called with a confidence score larger than 0.5 and an editing percentage less than 0.8 were remained as the input to FLARE. The Poisson model-based FLARE accounts for background editing rates to filter false positives from truly edited regions, and scores identified clusters for use in downstream applications (<https://github.com/YeoLab/FLARE>). False discovery rate (FDR) threshold was set to 0.1 and the maximum merge distance was set to 15 nt. Clusters from three replicates were intersected through bedtools (2.27.1). [bedtools.intersect(wa=True, u=True, s=True)] Then, buffer-only clusters were subtracted from the intersected edit clusters and termed “confident edit clusters”.

PacBio long-read sequencing and identification of RNA edits

Iso-seq library preparation of RBFOX2-INSCRIBE and enzyme-only samples was completed using the Kinnex full-length RNA kit (PN 103-072-000) and associated protocol. Libraries were sequenced on the Pacbio Revio system, with one SMRT Cell allocated per sample. Following sequencing, the reads were processed using PacBio's software toolkit. Consensus circular sequence (CCS) reads were assembled using the ccs tool (v8.0.0) and standard parameters. Reads were then de-concatenated using skera (v1.2.0), trimmed using lima (v2.9.0) and refined using refine (v4.1.1). Full-length reads were aligned to the hg38 reference using pbmm2 (v1.13.0) with parameters: --preset ISOSEQ and --sort. Aligned reads were filtered for only primary mapped alignments using the SAM Flag. NanoPlot (v1.32.1)²⁹ was used with parameters: --raw and --tsv_stats. Reads with a quality score below 20, unmapped reads, supplementary alignment reads, secondary alignment reads, and reads aligned to the incorrect strand were excluded from the analysis. After read filtering, mRNA isoform-level counts matrices and read assignments were obtained using IsoQuant v3.3.0³⁰ with the following parameters: --data_type pacbio, --transcript_quantification unique_only, and --gene_quantification unique_only. Using custom scripts, cytosine-to-uracil edits were then called using the data. Putative SNPs were removed from the samples by identifying edits that occur across all replicates and samples. Remaining annotated SNPs were removed using annotations from the dbSNP database³¹. The remaining edits were used to calculate gene and isoform-level editsC, the proportion of edited cytosines to all cytosines in a gene or isoform, respectively.

INSCRIBE edit cluster evaluations

De novo motif discovery was performed using HOMER (v4.9.1)¹⁹ using FLARE edit clusters and a shuffled background for each UTR, CDS, intron and total genic region. A kernel

density estimation (KDE) plot was used to show the distribution of the closest motif (UGCAUG or UGUGUG) distance to confident edit clusters. Edit clusters were shuffled randomly within their respective exons or introns 20 times to derive the shuffled edit clusters. The STAMP experiments serving as comparisons to INSCRIBE were conducted in three replicates in HEK293T cells and computationally processed similarly. The eCLIP experiments compared to INSCRIBE were IDR (irreproducible discovery rate) peaks from RBFOX2 eCLIP in HEK293T cells (GSE77634), TDP-43 eCLIP in K562 cells (ENCODE Data Coordination Center, <https://www.encodeproject.org>) and RBFOX2 eCLIP in mouse whole brain tissue (GSE240326).

Code availability

All custom software packages used for data analysis in this work, including SAILOR, HOMER, and FLARE, is available at <https://github.com/YeoLab/>; other computational methods are described in **Methods**.

Data Availability

All INSCRIBE data sets have been deposited at the Gene Expression Omnibus (GSE240014, GSE240326, GSE263371).

Acknowledgements

Chapter 2, in full, is published at *Nature Communications*: Liang, Qishan; Yu, Tao; Kofman, Eric; Jagannatha, Pratibha; Rhine, Kevin; Yee, Brian A.; Corbett, Kevin D.; Yeo, Gene W., “*High-sensitivity in situ capture of endogenous RNA-protein interactions in fixed cells and primary tissues*”. The dissertation author was the primary investigator and author of this paper.

The authors thank Joshua Schwartz for assistance with mouse brain tissue, and Yan Song for providing eCLIP data for comparisons. We also thank the members in Yeo lab and Corbett lab for the support in preparing the manuscript, especially Karen Mei and Hema Kopalle. This work was supported by the Waitt Advanced Biophotonics Core Facility of the Salk Institute with funding from NIH-NCI CCSG: P30 014195 and the Waitt Foundation. K.R. was supported by the Milton Safenowitz Postdoctoral Fellowship from the ALS Association. Q.L. was supported by an individual predoctoral fellowship from the American Heart Association. The authors acknowledge support from the National Institutes of Health (R35 GM144121 to K.D.C., RF1 MH126719, R01 HG011864, R01 HG004659 to G.W.Y.). This publication includes data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from a National Institutes of Health SIG grant (#S10 OD026929) and data generated at the Sequencing Core Facility, La Jolla Institute utilizing an Illumina NovaSeq 6000 that was acquired through the Shared Instrumentation Grant (SIG) Program (#S10OD025052). Computational resources were provided by the Department of Defense High Performance Computing Modernization Program (HPCMP) and the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center (SDSC)³² We would also like to thank the Stem Cell Genomics Core at the Sanford Stem Cell Institute for providing sequencing services.

Supplemental Table 2.1 DNA, RNA and protein sequences used in the study

Cloning and protein expression	
DNA sequence of	ATGAAATCTTCTCACCATCACCATCACCATGGTTCTTCTATGA
His-MBP-	AAATCGAAGAAGGTAAACTGGTAATCTGGATTAAACGGCGAT
APOBEC1-	AAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAG
Nanobody for	AAAGATACCGGAATTAAAGTCACCGTTGAGCATCCGGATAA
E.coli expression	ACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGCGATG
	GCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCT
	ACGCTCAATCTGGCCTGTTGGCTGAAATCACCCCGGACAAAG
	CGTTCCAGGACAAGCTGTATCCGTTTACCTGGGATGCCGTAC
	GTTACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAG
	CGTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGC
	CAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAAGCTG
	AAAGCGAAAGGTAAGAGCGCGCTGATGTTCAACCTGCAAGA
	ACCGTACTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTA
	TGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGACGT
	GGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCT
	GGTTGACCTGATTAAAAACAAACACATGAATGCAGACACCG
	ATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGAAACAG
	CGATGACCATCAACGGCCCCGTGGGCATGGTCCAACATCGACA
	CCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCA
	AGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAG
	GTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCAAAAGAG

Supplemental Table 2.2 DNA, RNA and protein sequences used in the study (continued)

Cloning and protein expression	
DNA sequence of His-MBP- APOBEC1- Nanobody for E.coli expression (continued)	TTCCTCGAAAACCTATCTGCTGACTGATGAAGGTCTGGAAGCG GTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCT TACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACT ATGGAAAACGCCCAGAAAGGTGAAATCATGCCGAACATCCC GCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGAT CAACGCCGCCAGCGGTGTCGACTGTCGATGAAGCCCTGAA AGACGCGCAGACTAATTCGAGCTCGAACAACAACAATA ACAATAACAACAACCTCGGGATCGAGGAAAACCTGTACTTCC AATCCAATGCAGGTGGTAGTAGCGAGACGGGCCCAGTTGCG GTGGATCCGACTTTACGCCGTCGCATCGAACCCACGAGTTC GAGGTGTTTTTTGACCCTCGTGAAGTGCCTAAAGAAACCTGC CTTTTGTATGAAATCAACTGGGGTGGTCGCCATTCCATCTGGC GTCACACGTCCCAAATACCAATAAGCATGTCGAGGTTAATT TCATCGAAAAATTTACTACCGAACGTTATTTCTGCCCTAATAC ACGCTGTTCCATCACGTGGTTCCTGTCATGGTCACCCTGCGGA GAATGTTCCCGCGCCATCACTGAATTCCTTAGCCGCTACCCG CACGTGACTTTATTCATCTACATTGCTCGCCTTTACCATCATG CTGACCCCCGTAATCGTCAGGGACTTCGCGATCTGATTAGCA GCGGCGTAACGATCCAGATTATGACAGAGCAGGAGTCCGGG TACTGTTGGCGCAACTTTGTGAATTATTCACCTTCCAATGAAG CCCATTGGCCACGTTACCCGCACCTGTGGGTGCGTTTATACGT

Supplemental Table 2.3 DNA, RNA and protein sequences used in the study (continued)

Cloning and protein expression	
DNA sequence of His-MBP- APOBEC1- Nanobody for E.coli expression (continued)	GCTTGAAGTGTACTGCATCATTCTTGGATTGCCCCCTGCCTG AACATCTTGCCTCGTAAACAGCCACAAGTACGTTCTTTACT ATCGCCTTACAATCTTGTTCATTATCAGCGCTTACCTCCTCATA TCCTTTGGGCAACCGGATTGAAGGGGAGTGGGACCTCAGGTG CCGGTTCCGCAACGGCCGGTAGTGGAGCCGGTGGTcaggtgcaatt ggtagagtctgggggaggcttgggtgcaggctgggtgactctctgcgactctcctgtgtagcctctgggc gctccttgatggggccaccatgcgctgggtaccgccaggctccagggaaggagcgagaatttgtgc aggcatttttgggatgagattgggtacggaatatgcagacactgcgaagggccgattcaccatctccag agacaatgccaagaacacgatatatctacaaatgaccaacctgaggtccgaagacacggccatgtatt attgtaatggattagtattcgggggtgaatactggggacaggggacccaggtcacagtctcctccggt ggaTGTTAA

Supplemental Table 2.4 DNA, RNA and protein sequences used in the study (continued)

Cloning and protein expression	
protein sequence of	MKSSHHHHHHGSSMKIEEGKLVWINGDKGYNGLAEVGKKFEK
His-MBP-	DTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQS
APOBEC1-	GLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIY
Nanobody	NKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLI
	AADGGYAFKYENGKYDIKDVGVNDNAGAKAGLTFLVDLIKNGH
	MNADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKVNNGVT
	VLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEG
	LEAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPN
	IPQMSAFWYAVRTAVINAASGRQTVDEALKDAQTNSSSSNNNNN
	NNNNNLGIEENLYFQSNAGGSSETGPVAVDPTLRRRIEPHEFEVF
	FDPRELKRETCCLLYEINWGGRRHSIWRHTSQNTNKHVEVNFIEKF
	TTERYFCPNTRCSITWFLSWSPCGECSRAITEFLSRYPHVTLFIYI
	ARLYHHADPRNRQGLRDLISSGVTIQIMTEQESGYCWRNRFVNYS
	PSNEAHWPYPHLLWVRLYVLELYCIILGLPPCLNILRRKQPQLTF
	FTIALQSCHYQRLPPHILWATGLKSGTSGAGSATAGSGAGGQV
	QLVESGGGLVQAGDSLRLSCVASGRSLDGATMRWYRQAPGKE
	REFVAGIFWDEIGTEYADTAKGRFTISRDNANKNTIYLQMTNLR
	EDTAMYYCNGLVFGGEYWGQGTQVTVSSGGC

Supplemental Table 2.5 DNA, RNA and protein sequences used in the study (continued)

<i>In vitro</i> RNA editing assay	
Human APP gene 3'-UTR RNA sequence (substrate)	CUGUGCUGUAACACAAGUAGAUGCCUGAACUUGAAUUAU CCACACAUCAGUAAUGUAUUCUAUCUCUCUUUACAUUUUG GUCUCUAUACUACAUAUUAUAAUGGGUUUUGUGUACUGUAA AGAAUUUAGCUGUAUCAAAACUAGUGCAUGAAUAGAUUCUC UCCUGAUUAUUUAUCACAUAGCCCCUAGCCAGUUGUAUA UUAUUCUUGUGGUUUGUGACCCAAUUAAGUCCUACUUUAC AUAUGCUUUAAGAAUCGAUGGGGGAUGCUUCAUGUGAACG UGGGAGUUCAGCUGCUUCUC
Reverse Transcription Primer	GAGAAGCAGCTGAACTCCCA
PCR amplification primer (For)	CTGTGCTGTAACACAAGTAGATGC
PCR amplification primer (Rev)	GAGAAGCAGCTGAACTCCCA

Supplemental Table 2.6 Number of editing clusters in individual replicates in the down-sampling experiment

Experiment	100M reads	80M reads	60M reads	40M reads	20M reads
RBFOX2 rep1	332054	274224	208869	137308	59855
RBFOX2 rep2	320945	263340	198986	128527	55096
RBFOX2 rep3	323035	265341	201397	130470	56374
Enzyme only rep1	209515	171721	129981	84215	36624
Enzyme only rep2	275200	224472	169029	108172	46166
Enzyme only rep3	251653	203664	151400	95300	38895
Buffer only rep1	218160	179850	137234	90161	39929
Buffer only rep2	222538	182473	138456	90058	39399
Buffer only rep3	179257	146538	109972	70781	30162

Supplemental Table 2.7 Number of edit clusters prior and after overlapping individual replicates

Experiment	Cluster counts in 3 replicates	Cluster counts after rep1 x rep2:	Cluster counts after rep1 x rep2 x rep3:	Cluster counts after subtracting Buffer-only	Percentage of confident cluster over original cluster per replicate
RBFOX2	231565	25329	6396	5602	0.0266
	198986				
	201397				
TDP43	160067	21573	6432	5751	0.0289
	210273				
	226663				
Enzyme only	129981	13435	2845	2168	0.0144
	169029				
	151400				
Buffer only	137234	12176	2704	N/A	N/A
	134516				
	119376				

Supplemental Table 2.8 Number of edit clusters in each RNA regions associated with sequencing depth (RBFOX2 INSCRIBE)

Category	RBFOX2_2	RBFOX2_4	RBFOX2_6	RBFOX2_8	RBFOX2_10
	0M	0M	0M	0M	0M
3'UTR	38.1	37.1	32.7	28.2	24.4
5'UTR	1.4	1.4	1.4	1.2	1.0
CDS	33.0	23.9	17.1	12.8	9.7
Distant Intron	10.3	19.6	29.1	36.1	42.7
Distant Non-coding Intron	5.0	4.8	4.9	5.1	5.1
Non-coding Exon	6.0	3.9	3.0	2.6	2.6
Proximal Intron	4.9	7.6	10.7	12.7	13.4
Proximal Non-coding Intron	1.3	1.6	1.2	1.2	1.1
All introns	21.5	33.6	45.9	55.1	62.2

Supplemental Figure 2.1 Establishment and optimization of APO1-Nb mediated RNA editing *in situ*

(A) Size Exclusion Chromatography at the final step of APO1-Nb purification. The green arrow indicated the dimerized His-MBP-APO1-Nb.

(B) SDS-PAGE image of the dimerized His-MBP-APO1-Nb confirmed the purity of the protein.

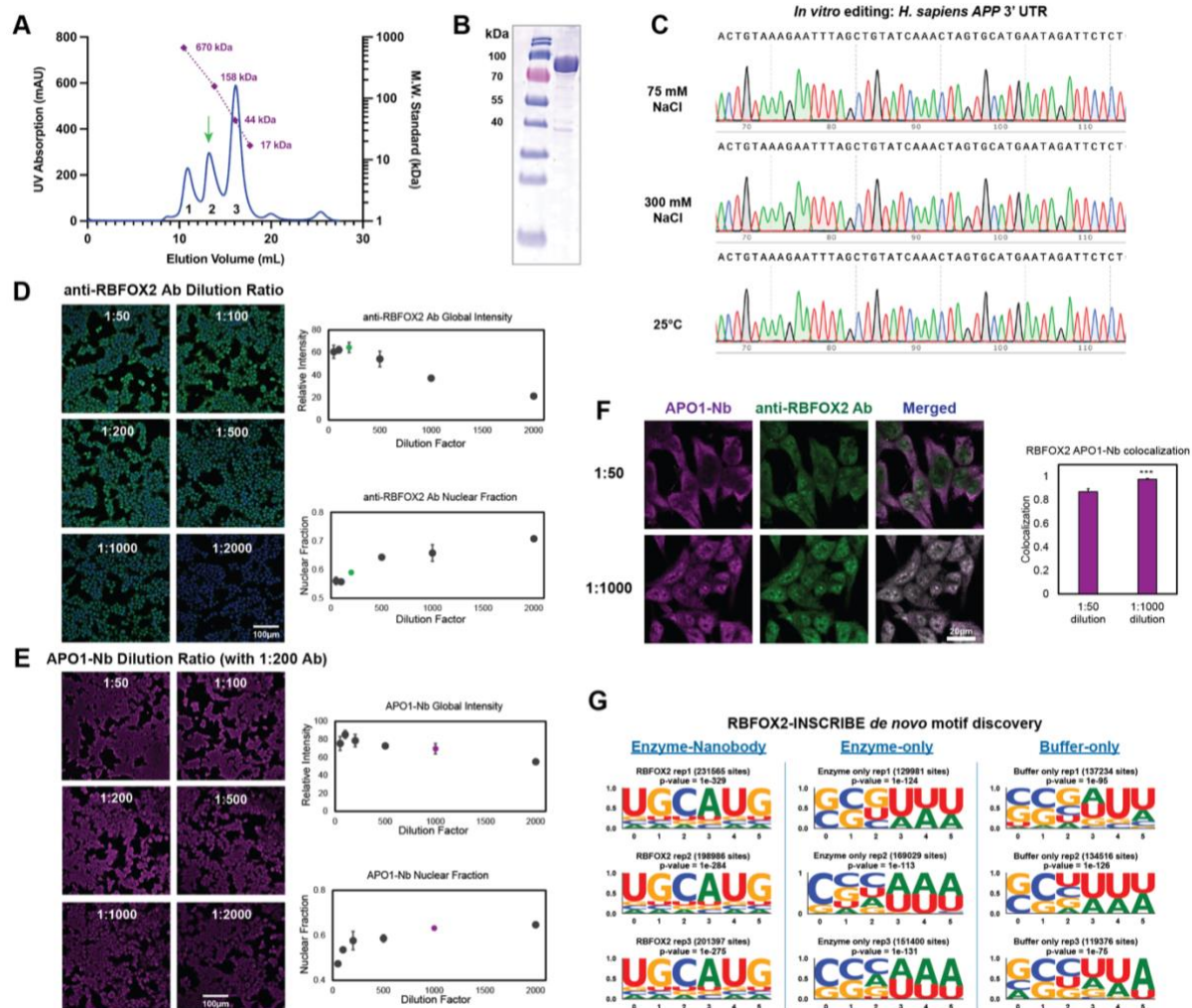
(C) Additional sanger sequencing results of the *in vitro* RNA editing assay. No obvious C-to-U editing was observed at high salt (75mM or 300mM NaCl) at 37°C or low salt at 25°C.

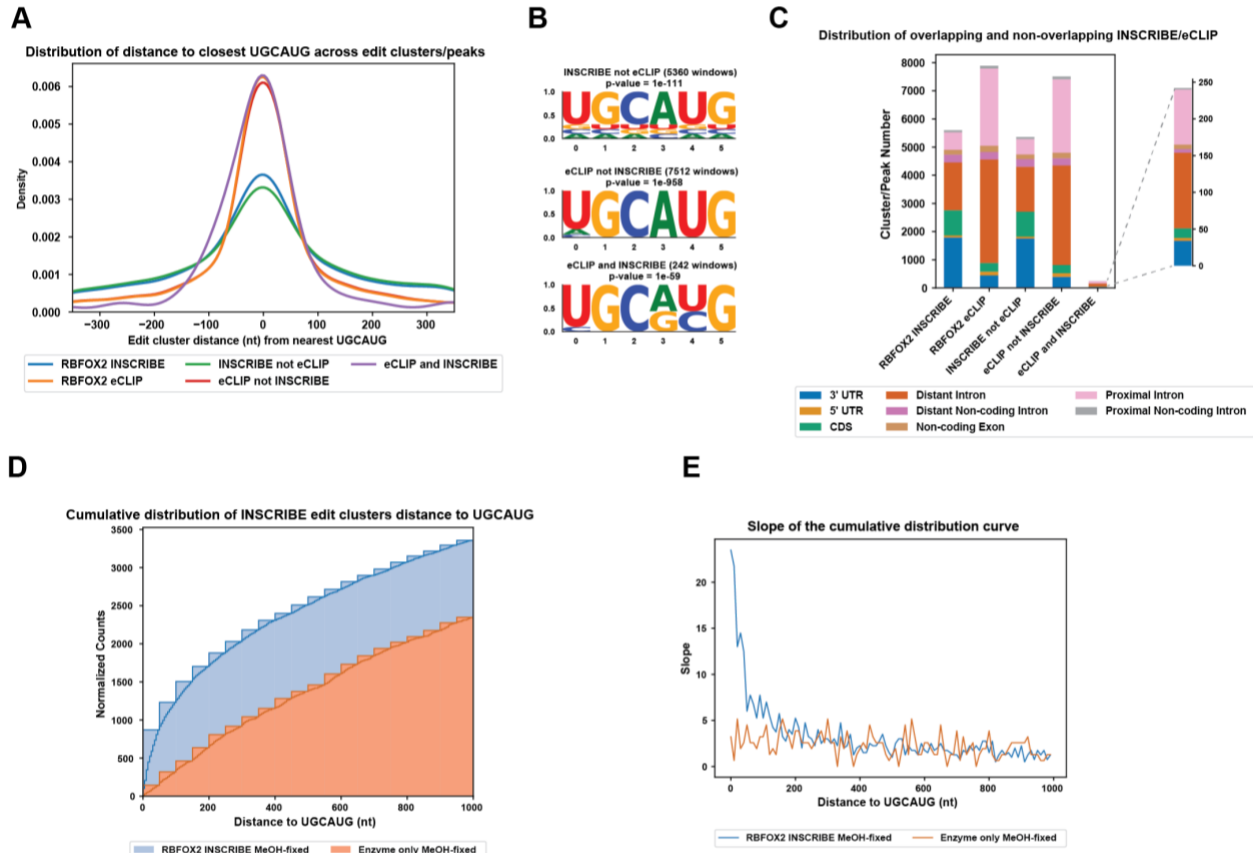
(D) Immunofluorescence (IF) images and quantified global intensities and nuclear fractions of anti-RBFOX2 Ab in HEK293T cells with different Ab dilution ratios. 1:200 dilution (highlighted as green dot) from 1mg/mL stock was chosen for the highest global intensity. (Green: anti-RBFOX2 Ab. Blue: DAPI). The dots indicate the average of global intensities or nuclear fractions of individual cells while the error bars indicate standard deviations, $n > 400$.

(E) IF images and quantified global intensities and nuclear fractions of APO1-Nb in HEK293T cells with fixed anti-RBFOX2 Ab concentration (1:200 dilution). (Magenta: APO1-Nb) The global intensity of APO1-Nb decreased at higher APO1-Nb dilution factors, whereas the nuclear fraction increased and saturated at around 1:1000 dilution (from 1mg/mL stock). The 1:1000 dilution (highlighted in magenta) was chosen with both relatively high global intensity and nuclear fraction since RBFOX2 localized in the nucleus. The dots indicate the average of global intensities or nuclear fractions of individual cells while the error bars indicate standard deviations, $n > 400$.

(F) Quantified colocalization between APO1-Nb and anti-RBFOX2 Ab at fixed anti-RBFOX2 Ab concentration. The colocalization was significantly higher ($p = 0.000484$, Z score = 4.17) at 1:1000 dilution than at 1:50 dilution of APO1-Nb from 1mg/mL stock. (Green: anti-RBFOX2 Ab. Magenta: APO1-Nb). Bar plot indicate the average of Pearson's correlation coefficient of individual field of images, the error bars indicate standard deviations (CI = 0.8412-0.8996; 0.9691-0.9833, $\alpha = 0.01$), $n = 5$.

(G) *De novo* motif discovery of INSCRIBE replicates identified the canonical RBFOX2 motif UGCAUG as the top motif in all RBFOX2-INSCRIBE replicates, but not the negative controls (enzyme-only and buffer-only).





Supplemental Figure 2.3 Comparison of INSCRIBE to eCLIP for RBFOX2

(A) Density plot of the distribution of the closest UGCAUG distance to confident edit clusters for RBFOX2-INSCRIBE, peaks for RBFOX2-eCLIP, clusters in INSCRIBE but not eCLIP, peaks in eCLIP but not INSCRIBE, and INSCRIBE clusters that overlap with eCLIP peaks.

(B) The canonical RBFOX2 binding motif UGCAUG was identified as the top motif in *de novo* motif analysis of clusters in INSCRIBE but not eCLIP, peaks in eCLIP but not INSCRIBE, and INSCRIBE clusters that overlap with eCLIP peaks.

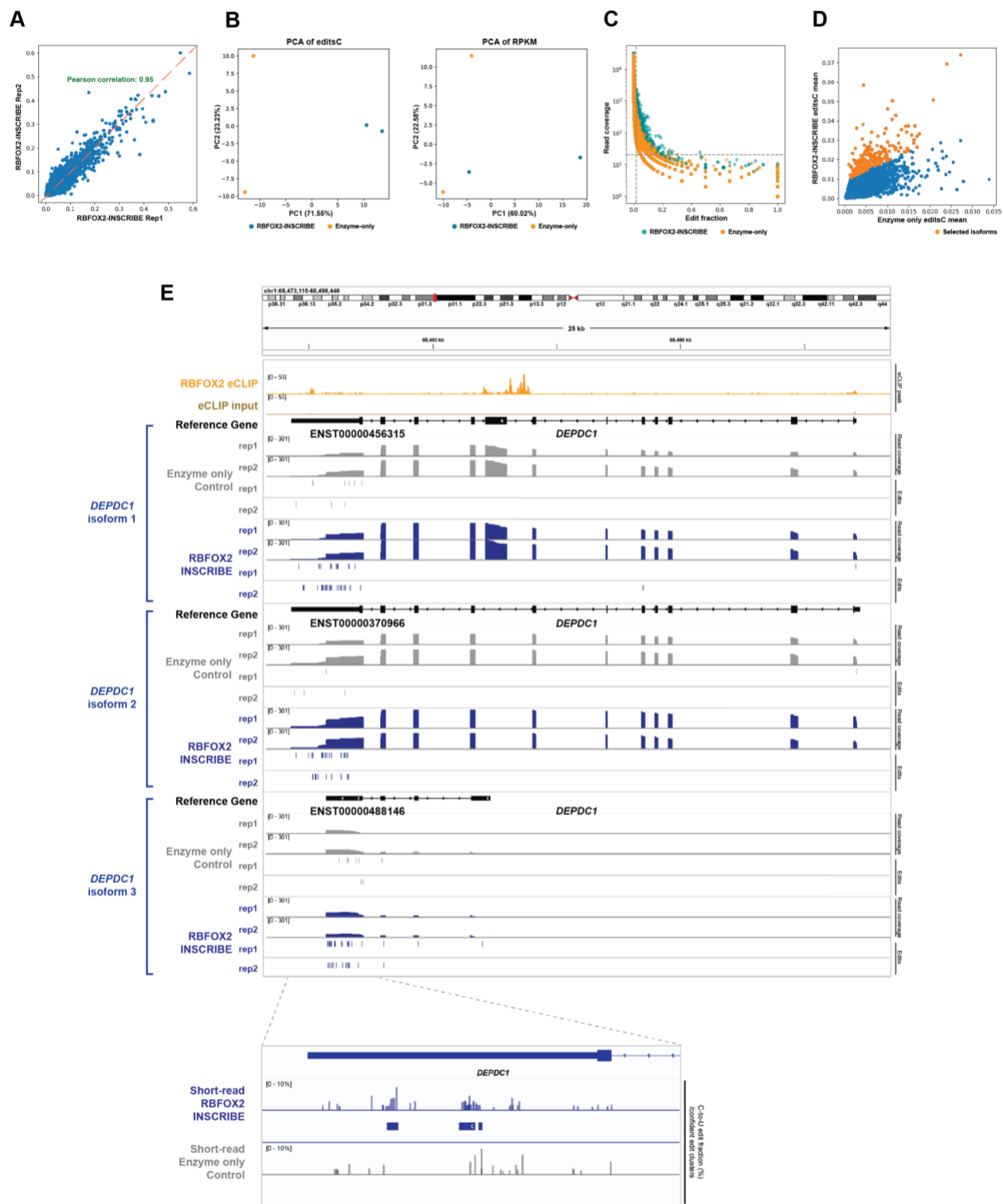
(C) Transcriptomic regions distribution of clusters/peaks of RBFOX2-INSCRIBE, RBFOX2-eCLIP, clusters/peaks in INSCRIBE but not eCLIP, clusters/peaks in eCLIP but not INSCRIBE, and clusters/peaks in both INSCRIBE and eCLIP.

(D) Cumulative distribution of INSCRIBE clusters or enzyme-only clusters with increased distance to UGCAUG. Counts were normalized to the total number of edit clusters in either the RBFOX2-INSCRIBE or enzyme-only control.

(E) The slopes of the cumulative distribution curve of RBFOX2-INSCRIBE and enzyme-only control, which converge at ~200nt from UGCAUG and indicate the RBFOX2-INSCRIBE edit radius is ~200nt from actual binding sites.

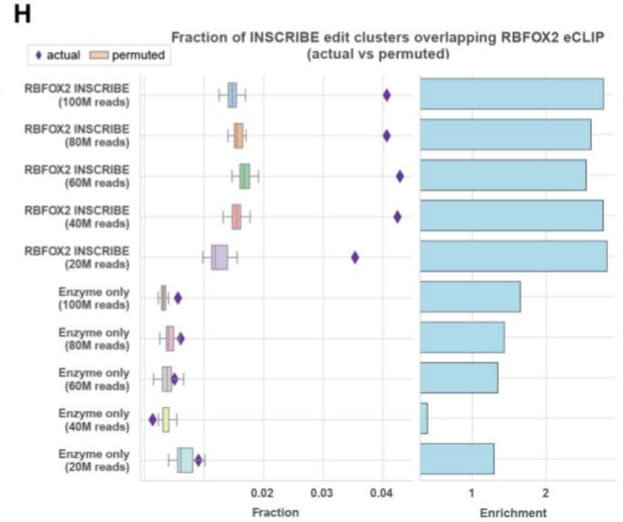
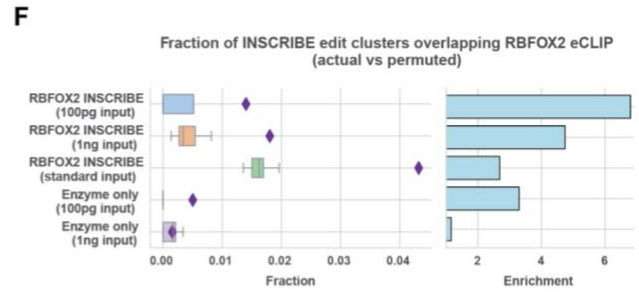
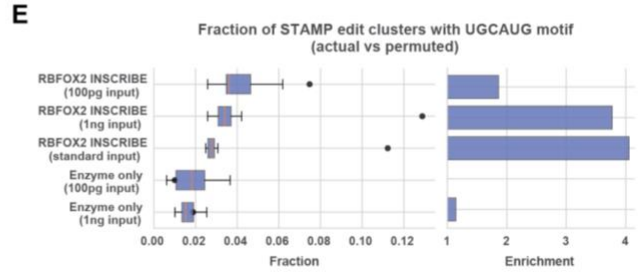
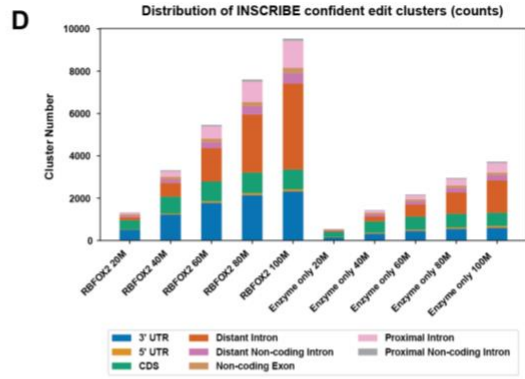
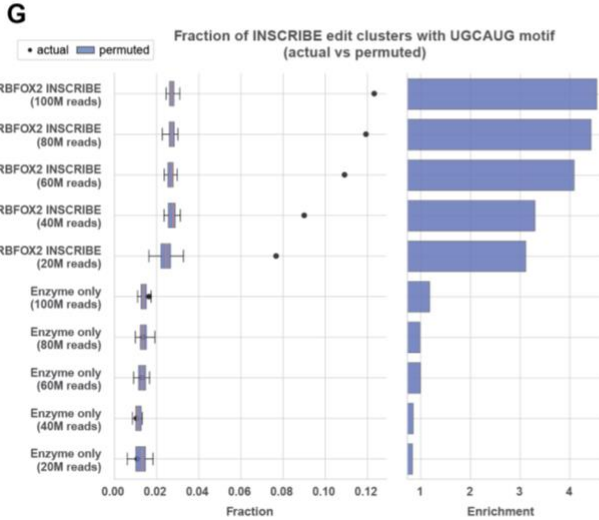
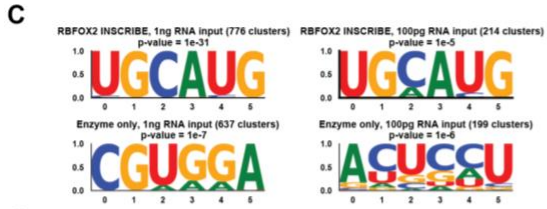
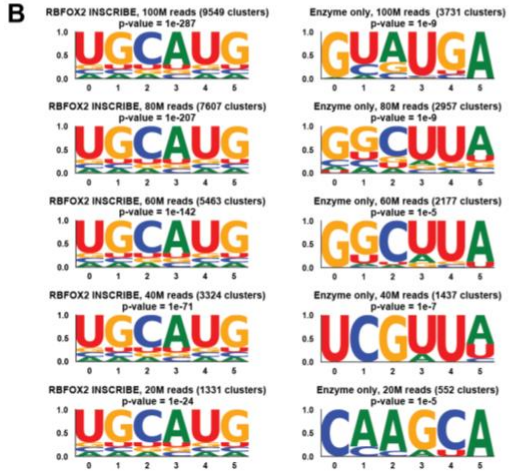
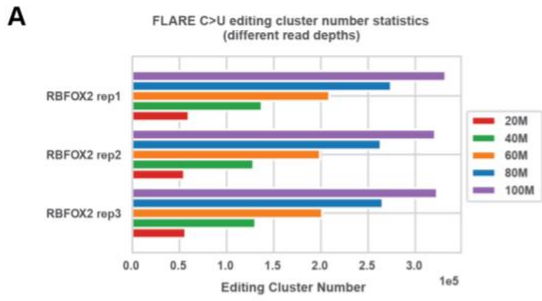
Supplemental Figure 2.4 PacBio sequenced INSCRIBE enables RNA isoform distinctions

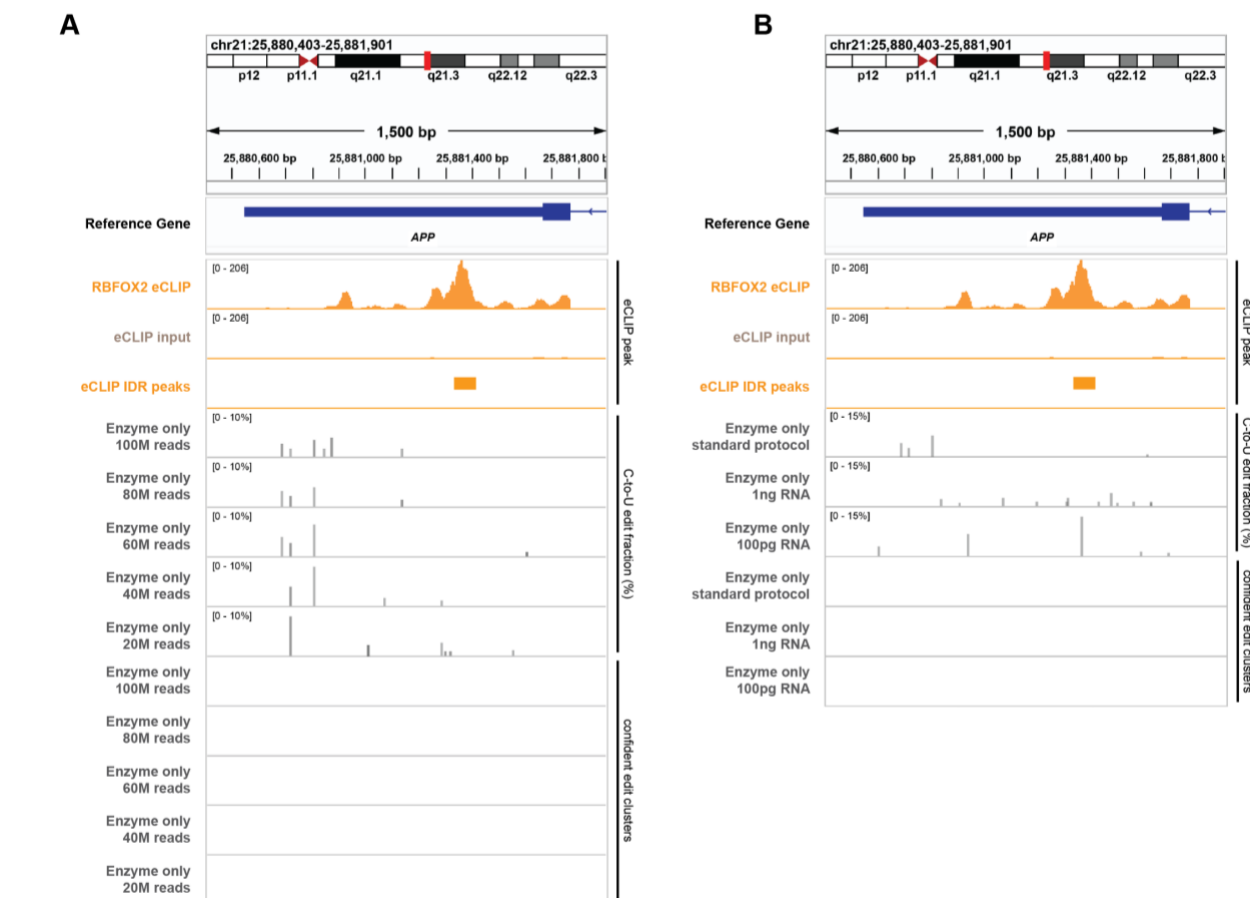
- (A) Correlation of editsC (ratio of number of edited Cs relative to the number of Cs across exons and UTRs of isoforms) of RNA isoforms in different replicates. Pearson correlation: 0.95 between replicate 1 and 2.
- (B) Principal component analysis (PCA) on editsC or RPKM of RBFOX2-INSCRIBE and enzyme-only control both replicates.
- (C) Association between read coverage and edit fraction of individual RNA isoforms. Orange: enzyme-only control; Cyan: RBFOX2-INSCRIBE.
- (D) Mean editsC of each isoform in both RBFOX2-INSCRIBE and enzyme-only control. Orange: isoforms with over 1.5 standard deviation higher mean editsC in RBFOX2-INSCRIBE than enzyme-only control.
- (E) GV tracks showing PacBio sequenced RBFOX2-INSCRIBE vs enzyme-only control edits on three isoforms of *DEPDC1* in a 20kb window, alongside with INSCCRIBE read coverage, eCLIP peaks (HEK293T). The short-read sequenced RBFOX2-INSCRIBE vs enzyme-only control edit fractions and the confident edit clusters are shown at the bottom panel for comparison.



Supplemental Figure 2.5 The performance of INSCRIBE at various sequencing depths and ultra-low RNA input (continued)

- (A) Comparison of number of C-to-U edit clusters identified by FLARE for RBFOX2-INSCRIBE in MeOH-fixed HEK293T cells with different sequencing depths. The edit cluster number was proportionally associated with sequencing depth.
- (B) Irrespective of sequencing depths, the canonical RBFOX2 binding motif UGCAUG was identified as the top motif in *de novo* motif analysis of RBFOX2-INSCRIBE confident edit clusters, as opposed to the enzyme-only controls.
- (C) The canonical RBFOX2 binding motif UGCAUG was discovered in RBFOX2-INSCRIBE when the amount of input RNA was ultra-low (1 ng RNA, top motif; 100 pg RNA, 3rd top motif). UGCAUG motif was not discovered in all enzyme-only controls.
- (D) Comparison of the distribution of INSCRIBE edit clusters with different sequencing depths. Both the total counts and the percentage of intronic edit clusters of in RBFOX2-INSCRIBE increased upon deeper sequencing, indicating the intronic RBFOX2 binding sites were revealed more thoroughly at higher sequencing depths.
- (E) The actual fraction of UGCAUG-containing RBFOX2-INSCRIBE edit clusters (dots) compared with that of the 20 permuted clusters (box plot) with starting total RNA materials. Both lower input RNA libraries enriched UGCAUG as in the standard input library. The box shows the quartiles while the whiskers extend to show the rest of the distribution; the median is represented by the center line; outliers are not shown. Z-scores: RBFOX2-INSCRIBE 100 pg input, 1 ng input, standard input: 3.55, 16.94, 49.93; enzyme-only control 100 pg input, 1 ng input: -0.78, 0.61.
- (F) Comparison of eCLIP-overlapping fraction with different sequencing depths (box plot). Regardless of the amount of RNA starting material, the same INSCRIBE experiment exhibited similar enrichment. The box shows the quartiles while the whiskers extend to show the rest of the distribution; the median is represented by the center line; outliers are not shown. Z-scores: RBFOX2-INSCRIBE 100 pg input, 1 ng input, standard input: 4.73, 7.30, 18.84; enzyme-only control 100 pg input, 1 ng input: 1.07, 7.30.
- (G) The actual fraction of UGCAUG-containing RBFOX2-INSCRIBE edit clusters (dots) compared with that of the 20 permuted clusters (box plot) with different sequencing depths. The box shows the quartiles while the whiskers extend to show the rest of the distribution; the median is represented by the center line; outliers are not shown. The actual fraction and the enrichment were both increased in RBFOX2-INSCRIBE and remained similar in enzyme-only control with deeper sequencing. Z-scores: RBFOX2-INSCRIBE 100M, 80M, 60M, 40M, 20M: 60.88, 43.00, 49.20, 29.63, 12.28. Enzyme only 100M, 80M, 60M, 40M, 20M: 1.32, -0.11, -0.05, -0.50, -0.56.
- (H) Comparison of eCLIP-overlapping fraction with different sequencing depths (box plot). Regardless of sequencing depth, the same INSCRIBE experiment exhibited similar overlapped fraction and enrichment. The box shows the quartiles while the whiskers extend to show the rest of the distribution; the median is represented by the center line; outliers are not shown. Z-scores: RBFOX2-INSCRIBE 100M, 80M, 60M, 40M, 20M: 25.44, 25.95, 19.45, 19.12, 12.32. Enzyme only 100M, 80M, 60M, 40M, 20M: 2.49, 2.07, 0.97, 19.12, 0.73.

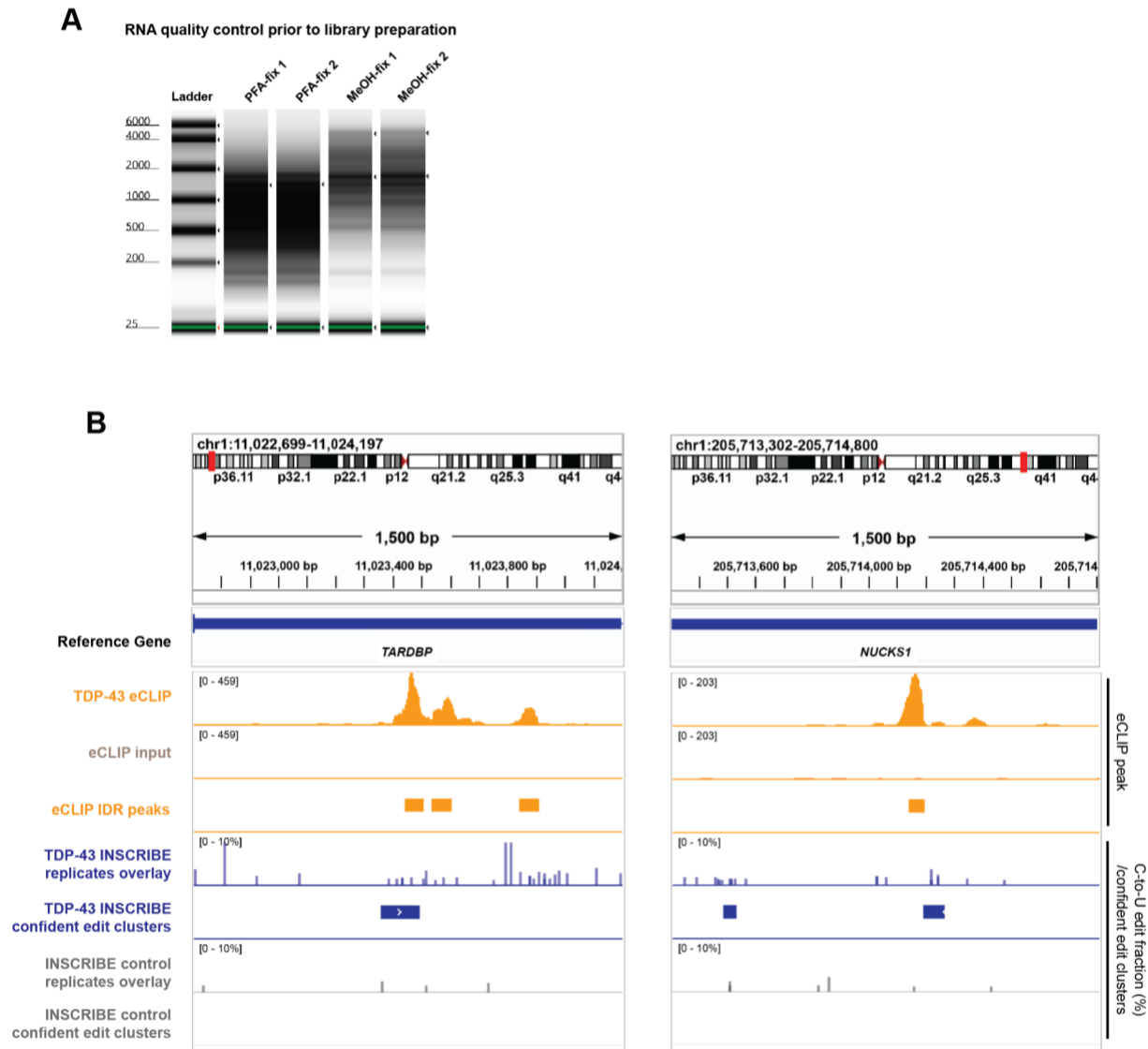




Supplemental Figure 2.6 Additional IGV tracks of INSCRIBE controls with various sequencing depths and ultra-low RNA input

(A) IGV tracks of representative RBFOX2 target gene (*APP*) examining the effect of sequencing depth on SAILOR-quantified C-to-U conversion pattern and FLARE confident edit clusters. The tracks include eCLIP peaks (orange), 3-replicate overlay of SAILOR-quantified edit fraction (gray, upper tracks) and FLARE confident edit clusters (gray, lower tracks) of enzyme-only control.

(B) IGV tracks of representative RBFOX2 target gene (*APP*) examining the effect of input starting total RNA on SAILOR-quantified C-to-U conversion pattern and FLARE confident edit clusters. The tracks include eCLIP peaks (orange), 3-replicate overlay of SAILOR-quantified edit fraction (gray, upper tracks) and FLARE confident edit clusters (gray, lower tracks) of enzyme-only control.

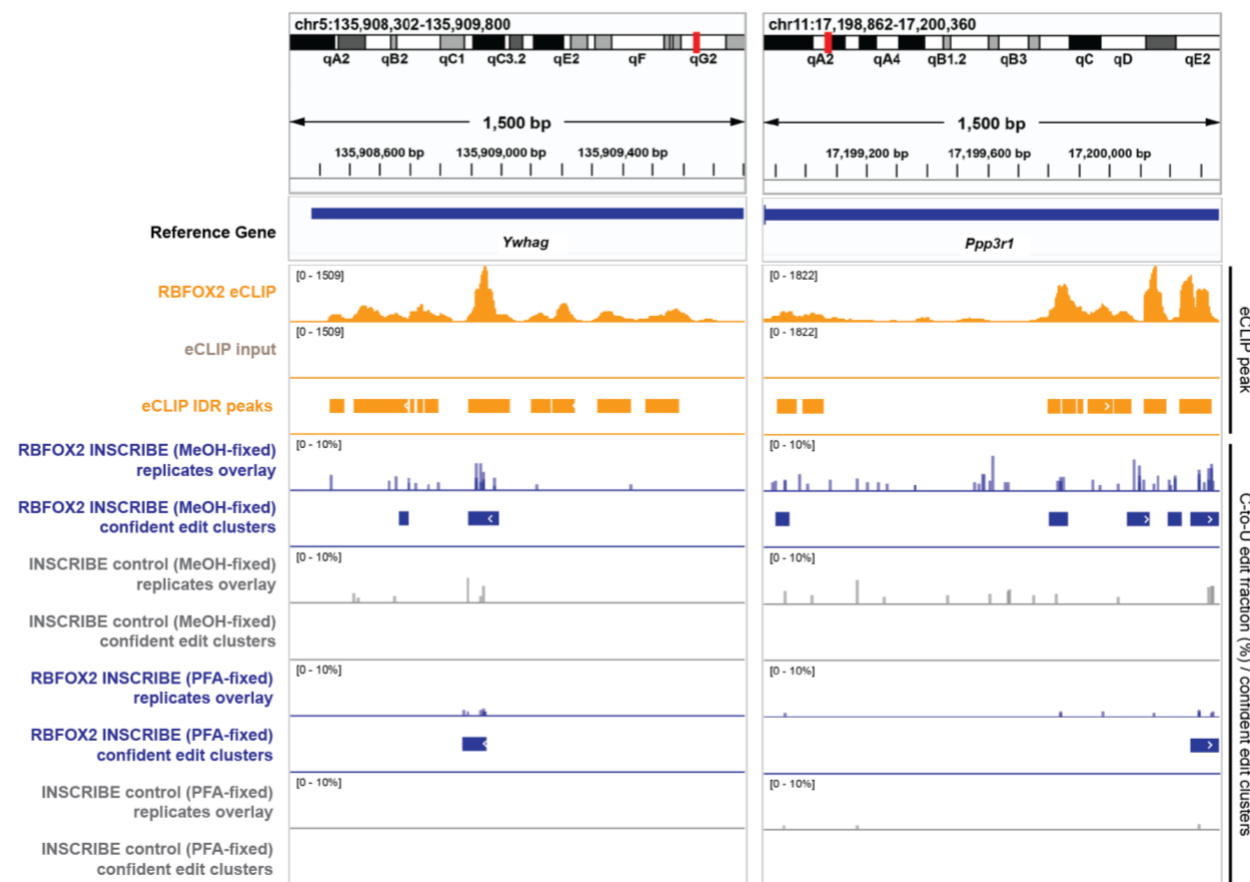


Supplemental Figure 2.7 INSCRIBE exhibited comparable performance in PFA-fixed cells (continued).

(A) RNA quality examination of harvested RNA from PFA-fixed HEK293T cells and MeOH-fixed HEK293T cells by electrophoresis analysis with a TapeStation system post *in situ* RNA editing. Despite exhibiting a lower average size, the RNAs derived from PFA-fixed cells remain suitable for library preparation for Illumina short read sequencing.

(B) IGV tracks of example TDP-43 target genes (*TARDBP*, *NUCKS1*) with 1500bp windows showing eCLIP peaks (orange), INSCRIBE in PFA-fixed cells (blue) and the enzyme-only control (gray). The overlay displays the edit fraction quantified by SAILOR in 3 replicates of INSCRIBE, alongside with the confident edit clusters identified by FLARE.

A



Supplemental Figure 2.8 Additional IGV tracks of RBFOX2-mouse brain RBFOX2-INSCRIBE (A) IGV tracks of example mouse RBFOX2 target genes (*Ywhag*, *Ppp3r1*) with 1500 bp windows showing mouse RBFOX2 eCLIP peaks (orange), MeOH-fixed and PFA-fixed mouse brain RBFOX2-INSCRIBE (blue) and the respective enzyme-only controls (gray). The overlay displays the edit fraction quantified by SAILOR in 3 replicates of INSCRIBE, alongside with the confident edit clusters identified by FLARE.

REFERENCES

1. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat Rev Genet.* 2014 Dec;15(12):829–45.
2. Ule J, Blencowe BJ. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Mol Cell.* 2019 Oct 17;76(2):329–45.
3. Das S, Vera M, Gandin V, Singer RH, Tutucci E. Intracellular mRNA transport and localized translation. *Nat Rev Mol Cell Biol.* 2021 Jul;22(7):483–504.
4. Prashad S, Gopal PP. RNA-binding proteins in neurological development and disease. *RNA Biol.* 2021 Jul 3;18(7):972–87.
5. Pereira B, Billaud M, Almeida R. RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends Cancer.* 2017 Jul 1;3(7):506–28.
6. Wolozin B, Ivanov P. Stress granules and neurodegeneration. *Nat Rev Neurosci.* 2019 Nov;20(11):649–66.
7. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell.* 2010 Apr 2;141(1):129–41.
8. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, Stanton R, Rigo F, Guttman M, Yeo GW. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods.* 2016 Jun;13(6):508–14.
9. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP identifies Nova-regulated RNA networks in the brain. *Science.* 2003 Nov 14;302(5648):1212–5.
10. König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol.* 2010 Jul;17(7):909–15.
11. Lee FCY, Ule J. Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Mol Cell.* 2018 Feb 1;69(3):354–69.
12. Brannan KW, Chaim IA, Marina RJ, Yee BA, Kofman ER, Lorenz DA, Jagannatha P, Dong KD, Madrigal AA, Underwood JG, Yeo GW. Robust single-cell discovery of RNA targets of RNA-binding proteins and ribosomes. *Nat Methods* 2021 185. 2021 May 7;18(5):507–19.
13. McMahon AC, Rahman R, Jin H, Shen JL, Fieldsend A, Luo W, Rosbash M. TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. *Cell.* 2016 Apr 21;165(3):742–53.

14. Wolfe AD, Li S, Goedderz C, Chen XS. The structure of APOBEC1 and insights into its RNA and DNA substrate selectivity. *NAR Cancer*. 2020 Dec 1;2(4):zca027.
15. Pleiner T, Bates M, Görlich D. A toolbox of anti-mouse and anti-rabbit IgG secondary nanobodies. *J Cell Biol*. 2017 Dec 20;217(3):1143–54.
16. Wang X, Yu L, Wu AR. The effect of methanol fixation on single-cell RNA sequencing data. *BMC Genomics*. 2021 Dec 5;22(1):420.
17. Deffit SN, Yee BA, Manning AC, Rajendren S, Vadlamani P, Wheeler EC, Domissy A, Washburn MC, Yeo GW, Hundley HA. The *C. elegans* neural editome reveals an ADAR target mRNA required for proper chemotaxis. *eLife*. 2017 Sep 19;6:e28625.
18. Kofman E, Yee B, Medina-Munoz HC, Yeo GW. FLARE: a fast and flexible workflow for identifying RNA editing foci. *BMC Bioinformatics*. 2023 Oct 2;24(1):370.
19. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell*. 2010 May 28;38(4):576–89.
20. Begg BE, Jens M, Wang PY, Minor CM, Burge CB. Concentration-dependent splicing is enabled by Rbfox motifs of intermediate affinity. *Nat Struct Mol Biol*. 2020 Oct;27(10):901–12.
21. Fu S, Wang A, Au KF. A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biol*. 2019 Feb 4;20(1):26.
22. Tziortzouda P, Van Den Bosch L, Hirth F. Triad of TDP43 control in neurodegeneration: autoregulation, localization and aggregation. *Nat Rev Neurosci*. 2021 Apr;22(4):197–208.
23. Doerr A. A fix for formaldehyde fixation. *Nat Methods*. 2015 Oct;12(10):906–7.
24. Evers DL, Fowler CB, Cunningham BR, Mason JT, O’Leary TJ. The Effect of Formaldehyde Fixation on RNA: Optimization of Formaldehyde Adduct Removal. *J Mol Diagn*. 2011 May 1;13(3):282–8.
25. Mullari M, Fossat N, Skotte NH, Asenjo-Martinez A, Humphreys DT, Bukh J, Kirkeby A, Scheel TKH, Nielsen ML. Characterising the RNA-binding protein atlas of the mammalian brain uncovers RBM5 misregulation in mouse models of Huntington’s disease. *Nat Commun*. 2023 Jul 19;14(1):4348.
26. Cao L, Coventry B, Goreshnik I, Huang B, Sheffler W, Park JS, Jude KM, Marković I, Kadam RU, Verschueren KHG, Verstraete K, Walsh STR, Bennett N, Phal A, Yang A, Kozodoy L, DeWitt M, Picton L, Miller L, Strauch EM, DeBouver ND, Pires A, Bera AK, Halabiya S, Hammerson B, Yang W, Bernard S, Stewart L, Wilson IA, Ruohola-Baker H, Schlessinger J, Lee S, Savvides SN, Garcia KC, Baker D. Design of protein-binding proteins from the target structure alone. *Nature*. 2022 May;605(7910):551–60.

27. Pecori R, Di Giorgio S, Paulo Lorenzo J, Nina Papavasiliou F. Functions and consequences of AID/APOBEC-mediated DNA and RNA deamination. *Nat Rev Genet.* 2022 Aug;23(8):505–18.

28. Medina-Munoz HC, Kofman E, Jagannatha P, Boyle EA, Yu T, Jones KL, Mueller JR, Lykins GD, Doudna AT, Park SS, Blue SM, Ranzau BL, Kohli RM, Komor AC, Yeo GW. Expanded palette of RNA base editors for comprehensive RBP-RNA interactome studies. *Nat Commun.* 2024 Jan 29;15(1):875.

29. De Coster W, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics.* 2023 May 1;39(5):btad311.

30. Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, Tilgner HU. Accurate isoform discovery with IsoQuant using long reads. *Nat Biotechnol.* 2023 Jul;41(7):915–8.

31. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001 Jan 1;29(1):308–11.

32. San Diego Supercomputer Center (2022): Triton Shared Computing Cluster. University of California, San Diego. Service. 2022 [cited 2024 Mar 4]; Available from: <https://tritoncluster.sdsc.edu>

CONCLUSION

The current thesis aims to unveil RBP-RNA recognition and interaction, by biochemistry investigation and by technology development. In summary, I have demonstrated the followings in this thesis:

- In the engineering of the ZRANB2 ZnF, an N-to-R mutant was identified to have altered RNA binding preference by rational design and screening with RNA bind-n-seq assay. (Chapter 1)
- The conformation and hydrogen bonding pattern of ZRANB2 ZnF in RNA recognition was characterized by molecular dynamics simulation. (Chapter 1)
- The underlying knowledge and principles of engineering of compact, designer RNA-targeting proteins for a modular RNA-targeting system are expanded. (Chapter 1)
- A novel technology INSCRIBE was developed to enable high sensitivity, isoform-aware, and endogenous RBP-RNA interaction profiling *in situ*. (Chapter 2)
- INSCRIBE is broadly applicable to diverse samples, including methanol- or formaldehyde-fixed HEK293T cells and mouse brain tissue. (Chapter 2)
- INSCRIBE has the potential to revolutionize our understanding in disease development and therapeutic responses by enabling applications in non-engineerable cells, primary tissues, and valuable clinical samples. (Chapter 2)

Although the aim of the research project described in Chapter 1 was to develop a modular RNA-targeting toolbox using a series of ZnF protein, the goal was not achieved during the course of my Ph.D. study. Only one zinc finger mutant was identified as a potentially novel module within the toolbox. Further exploration of additional ZRANB2 ZnF mutants and other

RanBP-type ZnF proteins is required to expand the toolbox. Besides, cooperative RNA binding of the modular ZnF array needs to be characterized to ensure relatively high binding affinity of ZnF array-RNA binding. Subsequently, to advance the toolbox for therapeutic applications, functional characterizations in human cell lines and model organisms that evaluate the synergistic effects and compatibility of the modular ZnF array are necessary.

In contrast to the challenges faced in Chapter 1, the technology INSCRIBE in Chapter 2 is developed more maturely. The experimental protocol and data analysis pipeline are intricately designed and well-integrated. At this stage, it readily profiles the whole-transcriptomic RNA binding of two RBPs in various samples with diverse fixation methods, although the low enzymatic activity of wild-type APOBEC1 limits the improvement of signal-to-noise ratio.

To enhance signal-to-noise ratio and thus further refine INSCRIBE technology, it's crucial to explore of alternative RNA modifying enzymes (e.g., other wild-type or engineered APOBEC enzymes as well as TadA enzymes) with high RNA editing specificity and moderate deamination activity. Such enzymes as TadA8r¹ have the potential to significantly enhance the signal-to-noise ratio of RNA editing, and therefore allows INSCRIBE to identify RBP binding sites with higher confidence.

Excitingly, to unleash the full potential of the technology, the future directions for INSCRIBE applications are the following: (1) develop the enzyme-nanobody reagent that are applicable for primary antibodies from different animal origins to broaden the range of available for profiling; (2) apply INSCRIBE in protein translation profiling by directing precise RNA editing in adjacent to ribosomes (“ribo-INSCRIBE”), which allows sensitive and convenient ribosomal profiling; (3) couple INSCRIBE with spatial transcriptomic technologies such as 10x

Genomics² to allow RBP-RNA interaction, RBP 3D location and RNA 3D location detection all in one sample.

Through the above efforts, I envision a future world with a deeper understanding of the intricate symphony orchestrated by RNAs and RNA-binding proteins in cellular functions. This knowledge will greatly advance novel targeted therapies, diagnostic tools, and personalized medicine for RNA-related and RBP-related diseases.

.

REFERENCES

1. Xiao YL, Wu Y, Tang W. An adenine base editor variant expands context compatibility. *Nat Biotechnol.* 2024 Jan 2;1–12.
2. Janesick A, Shelansky R, Gottscho AD, Wagner F, Williams SR, Rouault M, Beliakoff G, Morrison CA, Oliveira MF, Sicherman JT, Kohlway A, Abousoud J, Yu Drennon T, Mohabbat SH, 10x Development Teams, Taylor SEB. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat Commun.* 2023 Dec 19;14(1):8353.